BAYESIAN DENSITY REGRESSION AND PREDICTOR-DEPENDENT CLUSTERING

by
Ju-Hyun Park

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. Michael Kosorok, Reader
Dr. Amy Herring, Reader
Dr. Hongtu Zhu, Reader
Dr. Kenneth Bollen, Reader

ABSTRACT

JU-HYUN PARK: Bayesian Density Regression and Predictor-Dependent Clustering.
(Under the direction of Dr. David Dunson.)

Mixture models are widely used in many application areas, with finite mixtures of Gaussian distributions applied routinely in clustering and density estimation. With the increasing need for a flexible model for predictor-dependent clustering and conditional density estimation, mixture models are generalized to incorporate predictors with infinitely many components in the semiparametric Bayesian perspective. Much of the recent work in the nonparametric Bayes literature focuses on introducing predictor-dependence into the probability weights.

In this dissertation we propose three semiparametric Bayesian methods, with a focus on the applications of predictor-dependent clustering and condition density estimation. We first derive a generalized product partition model (GPPM), starting with a Dirichlet process (DP) mixture model. The GPPM results in a generalized Pólya urn scheme. Next, we consider the problem of density estimation in cases where predictors are not directly measured. We propose a model that relies on Bayesian approaches to modeling of the unknown distribution of latent predictors and of the conditional distribution of responses given latent predictors. Finally, we develop a semiparametric Bayesian model for density regression in cases with many predictors. To reduce dimensionality of data, our model is based on factor analysis models with the number of latent variables unknown. A nonparametric prior for infinite factors is defined.

ACKNOWLEDGMENTS

This thesis is the end of the first act of a play entitled "Ju-Hyun, Professional Biostatistician." I don't think that it would have been possible for me to have finished the first act without help and support from many people.

First, I am very grateful to my thesis advisor, Dr. David Dunson. In addition to his insightful comments and advice, he gave me the encouragement and confidence to get through all the hardships that I encountered in my research. In Korea it is believed that a person has three major life-changing opportunities in his/her life. I believe that this was one of my three opportunities to have Dr. Dunson as my thesis advisor.

I thank Dr. Amy Herring who encouraged me to pursue to the Ph.D program in Biostatistics at UNC-CH. Whenever I was depressed or had troubles in my life, she was my personal counselor. Her warm words of wisdom and encouragement rescued me out of the swamp of emotional instability.

I thank my committee members, Dr. Michael Kosorok, Dr. Hongtu Zhu, and Dr. Kenneth Bollen, for their support. Their valuable comments made my thesis more fruitful. I also thank the faculty and staff in Biostatistics for their kind assistance and guidance.

I am grateful to the Collaborative Studies Coordinating Center, UNC-CH, and the Biostatistics Branch in the National Institute of Environmental Health Sciences for the experience and financial support they provided me.

Finally, I thank my entire extended family for their support. Although they live in Korea, I feel as though they are here beside me. Their warm words in emails and calls gave me strength to carry on my studies. I dedicate my thesis to my parents, Youngcho Park and Jungnim Shin.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

1.1 Introduction

Mixture models are widely used in many application areas, with finite mixtures of Gaussian distributions applied routinely in clustering (Day, 1969; Binder, 1978; Symons, 1981) and density estimation (Roeder and Wasserman, 1997; Richardson and Green, 1997). A recent review of the use of mixture models in clustering and density estimation can be found in Fraley and Raftery (2002).

In recent years, there has been an increasing need for flexible models for predictor-dependent clustering and conditional density estimation. For example, in microarray analysis for disease diagnosis in clinical research, an interest focuses on identifying differentially expressed genes accounting for experimental design information, such as disease status and other interesting covariates. The scientific interest is extended to group genes that have similarities in the expression levels and the relationship to the covariates (Qin and Self, 2006). Another example of this need has applications to epidemiology, where response to treatment differs due to unmeasured risk factors, causing the distribution of a health outcome to vary with dose of a drug or chemical (Dunson, 2007).

Motivated by such research problems and the successful use of mixture models characterizing univariate and multivariate distributions, several attempts have been made to generalize the mixture models to incorporate predictors. A finite mixture of linear regressions has been the most widely used, with the probability weights assigned to components being either fixed (Viele and Tong, 2002) or modeled by a parametric regression model, commonly polytomous logistic regression. Hierarchical mixtures-of-experts models (Jordan and Jacob, 1994) in the machine learning literature instead use a probabilistic decision tree to model the probability weights. An alternative is a mixture of multivariate normals, which induces a conditional response distribution given predictors as a locally weighted mixture of normal regression models, as in Müller et al. (1996).

There is a potential criticism on limited flexibility of a finite mixture of linear regression models due to its formulation. In finite mixture models, some a priori knowledge on the number of components is necessary and subjects are assigned to one of the preselected number of components. In this sense, the semiparametric Bayesian approach has been considered to provide a flexible mixture model that incorporates prior information but is free of the restriction on the number of components. From the Bayesian formulation, infinite mixture models can be obtained by assuming the mixture distribution to be generated from a stick-breaking prior, commonly the Dirichlet process (DP) (Ferguson, 1973; 1974) prior. With a remarkable advance in developing Monte Carlo Markov chain (MCMC) methods, rich literature has been contributed to the Dirichlet process mixture (DPM) models (Lo, 1984; Escobar, 1994; Escobar and West, 1995).

Despite its flexibility in incorporating infinite number of linear regression models, a DPM of linear regressions itself is not appropriate for predictor-dependent clustering and conditional density estimation. Under the model, the probability weights are fixed and constant over the predictor space, implying that subjects are interchangeable and predictors are not informative about the clustering. The fixed probability weights also restrict a conditional response mean to be linear.

In the nonparametric Bayesian literature, several methods have been proposed to introduce predictor-dependence in the probability weights. Griffin and Steel (2006) defined an order-based dependent Dirichlet process, where the ordering of beta variates in the stick-breaking construction depends on predictors. As an alternative, Dunson et al. (2007) proposed a kernel-weighted mixture of DPs (WMDP), where the weights are proportional to the product of spatial closeness in the predictor space measured through a kernel and random weights assigned at the predictor values observed in the sample. Dunson and Park (2008) defined the kernel-stick breaking process (KSBP) by introducing independent random probability measures and beta-distributed random weights at each of infinite sequence of random locations. The probability weights are sequentially allocated by randomly breaking a probability stick (starting with a stick of length 1) and allocating the probability broken off to a basis location, with the length of each break being proportional to the product of a kernel and the assigned random weight. In addition, the latter two approaches have a desirable sparsity-favoring structure in which the introduction of additional mixture components is automatically penalized and the base parametric model is used to interpolate across sparse data regions.

Although the WMDP and KSBP are very flexible and can be implemented with a straightforward Markov chain Monte Carlo (MCMC) algorithm, a potential criticism is expensive computation in posterior sampling. There are many unknown parameters needed in defining such priors, including smoothing parameters in a kernel. The number of smoothing parameters increases as there are more predictors, resulting in the computational burden. In Dirichlet process mixture models, a common strategy to reduce expensive computation is to rely on a marginalized model that integrates out the infinitely-many parameters characterizing the process to induce a finite parameter sampling distribution through a random partition of subjects into clusters (Quintana and Iglesias, 2003; Quintana, 2006). A similar approach has yet to be developed that allows predictor-dependence in partitioning.

With a focus on the applications of predictor-dependent clustering and conditional density estimation, this dissertation proposes three semiparametric Bayesian methods, which greatly reduce computational burden while facilitating simpler interpretations through the use of predictor-dependent partition models induced through marginalizing joint nonparametric process models. The first method focuses on generalizing the product partition model (Hartigan, 1990; Barry and Hartigan, 1992) to incorporate predictors in its clustering process. The underlying idea is based on marginalizing out the parameters characterizing the predictor-component of the joint modeling approach of Müller et al. (1996). This allows us to obtain a generalized Pólya urn scheme, which incorporates predictor-dependent weights in a simple and intuitive manner. The development of this urn scheme is the main theoretical contribution of the thesis, with the remainder of the thesis focusing on using the result to obtain methods that allow latent variable distributions in hierarchical models to be unknown and predictor-dependent.

Latent variable models (LVMs) including factor analysis and structural equation models (Bollen, 1989; Sánchez et al., 2005) provide a flexible modeling framework for characterizing dependence in observed multivariate variables having a variety of measurement scales. LVMs are very widely used in the social sciences and increasingly in biomedical applications in which one is interested in relationships among latent variables or wishes to apply a flexible framework for modeling mixed scale data in a parsimonious manner. A concern in many applications is sensitivity to the assumptions of normality and linearity in describing the joint distributions of the latent variables. Using our nonparametric Bayes methods, we develop a flexible class of semiparametric Bayes models that avoid these assumptions, while favoring a sparse structure that allows one to collapse on the base parametric model when limited information is available in the data about the latent variable distributions. Motivated by the problem of uncertainty in the number of latent variables, we will also consider alternative latent variable specifications that avoid assuming that each subject has a fixed, finite number of latent traits. Instead, decomposing latent traits into their occurrences and scores, we allow each subject to have different traits, with only some of the traits shared across subjects. This paradigm, which is related to the Indian Buffet process (IBP) (Griffiths and Ghahramani, 2005; Ghahramani et al., 2007), is seemingly more realistic in many applications and also directly allows uncertainty in the number of factors.

1.2 Literature Review

1.2.1 Nonparametric Bayes

In statistical modeling, one of the commonly-faced problems is how to model an unknown probability distribution for response y. The easiest way to do this is to find a class of distributions in a parametric family, which usually provide us with ease of implementation and interpretation of a statistical model. In analyzing real data, however, it is quite common to have a belief that observed data don't follow any known parametric distribution. This leads to modeling with an inadequate parametric assumption, often resulting in unreasonable inference. For this reason research interest has been taken to nonparametric methods as a way of getting very flexible models, typically defined by removing the parametric assumption.

Much of work on nonparametric inference has been achieved in the frequentist perspective. However, there are some attractive advantages of the Bayes formulation. First, it provides a full probabilistic characterization of the problem, which automatically allows for estimation uncertainty. It also provides a natural framework for inclusion of prior information allowing for shrinkage and centering on parametric models, limiting the curse of dimensionality. Finally, it allows embedding in larger hierarchical models, so that one can easily account for complicating features of the analysis, such as missing data, censoring, and model uncertainty. In this respect, nonparametric Bayesian inference is accomplished by defining probability models with infinite-dimensional parameters (Bernardo and Smith, 1994; Müller and Quintana, 2004). A collection of nonparametric Bayesian papers can be found in Ghosh and Ramamoorthi (2003). For a recent review on nonparametric inferences, refer to Müller and Quintana (2004).

1.2.2 Random Probability Measure

In the nonparametric Bayesian formulation, one can allow an unknown distribution by defining a probability measure on a collection of distribution functions. More formally, a prior can be induced on a distribution by defining a random probability measure (RPM). According to

Ferguson (1973) and Antoniak (1974), there are two desirable properties of RPMs: (I) the prior distribution should have large support; (II) given a sample of observations, posterior distribution should be analytically manageable. In addition to the desirable properties mentioned in these earlier papers, there are a number of other properties considered in more recent work, such as posterior consistency (Ghosal et al., 1999; Lijoi et al., 2005; Walker et al., 2007) and existence of Bernstein von Mises theorems (Freedman, 1999).

1.2.2.1 Dirichlet Process

The Dirichlet process (DP) (Ferguson, 1973; 1974) has been most popular and playing a key role in the nonparametric Bayesian literature. Let $\mathcal{Y}$ be a space and $\mathcal{A}$ a σ-field of subsets, and let F_0 be a finite non-null measure on $(\mathcal{Y}, \mathcal{A})$. Suppose that $\mathbf{y} = (y_1, \ldots, y_n)$ follow the following model:

$$y_i \overset{iid}{\sim} F, \quad F \sim DP(\alpha F_0), \tag{1.1}$$

where $DP(\alpha F_0)$ denotes a DP with precision α and base measure F_0. RPM F is said to follow a DP prior, $DP(\alpha F_0)$, if for any measurable partition $(A_1, \ldots, A_k)$ of $\mathcal{Y}$, the vector of $(F(A_1), \ldots, F(A_k))$ follows a Dirichlet distribution, $D(\alpha F_0(A_1), \ldots, \alpha F_0(A_k))$. For $A \in \mathcal{A}$, the DPP has the following properties:

1) $E(F(A)) = F_0(A)$ and $V(F(A)) = F_0(A)(1 - F_0(A))/(1 + \alpha)$

2) $F|\mathbf{y} \sim DP(\alpha^* F_0^*)$, where $\alpha^* = \alpha + n$ and $F_0^* = (F_0 + \sum_{i=1}^n \delta_{y_i})$

These properties imply that precision α controls the concentration on base measure F_0 with a large value expressing confidence that F_0 provides a good approximation, so that there is a high degree of shrinkage toward F_0 in the posterior, which has a conjugate form being also a DP.

One of the notable features of the DP is that the DP induces a partitioning among subjects according to their response values. From the DP prediction rule, also known as Blackwell and

MacQueen (1973) Pólya urn scheme, obtained upon marginalizing over the prior over F:

$$\begin{aligned} P(y_1 \in \cdot) &= F_0(\cdot), \\ P(y_i \in \cdot | y_1, \ldots, y_{i-1}) &= \left(\frac{\alpha}{\alpha + i - 1}\right) F_0(\cdot) + \sum_{j=1}^{i-1} \left(\frac{1}{\alpha + i - 1}\right) \delta_{y_j}(\cdot), \end{aligned} \tag{1.2}$$

it is obvious that the response y_i for subject i either takes a value newly generated from nonatomic base measure F_0 with probability $\alpha/(\alpha + i - 1)$ or is set equal to one of the existing values $(y_1, \ldots, y_{i-1})$ chosen by sampling from a discrete uniform. This induced clustering process is also known as the CRP (Chinese Restaurant Process) (Aldous, 1985), the name of which originates from a sequential seating arrangement in a Chinese restaurant: customers sequentially enter a restaurant which have an infinite number of tables capable of having an unlimited number of customers, the first customer is seated at an empty table, and the ith customer can be seated either at a new table with probability $\alpha/(\alpha + i - 1)$ or at one of the tables occupied by the first $i - 1$ customers with probability proportional to the number of customers seated at that table.

The DP can be alternatively constructed, based on Sethuraman's (1994) stick-breaking representation. RPM F in expression (1.1) can be equivalently expressed as

$$F = \sum_{h=1}^{\infty} \pi_h \delta_{y_h^*}, \quad \pi_h = V_h \prod_{l=1}^{h-1} (1 - V_l), \quad V_h \overset{iid}{\sim} \text{beta}(1, \alpha), \quad y_h^* \overset{iid}{\sim} F_0. \tag{1.3}$$

and from this expression, it follows that the DP is discrete with probability one.

1.2.2.2 Other RPMs

There have been many RPMs in the literature, which have a more general form than that of the DP, but we focus ourselves on a few most popular ones.

In expression (1.1), one can have a RPM associated with a species sampling model (SSM) (Pitman, 1996) by replacing the DPP with a random distribution, which has a functional form as

$$F = \sum_{h=1}^{\infty} w_h \delta_{y_h^*} + \left(1 - \sum_{h=1}^{\infty} w_h\right) F_0, \tag{1.4}$$

where atoms $\{y_h^*\}_{h=1}^{\infty}$ are sampled from base measure F_0 independent of weights $\{w_h\}_{h=1}^{\infty}$, with $P(\sum_{h=1}^{\infty} w_h \leq 1) = 1$ and w_h being interpreted as the relative frequency of the hth species with tag equal to y_h^* in a certain large population of various species. If $P(\sum_{h=1}^{\infty} w_h = 1) = 1$ and F_0 is nonatomic, then the distribution in expression (1.4) is discrete. SSMs are flexible and rich, including finite-dimensional Dirichlet-multinomial process (Muliere and Secchi, 1995), the DP, its two-parameter extension, the two-parameters Poisson-Dirichlet process (Pitman and Yor, 1997), and the beta two-parameter process (Ishwaran and Zarepour, 2000) as special cases.

Another important class of RPMs is stick-breaking priors (Muliere and Tardella, 1998; Ishwaran and James, 2001), which can be also seen as a special case of expression (1.4). According to their definition, a stick-breaking prior has a form of

$$F = \sum_{h=1}^{N} w_h \delta_{y_h^*},$$

where for $1 \leq N \leq \infty$, $\{y_h^*\}_{h=1}^{N}$ are independent and identically distributed draws from F_0, $\{w_h\}_{h=1}^{N}$ are probability weights with $w_h = V_h \prod_{l<h}(1 - V_l)$, where $V_h \overset{ind}{\sim} Beta(a_h, b_h)$. Some notable examples of stick-breaking priors are the two-parameters Poisson-Dirichlet process (with parameters $a_h = 1 - a$ and $b_h = b + ka$), the beta two-parameter process (with parameters $a_h = a$ and $b_h = b$), and the DP (with parameters $a_h = 1$ and $b_h = \alpha$) for $N \to \infty$.

The last, but not least, RPM we consider is the Pólya tree (PT) (Lavine, 1992; 1994), which is a generalization of the DP. The PT is defined by a set $\Pi = \{\pi_l, l = 1, 2, \dots\}$ of nested binary partitions of the sample space $\mathcal{Y}$. The PT is initiated by splitting the space $\mathcal{Y}$ into two disjoint subsets $\pi_1 = \{B_0, B_1\}$ and continues to partition the subsets, with the partition at level l being represented as $\pi_l = \{B_\epsilon,\ \epsilon = \epsilon_1 \dots \epsilon_l\}$, where ϵ is a binary string of length l with $\epsilon_j \in \{0, 1\}$. It is said that a RPM F follows a PT prior, denoted by $PT(\Pi, \mathcal{C})$, if there is a sequence of nonnegative numbers $\mathcal{C} = \{c_\epsilon\}$ and random variables $\mathcal{Z} = \{Z_\epsilon\}$ with $Z_\epsilon \overset{ind}{\sim} Beta(c_{\epsilon 0}, c_{\epsilon 1})$ and for every $l = 1, 2, \ldots$ and every $\epsilon = \epsilon_1 \ldots \epsilon_l$,

$$F(B_{\epsilon_1 \ldots \epsilon_l}) = \Big(\prod_{j=1; \epsilon_j = 0}^{l} Z_{\epsilon_1 \ldots \epsilon_{j-1}} \Big) \Big(\prod_{j=1; \epsilon_j = 1}^{l} (1 - Z_{\epsilon_1 \ldots \epsilon_{j-1}}) \Big).$$

The PT has the following properties: 1) Different from the other RPMs we considered in this subsection, continuous distributions can be generated from the PT with a certain choice of parameters, such as $c_{\epsilon_1 \ldots \epsilon_l} = l^2$; 2) The PT is a conjugate prior, $F|\mathbf{y} \sim PT(\Pi, \mathcal{C}^*)$, where $\mathcal{C}^* = c_\epsilon + n_\epsilon$ and n_ϵ is the number of observations in B_ϵ (Ferguson, 1974; Lavine, 1994) The DP is a special case of the PT with $c_\epsilon = c_{\epsilon 0} + c_{\epsilon 1}$.

1.2.3 Mixture models

1.2.3.1 Finite mixture models

Suppose that the response Y_i follows one of k (usually $< \infty$) group-specific densities $f_h(\cdot) = f_h(\theta_h)$, characterized by finite dimensional parameter $\theta_h \in \Psi$ (often $\Psi = R^d$) in a parametric family, with probability p_h, and then a (parametric) mixture model for Y_i is defined as

$$p(y_i) = \sum_{h=1}^{k} p_h f_h(y_i | \theta_h). \tag{1.5}$$

By introducing an unobservable (latent) variable S_i, with $S_i = h$ denoting that subject i belongs to the hth mixture component, expression (1.5) can be equivalently expressed in hierarchial form as

$$\begin{aligned} Y_i | S_i &\sim f(\theta_{S_i}) \\ S_i &\sim \text{Multinomial}(\{1, \ldots, k\}; \mathbf{p}), \end{aligned} \tag{1.6}$$

where $\mathbf{p} = (p_1, \ldots, p_k)$, and thus $\mathbf{S} = (S_1, \ldots, S_n)$ completely determines a partition of n subjects into $k \leq n$ clusters. With respect to the interpretation of these models, the former can be viewed as a semiparametric construction as an alternative to nonparametric models, whereas the latter is the missing data formulation (Jasra et al., 2005). For a recent review of the use of finite mixture models in various applications, refer to Fraley and Raftery (2002).

In order to generalize (1.5) to incorporate predictors $\mathbf{x} = (x_1, \ldots, x_p)$, one can model predictor dependence in $\pi = (\pi_1, \ldots, \pi_k)'$ and/or $f(\theta_h)$, $h = 1, \ldots, k$, as follows:

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

For example, hierarchical mixtures-of-experts models (Jordan and Jacobs, 1994) characterize $\pi_h(\mathbf{x})$ using a probabilistic decision tree, while letting $f_(y \,|\, \theta_h) = N(y; \mathbf{x}'\boldsymbol{\beta}_h, \tau_h^{-1})$ with $\theta_h = (\boldsymbol{\beta}_h', \tau_h)'$ correspond to the conditional density for a normal linear model. The term "expert" corresponds to the choice of $f(y \,|\, \theta_h, \mathbf{x})$, as different experts in a field may have different parametric models for the conditional distribution. A number of authors have considered alternative choices of regression models for the weights and experts (e.g., Jiang and Tanner, 1999). For recent articles, refer to Carvalho and Tanner (2005), Ge and Jiang (2006), and Geweke and Keane (2007).

1.2.3.2 Mixture Models with RPMs

Due to the discrete feature of the RPMs in section 1.2.2, they are not suitable for use as a prior on continuous densities. To avoid this constraint, discrete RPMs are often used as a mixing distribution in the mixture model framework, and the model can be expressed in hierarchical form as

$$\begin{aligned} y_i|\phi_i &\overset{ind}{\sim} f(\phi_i) \\ \phi_i &\overset{iid}{\sim} G \\ G &\sim RPM, \end{aligned} \tag{1.7}$$

often resulting in an infinite mixture model. For example, using Sethuraman's (1994) stick-breaking representation of the DP in (1.3), the mixture model with a choice of $G \sim DP(\alpha G_0)$ in expression (1.7), can be expressed as

$$p(y_i) = \sum_{h=1}^{\infty} \pi_h f_h(y_i|\theta_h), \tag{1.8}$$

where π_h and V_h are the same as in expression (1.3), but $\{\theta_h\}_{h=1}^{\infty}$ are an iid sample from base measure G_0, and this model is often referred to as the Dirichlet process mixture (DPM) models (Lo, 1984; Escobar and West, 1995). For mixture models with other RPMs, refer to the following papers: Ishwaran and Jaems (2003) and Navarrete et al. (2008) (for the RPM associated with the SSM), Ishwaran and James (2001) (for the stick-breaking prior), and Hanson and Johnson (2002), Paddock et al. (2003), and Hanson (2006) (for the PT).

With a focuss on the DPM models, there have been rich contributions to developing algorithms for posterior computation, and most of algorithms follows one of the three main approaches: the marginal approach, the conditional approach, and the split-merge approach. In order to avoid the need for expensive computation for the infinite-dimensional G, the marginal approach is to marginalize over the Dirichlet process, resulting in the Pólya urn scheme (Blackwell and MacQueen, 1973), which plays a key role in the Gibbs sampling methods (Escobar, 1994). Based on expression (1.2), the Gibbs sampling proceeds by sequentially sampling ϕ_i from its full conditional distribution

$$(\phi_i \,|\, \boldsymbol{\phi}^{(i)}, \mathbf{y}, \alpha) \;\; \sim \;\; q_{i0} G_{i,0} + \sum_{h=1}^{k^{(i)}} q_{ih} \delta_{\theta_h^{(i)}}, \tag{1.9}$$

where $\boldsymbol{\theta}^{(i)} = (\theta_1^{(i)}, \ldots, \theta_{k^{(i)}}^{(i)})$ is the $k^{(i)}$ unique values of $\boldsymbol{\phi}^{(i)} = (\phi_1, \ldots, \phi_{i-1}, \phi_{i+1}, \ldots, \phi_n)$, the posterior $G_{i,0}(\phi)$ is

$$G_{i,0}(\phi) = \frac{G_0(\phi) f(y_i|\phi)}{\int f(y_i|\phi) dG_0(\phi)} = \frac{G_0(\phi) f(y_i|\phi)}{h_i(y_i)},$$

$q_{i0} = c\, w_{i0}\, h_i(y_i)$, $q_{ih} = c\, w_{ih}\, f(y_i|\theta_h)$, and c is a normalizing constant. Due to a possibility of slow mixing, Bush and MacEachern (1996) modified the Gibbs sampling algorithm by updating the cluster specific parameters $\boldsymbol{\theta}$ separately from the cluster membership indicators $\mathbf{S} = (S_1, \ldots, S_n)$, with $S_i = h$ if $\phi_i = \theta_h$. Jain and Neal (2004) mentioned that in a case where mixture components are similar in terms of parameter values, the Gibbs sampler can become trapped in local modes, resulting in inefficient sampling and an inappropriate clustering. The authors suggested a split-merge MCMC method based on a Metropolis-Hastings procedure as a remedy for such problem.

One disadvantage of using the Pólya urn Gibbs sampler is that one cannot directly sample from the posterior of the DP, which is a motivation of the conditional approach. Ishwaran and Zarepour (2000) proposed a MCMC algorithm for a truncation approximation to the DP in the DPM model in (1.8), and the truncation approximation is alleviated by Papaspiliopoulos and Roberts's (2007) retrospective algorithm and Walker's (2007) slice sampling approach. Dunson and Park (2008) used both marginal and conditional approaches to posterior computation for kernel stick-breaking process models.

Parallel to the attempt to incorporate predictors within the finite mixture model framework, there has been considerable recent interest in the Bayesian nonparametric literature on developing priors for predictor-dependent collections of random probability measures. Starting with the Sethuraman (1994) stick-breaking representation of the DP, MacEachern (1999; 2001) proposed a class of dependent DP (DDP) priors, which is defined with common fixed weights π_h, but with atoms θ_h varying with predictors $\mathbf{x}$ in a stochastic process. DDP priors have been successfully implemented in ANOVA modeling (De Iorio et al., 2004), spatial data analysis (Gelfand et al., 2005), time series (Caron et al., 2006) and stochastic ordering (Dunson and Peddada, 2008) applications. Noting limited flexibility due to the fixed weight assumption, Griffin and Steel (2006) proposed an order-based DDP, where the ordering of the beta variates in the stick-breaking construction depends on $\mathbf{x}$.

1.2.4 Applications of Mixture Models

1.2.4.1 Density Regression

Conditional density estimation has had an increasing attention in the frequentist literature with an attempt to provide information on the relationship between a response Y and predictors $\mathbf{X} = (X_1, \ldots, X_n)$. Since pioneered by Rosenblatt (1969), kernel-based estimation methods with a basis on iid observations have played a key role in nonparametric conditional density estimation, having a similar functional form to expression (1.5). Hyndman et al. (1996) modified Rasenblatt's estimator using Nadaraya-Watson kernel regression, showing the properties of their estimator. An alternative approach was proposed by Fan et al. (1996), where local polynomial regression is instead used to generalize the Rasenblatt's estimator, later further improved by Hyndman and Yao (1998). Hall et al. (1999) proposed two improved methods, one based on a local logistic model and the other modifying the Nadaraya-Watson estimator. Bashtannyk and Hyndman (2001) and Fan and Yim (2004) handled the bandwidth selection problem in kernel conditional density estimation. For a summary of current methods on error criteria, kernel functions, and a bandwidth selection in general kernel density estimation, refer to Ahmad and Ran (2004) and references therein.

As opposed to rich literature on conditional density estimation, the Bayesian literature on the topic of conditional density estimation, referred to as density regression (Dunson, 2007; Dunson et al., 2007), is sparse. In the Bayesian literature, most of attentions have been focussed on density estimation (Ferguson, 1973; Lo, 1984; West, 1992; Escobar and West, 1995; Roeder and Wasserman, 1997; Richardson and Green, 1997; Ker and Ergün, 2005). A difficulty in density regression arises from the need of defining a prior for a collection of dependent random probability measures, referred to as a RPM field (RPMF). Starting from a DPM of normals for the joint distribution of Y and $\mathbf{X}$, Müller et al. (1996) expressed the conditional distribution of Y given $\mathbf{X}$ as a locally weighted mixture of normal regression models. Recently, Dunson et al. (2007) proposed a kernel-weighted mixture of independent DPs (WMDP) by placing a DP at each sampled predictor values, which resulted in a nonparametric mixture of regression models for the conditional distribution of Y given $\mathbf{X}$, with the mixture distribution varying with predictors. Motivated by a generalized urn scheme implied by the WMDP, Dunson (2007) dealt with the density regression problem in the empirical Bayesian approach, where hyperparameters were estimated by generalized maximum likelihood estimation. Noting a limited flexibility due to sample dependency of the WMDP prior, Dunson and Park (2008) proposed the kernel-stick breaking processes (KSBP), which is conceptually similar to WMDP, but differs in that independent RPMs and beta-distributed random weights are assigned to each of infinite sequence of random locations and the stick-breaking probability weights are expressed as a multiplication of a kernel by the beta weights.

With respect to density regression, the posterior consistency, asymptotic behavior of estimated densities to the true density, is a unrevealed research area, whereas Ghosal et al. (1999), Lijoi et al. (2005), and Walker et al. (2007) considered the property in density estimation. However, the results of Rodrigues et al. (2007) suggest that using the Müller et al. (1996) approach to induce a model for the conditional distributions does result in consistent estimates under some regularity conditions.

1.2.4.2 Clustering

As opposed to rich literature on conditional density estimation in both the frequentist and the Bayesian perspective, there is not much literature contributed to predictor-dependent clustering. Most of recent work focuses on clustering without predictors being involved.

Clustering based on mixture model in (1.5) is also known as Model-based clustering (MBC) (Murtagh and Raftery, 1984; Banfield and Raftery, 1993). A mixture model with $f_h(y_i|\theta_h)$ being multivariate Gaussian with a mean vector μ_h and a covariance matrix Σ_h has been successfully used in a broad variety of application areas (Murtagh and Raftery, 1984; Banfield and Raftery, 1993; Dasgupta and Raftery, 1998; Campbell et al., 1997; Celeux and Govaert, 1995; Mukerjee et al., 1998; Yeung et al., 2001). Banfield and Raftery (1993) further improved this model by parameterizing the covariance matrix Σ_h, determining the shape, volume, and orientation of the clusters, through eigenvalue decomposition, and showed that this general approach includes earlier methods based on Gaussian mixtures as special cases, such as the sum of squares criterion, as known as a heuristic (Ward, 1963), Friedman and Rubin (1967), Scott and Symons (1971), and Murtagh and Raftery (1984). For the detailed discussion concerning a class of models within this method, refer to Celeux and Govaert (1995).

With respect to implementing mixture models for clustering, there are two strategies: an agglomerative hierarchical approach based on the classification likelihood (Murtagh and Raftery, 1984; Banfield and Raftery, 1993) and an iterative relocation approach through the expectation-maximization (EM) algorithm for maximum likelihood estimation (methods based on likelihood ratio criteria, ; Celeux and Govaert, 1995). Dasgupta and Raftery (1998) and Fraley and Raftery (1998) showed good performance of the model through the EM algorithm, together with the Bayesian Information Criterion (BIC) approximation to determine the number of clusters. Refer to Fraley and Raftery (2002) for a recent review of the use of finite mixture models in clustering. For some other model selection criteria for mixture models, refer to Spiegelhalter et al. (2002) and Naik et al. (2007).

As an alternative model-based approach, Hartigan (1990) and Barry and Hartigan (1992) proposed the product partition models (PPMs), where the prior probability of the partition of $\{1, \ldots, n\}$ formed by $\mathbf{S}$ in expression (1.6) has a product form and the observations in group h are sampled from a common density $f_h(\theta_h)$, independently of other observations in different groups. The authors designated and successfully applied the model to address change point problems in time series and Crowley (1997) considered the model to obtain the product estimates of normal means. Recently Quintana and Iglesias (2003) showed that the DP is a special case of a PPM by recognizing that the DP induced marginal prior distribution on partitions is in the form of PPMs, and this relationship is further generalized by Quintana (2006) in that PPMs and the species sampling models (SSMs) (Pitman, 1996; Ishwaran and Jaems, 2003) induces the same partition probability model under exchangeability.

Using nonparametric Bayesian methods, clustering is formed by a prediction rule, which is obtained upon marginalizing over a random probability measure. As mentioned in Section 1.2.2.1, the DP induces clustering through Blackwell and MacQueen (1973) Pólya urn scheme.

Dunson et al. (2007) and Dunson and Park (2008) obtained a predictor dependent prediction rule for the WMDP and the KSBP, respectively. However, these urn schemes are not immediately useful for posterior computation.

Clustering and estimation for component-specific parameters based on the MCMC simulated samples is sensitive to the component labeling, on which density regression does not depend. This problem is known as the ”label switching” problem (Redner and Walker, 1984), caused by the symmetric prior on the parameters a of mixture model, which leads to the symmetric posterior distribution invariant to relabeling of the parameters. A common solution to this problem is artificial identifiability constraints (ICs) (Diebolt and Robert, 1994). Noting that in certain cases, ICs fail to remove the symmetry in the posterior distribution, Stephens(1997, 2000) and Celeux (1998) proposed “relabeling algorithms” to minimize the posterior expectation of some loss function. Celeux et al. (2000) dealt with label switching in the decision theoretic perspective. For a recent review on these algorithms, refer to Jasra et al. (2005).

1.3 Overview of Research

The focus of this dissertation is on developing flexible models for density regression and predictor-dependent clusterings. In each chapter, we handle different types of data structure, which researchers can frequently encounter. The methods described in Chapters 2 and 3 are applied to the data from the Collaborative Perinatal Project, where scientific interest lies in the effects of DDT (Dichlorodiphenyltrichloroethane) on health outcomes of children. Chapter 2 is a self-contained article proposing a class of nonparametric clustering models that incorporates predictors and illustrating it for density regression and predictor-dependent clustering. Chapter 3 is also a self-contained article, where a flexible model is derived, based on a joint modeling strategy using nonparametric Bayes approaches in cases, in which predictors are not directly measured. Chapter 4 describes extension of factor analysis to allow number of factors to vary, with applications to dimensionality reduction in predictive modeling. Chapter 5 discuss possible extension of the proposed methodologies and areas of future research.

CHAPTER 2

BAYESIAN GENERALIZED PRODUCT PARTITION MODEL

2.1 Introduction

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

$$f(y) = \sum_{h=1}^{k} \pi_h \, f_h(y \,|\, \theta_h), \tag{2.1}$$

where k is the number of mixture components, π_h is the probability weight assigned to component h, and $f_h(\cdot \,|\, \theta_h)$ is a distribution in a parametric family characterized by the finite-dimensional θ_h, for $h = 1, \ldots, k$. For a review of the use of (2.1) in clustering and density estimation, refer to Fraley and Raftery (2002).

In order to generalize (2.1) to incorporate predictors $\mathbf{x}$, one can model predictor dependence in $\pi = (\pi_1, \ldots, \pi_k)'$ and/or $f_h(\theta_h)$, $h = 1, \ldots, k$, as follows:

$$f(y \,|\, \mathbf{x}) = \sum_{h=1}^{k} \pi_h(\mathbf{x})\, f_h(y \,|\, \mathbf{x}, \theta_h). \tag{2.2}$$

For example, hierarchical mixtures-of-experts models (Jordan and Jacob, 1994) characterize $\pi_h(\mathbf{x})$ using a probabilistic decision tree, while letting $f_h(y \,|\, \mathbf{x}, \theta_h) = N(y; \mathbf{x}'\beta_h, \tau_h^{-1})$ correspond to the conditional density for a normal linear model. The term "expert" corresponds to the choice of $f_h(y \,|\, \mathbf{x}, \theta_h)$, as different experts in a field may have different parametric models for the conditional distribution. A number of authors have considered alternative choices of regression models for the weights and experts (e.g., Jiang and Tanner, 1999). For recent articles, refer to Carvalho and Tanner (2005) and Ge and Jiang (2006).

In this article, our goal is to develop a flexible semiparametric Bayes framework for predictor-dependent clustering and conditional distribution modeling. Potentially, we could simply rely on (2.2), as predictor-dependent clustering will naturally arise through the allocation of subjects sampled from (2.2) to experts. However, a concern is the sensitivity to the choice of the number of experts, k. A common strategy is to fit mixture models having different numbers of components, with the AIC or BIC used to select the model with the best fit penalized for model complexity. Unfortunately, these criteria are not appropriate for mixture models and other hierarchical models in which the number of parameters is unclear. For this reason, there has been recent interest in defining new model selection criteria that are appropriate for mixture models. Some examples include the DIC (Spiegelhalter et al., 2002) and the MRC (Naik et al., 2007).

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

There is a rich Bayesian literature on infinite mixture models, which let $k \to \infty$ in expression (2.1). This is accomplished by letting $y_i \sim f(\phi_i)$, with $\phi_i \sim G$, where $G = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h}$, with $\pi = \{\pi_h\}_{h=1}^{\infty}$ an infinite sequence of probability weights, δ_θ a probability measure concentrated at θ, and $\theta = \{\theta_h\}_{h=1}^{\infty}$ an infinite sequence of atoms. A wide variety of priors have been proposed for G, with the most common choice being the Dirichlet process (DP) prior (Ferguson, 1973; 1974). When a DP prior is used for the mixture distribution, G, one obtains a DP mixture (DPM) model (Lo, 1984; Escobar and West, 1995).

In marginalizing out G, one induces a prior on the partition of subjects $\{1, \ldots, n\}$ into clusters, with the cluster-specific parameters consisting of independent draws from G_0, the base distribution in the DP. As noted by Quintana and Iglesias (2003), this induced prior is a type of product partition model (PPM) (Hartigan, 1990; Barry and Hartigan, 1992). When the focus is on clustering or generating a flexible partition model for prediction, as in Holmes et al. (2005), it is appealing to marginalize out G in order to simplify computation and interpretation. The DP induces a particular prior on the partition and one can develop alternative classes of PPMs by replacing the DP prior on G with an alternative choice. Quintana (2006) applied this strategy for species sampling models (SSMs) (Pitman, 1996; Ishwaran and Jaems, 2003), which are a very broad class of nonparametric priors that include the DP as a special case.

Our focus is on further generalizing PPMs to include predictor-dependence by starting with (2.2) in the $k = \infty$ case, and attempting to obtain a prior which results in a PPM upon marginalization. There has been considerable recent interest in the Bayesian nonparametric literature on developing priors for predictor-dependent collections of random probability measures. Starting with the Sethuraman (1994) stick-breaking representation of the DP, MacEachern (1999, 2001) proposed a class of dependent DP (DDP) priors. In the fixed π case, DDP priors have been successfully implemented in ANOVA modeling (De Iorio et al., 2004), spatial data analysis (Gelfand et al., 2005), time series (Caron et al., 2006) and stochastic ordering (Dunson and Peddada, 2008) applications. Unfortunately, the fixed π case does not allow predictor-dependent clustering, motivating articles on order-based DDPs (Griffin and Steel, 2006), weighted mixtures of DPs (Dunson et al., 2007) and kernel stick-breaking processes (Dunson and Park, 2008).

In order to avoid the need for computation of the very many parameters characterizing these nonparametric priors, we focus instead on obtaining a generalized product partition model (GPPM) through relying on a related specification to Müller et al. (1996). Section 2.2 reviews the PPM and its relationship with the DP. Section 2.3 induces predictor-dependence in the PPM through a carefully-specified joint DPM model. Section 2.4 describes a simple and efficient Gibbs sampler for posterior computation. Section 2.5 contains an application, and Section 2.6 discusses the results.

2.2 Product Partition Models and Dirichlet Process Mixtures

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

$$f(\mathbf{y}|\mathbf{S}^*) = \prod_{h=1}^{k} f_h(\mathbf{y}_h), \quad \pi(\mathbf{S}^*) = c_0 \prod_{h=1}^{k} c(\mathbf{S}_h^*), \tag{2.3}$$

where $f_h(\mathbf{y}_h) = \int \prod_{i \in \mathbf{S}_h^*} f(y_i \,|\, \theta_h) dG_0(\theta_h)$, $f(\cdot \,|\, \theta)$ is a likelihood characterized by θ, the elements of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$ are independently and identically distributed with prior G_0, $c(\mathbf{S}_h^*)$ is a non-negative cohesion, and c_0 is a normalizing constant. The posterior distribution of the partition $\mathbf{S}^*$ given $\mathbf{y}$ also has a PPM form, but with the posterior cohesion $c(\mathbf{S}_h^*) f_h(\mathbf{y}_h)$.

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

$$\begin{aligned} y_i \,|\, \boldsymbol{\theta}, \mathbf{S} &\overset{ind}{\sim} f(\theta_{S_i}), \\ S_i &\overset{iid}{\sim} \sum_{h=1}^{k} \pi_h \delta_h, \quad \theta_h \overset{iid}{\sim} G_0, \end{aligned} \tag{2.4}$$

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

As motivated by Quintana and Iglesias (2003), a convenient choice corresponds to the Dirichlet process prior, $G \sim DP(\alpha G_0)$, with α a precision parameter and G_0 a non-atomic base measure. By the Dirichlet process prediction rule (Blackwell and MacQueen, 1973), the conditional prior of ϕ_i given $\boldsymbol{\phi}^{(i)} = (\phi_1, \ldots, \phi_{i-1}, \phi_{i+1}, \ldots, \phi_n)'$ and marginalizing out G is

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

which generates new values from G_0 with probability $\alpha/(\alpha + n - 1)$ and otherwise sets ϕ_i equal to one of the existing values $\boldsymbol{\phi}^{(i)}$ chosen by sampling from a discrete uniform. Hence, the joint distribution of $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)'$ is obtained as

$$\pi(\boldsymbol{\phi}) = \prod_{i=1}^n \left\{ \frac{\alpha G_0(\phi_i) + \sum_{j<i} \delta_{\phi_j}(\phi_i)}{\alpha + i - 1} \right\}. \tag{2.6}$$

Let $k = n(\mathbf{S}^*)$ denote the number of partition sets, with $k_h = n(\mathbf{S}_h^*)$ the cardinality of $\mathbf{S}_h^*$. Letting $\boldsymbol{\phi}_h = \{\phi_i : i \in \mathbf{S}_h^*\}$, with $\boldsymbol{\phi}_{h,l}$ being the parameter for the lth subject, ordered by the ids, in cluster h, Quintana and Iglesias (2003) show that (2.6) is equivalent to

$$\begin{aligned} \pi(\boldsymbol{\phi}) &= \sum_{\mathbf{S}^* \in \mathcal{P}} \frac{1}{\prod_{l=1}^n (\alpha + l - 1)} \prod_{h=1}^k \alpha (k_h - 1)! G_0(\boldsymbol{\phi}_{h,1}) \prod_{j=2}^{k_h} \delta_{\boldsymbol{\phi}_{h,1}}(\boldsymbol{\phi}_{h,j}) \\ &= c_0 \sum_{\mathbf{S}^* \in \mathcal{P}} \prod_{h=1}^k c(\mathbf{S}_h^*) \pi_h(\boldsymbol{\phi}_h), \end{aligned} \tag{2.7}$$

where $\mathcal{P}$ is the set of all partitions of $\{1, \ldots, n\}$, $c_0 = \prod_{l=1}^n (\alpha + l - 1)^{-1}$, $c(\mathbf{S}_h^*) = \alpha(k_h - 1)!$, and $\pi_h(\boldsymbol{\phi}_h)$ is the prior on $\boldsymbol{\phi}_h$. The marginal likelihood of $\mathbf{y}$ is then obtained as

$$f(\mathbf{y}) = c_0 \sum_{\mathbf{S}^* \in \mathcal{P}} \prod_{h=1}^{k} c(\mathbf{S}_h^*) \int \prod_{i \in \mathbf{S}_h^*} f(y_i|\theta) dG_0(\theta), \tag{2.8}$$

which is a special case of the form implied by (2.3) corresponding to a PPM with cohesion $c(\mathbf{S}_h^*) = \alpha(n(\mathbf{S}_h^*) - 1)!$. This implies that simple and efficient Markov Chain Monte Carlo (MCMC) algorithms developed for DPMs can be used for posterior computation in PPMs. However, the class of PPMs induced by the DPM specification above assumes that the subjects are exchangeable, and does not allow for the incorporation of predictors.

2.3 Predictor Dependent Product Partition Models

2.3.1 Proposed formulation

Our goal is to incorporate predictor values $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ into a class of PPMs, so that the prior on the partition $\mathbf{S}^*$ has the form

$$\pi(\mathbf{S}^*|\mathbf{X}) \propto \prod_{h=1}^{k} c(\mathbf{S}_h^*, \mathbf{X}_h), \tag{2.9}$$

where $\mathbf{X}_h = \{\mathbf{x}_i : i \in \mathbf{S}_h^*\}$, for $h = 1, \ldots, k$, and the cohesion $c(\cdot)$ depends on the subjects predictor values. Expression (2.9) has two appealing properties. First, the posterior distribution of the partition $\mathbf{S}^*$ updated with the likelihood of response $\mathbf{y} = (y_1, \ldots, y_n)'$ is still in a class of PPMs, but with updated cohesion $c(\mathbf{S}_h^*, \mathbf{X}_h) f_h(\mathbf{y}_h)$. Secondly, there is a direct influence of predictors $\mathbf{X}$ on the partition process. Previous incorporation of predictors in PPMs instead relies on replacing $f(y_i \,|\, \theta_h)$ with $f(y_i \,|\, \mathbf{x}_i, \theta_h)$ in expression (2.3), which allows the predictor effect to vary across clusters but does not allow the clustering process itself to be predictor dependent.

To specify cohesion $c(\mathbf{S}_h^*, \mathbf{X}_h)$, we exploit the connection between PPM and DPMs. For simplicity of notation, we focus on univariate response y, though multivariate generalizations are straightforward. Suppose $\mathbf{z}_i = (y_i, \mathbf{x}_i')'$ follows the hierarchical model:

$$\begin{aligned} f(\mathbf{z}_i \,|\, \boldsymbol{\phi}_i) &= f(y_i, \mathbf{x}_i|\varphi_i, \gamma_i) = f_1(y_i|\mathbf{x}_i, \varphi_i) f_2(\mathbf{x}_i|\gamma_i), \\ \boldsymbol{\phi}_i &\sim G, \quad G \sim DP(\alpha G_0), \end{aligned} \tag{2.10}$$

where $G_0 = G_{0\varphi} \bigotimes G_{0\gamma}$ is the product measure of $G_{0\varphi}$ and $G_{0\gamma}$, components inducing a base prior for φ_i and γ_i, respectively. This DPM model will induce partitioning of the subjects $\{1, \ldots, n\}$ into $k \leq n$ clusters, with $i \in \mathbf{S}_h^*$ denoting that subject i belongs to cluster h, which implies that $\varphi_i = \varphi_h^*$ and $\gamma_i = \gamma_h^*$, where $\boldsymbol{\gamma}^* = (\gamma_1^*, \ldots, \gamma_k^*)'$ and $\boldsymbol{\varphi}^* = (\varphi_1^*, \ldots, \varphi_k^*)'$ denote the unique values of $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)'$ and $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_n)'$, respectively.

Under (2.10), we can obtain a joint distribution of $\boldsymbol{\phi} = (\boldsymbol{\varphi}, \boldsymbol{\gamma})$ using the same approach used in deriving expression (2.7). If we then multiply by the conditional likelihood $\prod_{i=1}^n f_2(\mathbf{x}_i|\gamma_i)$ and marginalize out $\boldsymbol{\gamma}$, the joint distribution of $\boldsymbol{\varphi}$ and $\mathbf{X}$ is given by

$$\pi(\boldsymbol{\varphi}, \mathbf{X}) = \sum_{\mathbf{S}^* \in \mathcal{P}} c_0 \prod_{h=1}^{k} \alpha(k_h - 1)! \left\{ \int \prod_{i \in \mathbf{S}_h^*} f_2(\mathbf{x}_i|\gamma_h^*) dG_{0\gamma}(\gamma_h^*) \right\} G_{0\varphi}(\boldsymbol{\varphi}_{h,1}) \prod_{j=2}^{k_h} \delta_{\boldsymbol{\varphi}_{h,1}}(\boldsymbol{\varphi}_{h,j}), \tag{2.11}$$

where $\boldsymbol{\varphi}_{h,l}$ is the parameter for the response y of the lth subject, ordered by the ids, in cluster h, and therefore the conditional distribution of $\boldsymbol{\varphi}$ given $\mathbf{X}$ is

$$\begin{aligned} \pi(\boldsymbol{\varphi}|\mathbf{X}) &= c_0^* \sum_{\mathbf{S}^* \in \mathcal{P}} \prod_{h=1}^{k} \alpha(k_h - 1)! \left\{ \int \prod_{i \in \mathbf{S}_h^*} f_2(x_i|\gamma_h^*) dG_{0\gamma}(\gamma_h^*) \right\} G_{0\varphi}(\boldsymbol{\varphi}_{h,1}) \prod_{j=2}^{k_h} \delta_{\boldsymbol{\varphi}_{h,1}}(\boldsymbol{\varphi}_{h,j}) \\ &= c_0^* \sum_{\mathbf{S}^* \in \mathcal{P}} \prod_{h=1}^{k} c(\mathbf{S}_h^*, \mathbf{X}_h) \pi_h(\boldsymbol{\varphi}_h), \end{aligned} \tag{2.12}$$

where c_0^* is a normalizing constant, so that the sum over $\mathcal{P}$ is unity, $c(\mathbf{S}_h^*, \mathbf{x}_h) = \alpha(k_h - 1)! \int \prod_{i \in \mathbf{S}_h^*} f_2(\mathbf{x}_i|\gamma) dG_{0\gamma}(\gamma)$, and $\pi_h(\boldsymbol{\varphi}_h)$ is a prior on partitioned set $\boldsymbol{\varphi}_h$. Hence, we have induced a generalized PPM (GPPM) of the form shown in (2.9) starting with a joint DPM model for the response and predictors related to that proposed by Müller et al. (1996). A related idea was independently developed by Fernando Quintana and collaborators in recent work (unpublished communication), though our subsequent development differs from theirs.

2.3.2 Generalized Pòlya Urn Scheme

It is not obvious from expression (2.12) how the predictor and hyperparameter values impact clustering. However, as shown in Theorem 1, we can show that the proposed GPPM induces a simple predictor-dependent generalization of the Blackwell and MacQueen (1973) Pólya urn scheme, which should be useful both in interpretation and posterior computation.

Theorem 2.3.1. *Let superscript (i) on any matrix or vector indicate that the contribution of subject i has been removed. The full conditional prior of φ_i given α, $\boldsymbol{\varphi}^{(i)}$, and $\mathbf{X}$, or equivalently given α, $\boldsymbol{\varphi}^{*(i)}$, $\mathbf{S}^{(i)}$, and $\mathbf{X}$, has the form*

$$\left(\varphi_i \,|\, \alpha, \boldsymbol{\varphi}^{*(i)}, \mathbf{S}^{(i)}, \mathbf{X}\right) \sim w_0(\mathbf{x}_i) G_{0\varphi} + \sum_{h=1}^{k^{(i)}} w_h(\{\mathbf{x}_i, \mathbf{X}_h^{(i)}\}) \delta_{\varphi_h^{*(i)}}, \tag{2.13}$$

with the probability weights

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

where c is a normalizing constant and $G_{0\gamma}^*(\cdot|\mathbf{X}_h^{(i)})$ is the posterior distribution updated with the likelihood of predictor cluster h excluding the contribution from the ith subject.

The proof is in Appendix A. Theorem 2.3.1 implies that subject i is assigned to either a new generated value (creating a new cluster) or one of the existing unique values, with the probability weights being proportional to a product of the DP probability weights and the marginal likelihoods at its predictor value varying across clusters. Therefore, subject i is more likely to be grouped into cluster h if the predictor value of subject i, $\mathbf{x}_i$, is close to those of other subjects in the hth cluster, $\mathbf{X}_h$, with the measure of closeness depending on the scale of the data through the choice of $f_2(\cdot)$.

Conceptually, this idea is related to the Bayesian partition model (BPM) of Holmes et al. (2005) in that subjects close together in the predictor space will tend to have similar response distributions. However, instead of measuring closeness through assuming a particular distance metric, our specification automatically induces a distance metric through a flexible nonparametric model for the joint distribution of the predictors. This allows the measure of closeness to be adaptive depending on location in the predictor space, automatically producing spatially-adaptive bandwidth selection. In the special case of a degenerate distribution for $\mathbf{x}$, $f_2(\mathbf{x}|\gamma) = \delta_\gamma(\mathbf{x})$, formulation (2.13) reduces to the Blackwell and MacQueen Pòlya urn scheme of expression (2.5).

An apparent disadvantage of our formulation is that by inducing a prior for the conditional distribution of y_i given $\mathbf{x}_i$ through a prior for the joint distribution of y_i and $\mathbf{x}_i$, we are implicitly assuming that the predictors are random variables. In fact, in many applications one or more of the predictors may be fixed by design, representing spatial location, time of observation or an experimental condition. The predictor-dependent urn scheme shown in Theorem 2.3.1 is still useful and coherent in such cases, as this urn scheme is defined conditionally on the predictor values. This urn scheme clearly results in a coherent joint prior for $\boldsymbol{\varphi}$ conditionally on $\mathbf{X}$, which is invariant to permutations in the ordering of the subjects. It is in general very difficult to define a predictor-dependent urn scheme, which satisfies these conditions.

The use of the conjugacy simplifies the weights in (2.13), resulting in a closed and simple form for computation. Among many choices, we focus on two special cases: a normal-Wishart prior and a Poisson-gamma prior. Suppose that a normal-Wishart distribution is assumed for continuous $p \times 1$ predictors $\mathbf{x}$ and parameter $\gamma = (\boldsymbol{\mu}_\mathbf{x}, \boldsymbol{\Sigma}_\mathbf{x})'$:

$$\begin{aligned} \mathbf{x}|\boldsymbol{\mu}_\mathbf{x}, c_\mathbf{x}, \boldsymbol{\Sigma}_\mathbf{x} &\sim N(\boldsymbol{\mu}_\mathbf{x}, c_\mathbf{x}^{-1}\boldsymbol{\Sigma}_\mathbf{x}), \\ \boldsymbol{\mu}_\mathbf{x}|\mu_\mathbf{x}, c_\mu, \boldsymbol{\Sigma}_{0\mathbf{x}} &\sim N(\boldsymbol{\mu}_{0\mathbf{x}}, c_\mu^{-1}\boldsymbol{\Sigma}_\mathbf{x}) \\ \boldsymbol{\Sigma}_\mathbf{x}^{-1}|\nu_\mathbf{x}, \boldsymbol{\Sigma}_{0\mathbf{x}} &\sim \mathcal{W}(\boldsymbol{\Sigma}_{0\mathbf{x}}^{-1}, \nu_\mathbf{x}), \end{aligned} \qquad (2.14)$$

where $c_\mathbf{x}^{-1}$ and c_μ^{-1} are multiplicative constants, and $\mathcal{W}(\boldsymbol{\Sigma}_{0\mathbf{x}}^{-1}, \nu_\mathbf{x})$ is a Wishart with degrees of freedom $\nu_\mathbf{x}$ and expectation $\nu_\mathbf{x}\boldsymbol{\Sigma}_{0\mathbf{x}}^{-1}$. Then the marginal likelihood of $\mathbf{x}_i$ in probability weight $w_0(\mathbf{x}_i)$ in (2.13) is a noncentral multivariate t-distribution with degrees of freedom $\nu = \nu_\mathbf{x} - p + 1$, mean $\boldsymbol{\mu} = \boldsymbol{\mu}_{0\mathbf{x}}$, and scale $\boldsymbol{\Sigma} = (c_{\mathbf{x}} + c_\mu)/(\nu c_{\mathbf{x}} c_\mu)\boldsymbol{\Sigma}_{0\mathbf{x}}$:

$$f(\mathbf{x}|\boldsymbol{\mu}, \nu, \boldsymbol{\Sigma}) = \frac{\Gamma((\nu+p)/2)}{(\pi\nu)^{p/2}\Gamma(\nu/2)|\boldsymbol{\Sigma}|^{1/2}}\left(1 + \frac{1}{\nu}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)^{-(\nu+p)/2}, \tag{2.15}$$

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

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

where $\bar{\mathbf{x}}_h^{(i)} = \sum_{j:S_j^{(i)}=h} \mathbf{x}_j / k_h^{(i)}$. Note that the structure in expression (2.14) is slightly different from a commonly used normal-Wishart prior in that a multiplicative constant is multiplied not only to the variance of the expectation of $\mathbf{x}$ but also to the variance of $\mathbf{x}$. The reasoning for this is to induce local clustering by making the distribution of $\mathbf{x}$ denser around its expected value, while the expected value can be drawn over the range of $\mathbf{x}$, with $c_{\mathbf{x}}^{-1}$ restricted to be in $(0, 1]$ and $c_\mu^{-1} = 1$. Allowing $c_{\mathbf{x}}$ to vary across clusters gives us additional flexibility.

In the case of discrete predictors, we can also obtain a closed form marginal likelihood of $\mathbf{x}$. In order to simplify calculations in the discrete case, we assume a priori independence for the different predictors. Suppose that x_j for $j = 1, \ldots, p$ follow a Poisson distribution with mean Γ_j, which is assigned a Gamma prior with mean a_j/b_j, $\mathcal{G}(a_j, b_j)$, as the base measure $G_{0\gamma}$. The marginal distribution of $\mathbf{x}$ in w_0 is a product of negative binomials with the number of successes $r_j = a_j$ and success probability $p_j = b_j/(1 + b_j)$:

$$Pr(X_j = k) = \frac{\Gamma(r_j + k)}{k!\Gamma(r_j)} p_j^{r_j}(1 - p_j)^k \qquad j = 1, \ldots, p. \tag{2.16}$$

The marginal distribution in w_h, for $h = 1, \ldots, k^{(i)}$, is also a product of negative binomials, but with hyperparameters $a_j^* = a_j + \sum_{j:S_j^{(i)}=h} x_j$ and $b_j^* = b_j + k_h^{(i)}$. For bounded discrete predictors, we can instead use a multinomial likelihood with a Dirichlet prior for the category probabilities. The case of mixed discrete and continuous predictors can also be dealt with easily.

2.4 Posterior Computation

One of the appealing features of our predictor-dependent urn scheme is that we can rely on efficient Pólya urn Gibbs sampling algorithms developed for computation in marginalized DPMs (Bush and MacEachern, 1996) with minimal modifications. In addition, although we focus here on posterior computation through MCMC, our predictor-dependent urn scheme could similarly be used to develop sequential importance sampling (SIS) algorithms (MacEachern et al., 1999; Quintana and Newton, 2000), modified weighted Chinese restaurant (WCR) sampling algorithms (Ishwaran and Jaems, 2003), as well as fast variational Bayes approximations (Kurihara et al., 2006).

Following Bush and MacEachern (1996), our algorithm updates the cluster specific parameters $\boldsymbol{\varphi}^*$ separately from the cluster membership indicators $\mathbf{S}$. From Theorem 2.3.1, the full conditional posterior distribution of φ_i can be derived as follows:

$$\big(\varphi_i \,|\, \alpha, \boldsymbol{\varphi}^{*(i)}, \mathbf{S}^{(i)}, \mathbf{X}, \mathbf{y}\big), \sim q_{i,0} G_{0\varphi,i} + \sum_{h=1}^{k^{(i)}} q_{i,h} \delta_{\varphi_h^{*(i)}}, \tag{2.17}$$

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

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

$q_{i,0} = cw_0(\mathbf{x}_i)h_i(y_i|\mathbf{x}_i)$, $q_{i,h} = cw_h(\{\mathbf{x}_i, \mathbf{X}^{(i)}\}) f_1(y_i|\mathbf{x}_i, \varphi_h^{*(i)})$, and c is a normalizing constant. Instead of sampling directly from expression (2.17) in implementing the Gibbs sampling, we first sample S_i, for $i = 1, \ldots, n$, from its multinomial conditional posterior distribution with:

$$\Pr(S_i = h|\boldsymbol{\varphi}^{*(i)}, \mathbf{S}^{(i)}, \mathbf{X}, \mathbf{y}) = q_{i,h}, \qquad h = 0, 1, \ldots, k^{(i)}, \tag{2.18}$$

and when $S_i = 0$, φ_i is set to a new value generated from $G_{0\varphi,i}$. As a result of updating $\mathbf{S}$, the number of clusters, k is automatically updated. As a next step, we update $\boldsymbol{\varphi}^*$ conditional on $\mathbf{S}$ and k from

$$\big(\varphi_h|\boldsymbol{\varphi}^{*(h)}, \mathbf{S}, k, \mathbf{y}, \mathbf{X}\big) \propto \Big\{ \prod_{i:S_i=h} f_1(y_i|\mathbf{x}_i, \varphi_h) \Big\} G_{0\varphi}(\varphi_h). \tag{2.19}$$

In a case that there are some unknown parameters ψ characterizing the base measure $G_{0\varphi}$, we include an additional step for updating ψ based on the full conditional posterior distribution

$$\big(\psi|\boldsymbol{\varphi}, \mathbf{y}, \mathbf{x}\big) \propto \pi(\psi)\Big\{ \prod_{h=1}^{k} G_{0\varphi}(\varphi_h^*|\psi) \Big\}. \tag{2.20}$$

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

2.5 Simulation Examples

2.5.1 Model specification

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

$$f(y_i|\mathbf{x}_i^*) = \sum_{h=1}^{\infty} \pi_h(\mathbf{x}_i^*)\, f_1(y_i \,|\, \mathbf{x}_i^*, \varphi_h^*), \tag{2.21}$$

where $\mathbf{x}_i^* = (1, \mathbf{x}_i')' = (1, x_{i1}, \ldots, x_{ip})'$ and $f_1(y_i|\, \mathbf{x}_i^*, \varphi_h^*) = N(y_i; \mu_h, \sigma_{y,h}^2)$ with $\varphi_h^* = (\mu_h, \sigma_{y,h}^2)'$ for the first simulation and $f_1(y_i|\, \mathbf{x}_i^*, \varphi_h^*) = N(y_i; \mathbf{x}_i^{*\prime}\boldsymbol{\beta}_h, \sigma_y^2)$ with $\varphi_h^* = (\boldsymbol{\beta}_h, \sigma_{y,h}^2)'$ for the second simulation. The GPPM proposed in Section 3 is used to place a prior on the partition $\mathbf{S}^*$ and atoms $\boldsymbol{\varphi}^*$. Although there are $k \leq n$ mixture components represented in the sample of n subjects under the GPPM, there are conceptually infinitely many components, since the number of components increases stochastically as subjects are added.

In the absence of prior knowledge about the scale, it is recommended that continuous predictors be standardized to simplify prior elicitation. We require G_0 to correspond to a proper distribution, since marginal likelihoods will be used in calculating conditional posterior probabilities for partitioning. To simplify updating of the scale parameter, $c_{\mathbf{x}}$, we assume a discrete uniform prior on $(0, 1]$. For discrete predictors, we fix $a_j = b_j = 1$, for $j = 1, \ldots, p-1$. In addition, let $\sigma_{y,h}^{-2} \sim \mathcal{G}(a_y, b_y)$, $\mu_h \sim N(\mu, \kappa^{-1}\sigma_{y,h}^2)$, $\boldsymbol{\beta}_h \sim N(\boldsymbol{\beta}, \sigma_{y,h}^2 \mathbf{V})$ with $\mathbf{V} = \kappa^{-1} n (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}$ and $X^* = (\mathbf{x}_1^*, \ldots, \mathbf{x}_n^*)'$, $\mu \sim N(\mu_0, \kappa^{-1}\sigma_\mu^2)$, $\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, \kappa^{-1}\mathbf{V}_0)$, and $\kappa \sim \mathcal{G}(a_\kappa, b_\kappa)$. The last three prior distributions on $\psi = (\mu, \kappa)'$ or $\psi = (\boldsymbol{\beta}, \kappa)'$ are for additional flexibility. In the implementation, we let $\alpha = 1$, $\boldsymbol{\mu}_{0\mathbf{x}} = \mathbf{0}$, $\boldsymbol{\Sigma}_{0\mathbf{x}}^{-1} = 4I_{p \times p}$, $\nu_{\mathbf{x}} = p$, $\mu_0 = 0$, $\boldsymbol{\beta}_0 = \mathbf{0}$, $\mathbf{V}_0 = n(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}$, and $a_y = b_y = a_\kappa = b_\kappa = 1$. Other choices of these parameters are also considered to check sensitivity of models to our primary choice.

2.5.2 Implementation and Results

We consider two cases in which $n = 500$, $p = 1$, and x_{i1} is generated from a uniform distribution over $(0, 1)$. We first simulated data from a normal distribution with mean x_{i1}^2 and variance 0.04, $N(y_i; x_{i1}^2, 0.04)$. The data were analyzed using a mixture of normals with the prior specification of Section 2.5.1, and with the MCMC algorithm of Section 2.4 implemented for 10,000 iterations, discarding the initial 1,000 iterations as a burn-in. Figure 2.1 shows selected results. The algorithm converged rapidly and mixing was good based on trace plots of μ, the number of clusters, and $f(y = 1.5|\mathbf{x}^* = (1\, 0.25)')$, where the data point for y is randomly selected among possible values (the left panel of Figure 2.1). As shown in the right panel of Figure 2.1, the predictive densities and mean function of y (solid lines) well approximate the true values (dotted lines), which are completely embedded within pointwise 99% credible intervals (dashed lines). The posterior mean of the number of clusters was 2.4 with a 95% credible interval of $[2, 4]$ and the estimated normal means were almost equally spaced over $(0, 1)$.

As a more challenging second simulation case, we simulated data to approximately mimic the data in the reproductive epidemiology study considered in Section 2.6. In particular, we

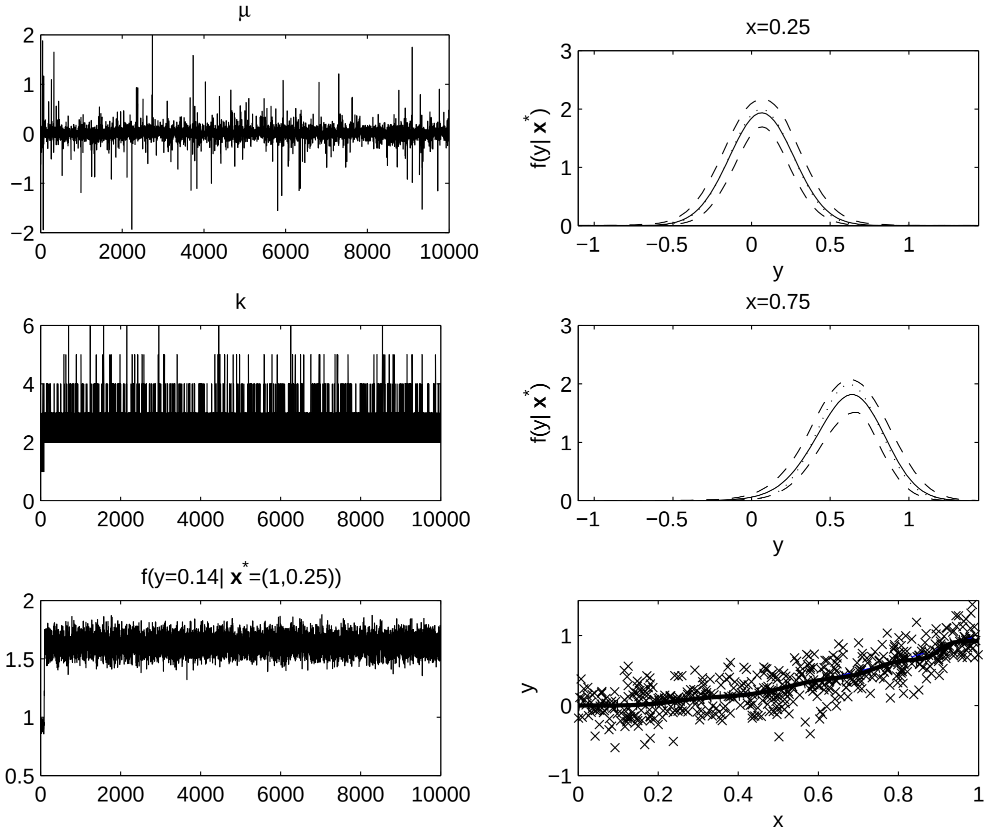

FIGURE 2.1: Results for the first simulation example. The left column provides trace plots for representative quantities, while the right panel shows the conditional distributions for two different values of x, as well as the mean function estimation along with the raw data. Posterior means are solid lines, pointwise 99% credible intervals are dashed lines, and true values are dotted lines.

generated data from the following mixture of two linear models:

$$f(y_i|\mathbf{x}_i) = (1 - x_{i1}^4)N(y_i; 1, 0.04) + x_{i1}^4 N(y_i; , 1 - x_{i1}^2, 0.01),$$

where a secondary peak appears in the left tail of the response distribution, moving closer to zero as x_{i1} increases. This behavior in which the tail of the distribution, corresponding to those subjects with the most extreme response, is particularly sensitive to changes in an exposure variable is common in toxicology and epidemiology studies. We analyzed the data using a mixture of regression models with the GPPM approach specified in Section 2.5.1, and also using

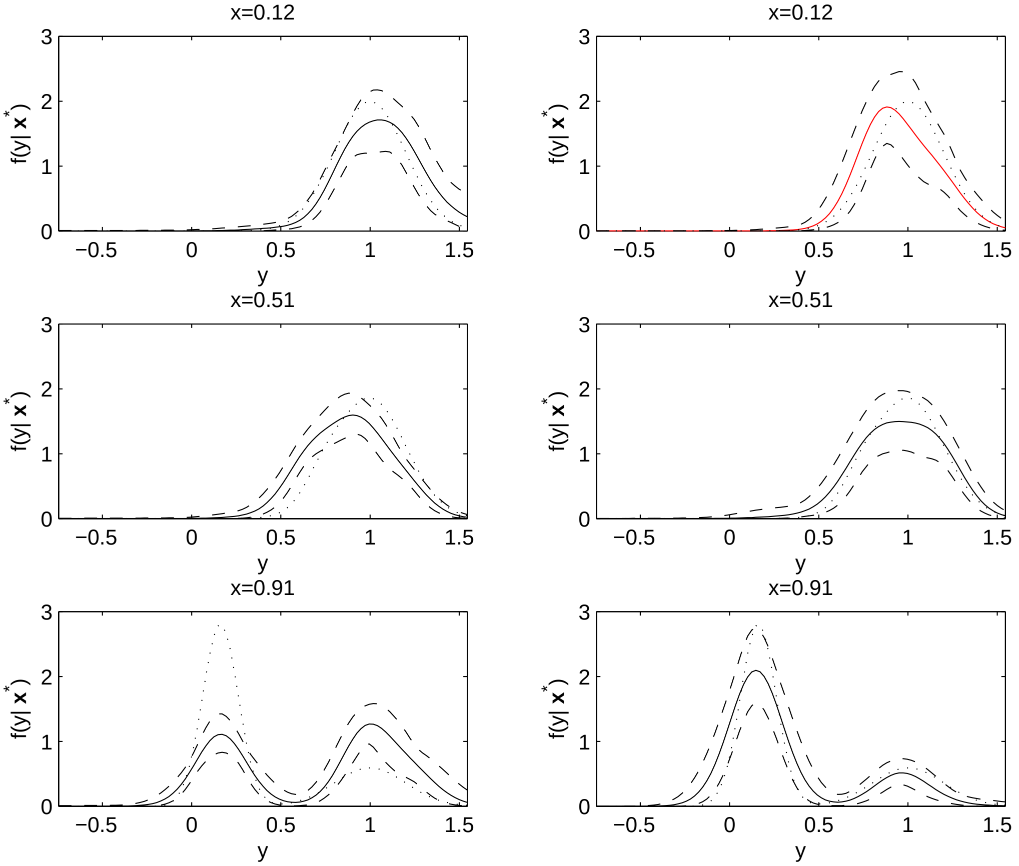

FIGURE 2.2: Estimated predictive densities from the PPM (left panel) and the GPPM (right panel) at the 10th, 50th and 90th percentiles of the empirical distribution of x: posterior means (solid lines), pointwise 99% credible intervals (dashed lines), and true values (dotted lines).

the DPM-based PPM described in Section 2.2. These two approaches result in mixtures of normal linear regressions, but the first approach allows the mixture weights to be predictor-dependent, while the second doesn't. The precision parameter α and base measure G_0 for the DPM-based PPM were set to be the same as those used in the GPPM approach. Both analyses were run for 30,000 iterations with a 10,000 iteration burn-in, with good mixing and convergence rates in both cases based on examination of trace plots and diagnostics.

From Figure 2.2, it is clear that the proposed approach provides a more flexible model capturing the rapid changes in the distribution across local regions of the predictor space even for the somewhat small sample size of $n = 500$. We also repeated the analysis of the second simulation including a discrete predictor, which was obtained by truncating the continuous predictor into l groups. It was observed that the proposed method worked well for a variety choices of l (results are not shown).

2.6 Epidemiologic Application

We apply the proposed method to the data used in Longnecker et al. (2001) and Dunson and Park (2008). DDT has been widely used and shown to be effective against malaria-transmitting mosquitoes, but several health-threatening effects of DDT have been also reported. Longnecker et al. (2001) used the data from the US Collaborative Perinatal Project to investigate the association between DDT and preterm birth, defined as delivery before 37 weeks of complete gestation. The authors showed that adjusted for other covariates, increasing concentrations of maternal serum DDE, a persistent metabolite of DDT, led to high rate of preterm birth by fitting a logistic regression model with categorized DDE levels. Dunson and Park (2008) applied a kernel stick-breaking process mixture of linear regression models to the same data with a focus on the predictive density of gestational age at delivery (GAD), concluding strong evidence of a steadily increasing left tail with DDE dose. For more information on the study design and data structure, refer to Longnecker et al. (2001).

We let x_{i1} and x_{i2} be the DDE dose for child i and the mother's age after normalization, respectively. There were $2,313$ children left in the study after removing children with GAD > 45 weeks, which are suspected as unrealistic values in reproductive epidemiology. By running the algorithm of the GPPM approach applied to the first simulation example for 30,000 iterations with a 10,000 iteration burn-in, we obtained the estimated predictive densities of GAD at selected percentiles $(10, 30, 70, 90)$ of the empirical distribution of DDE (Figure 2.3) with the maternal age being fixed at its mean. The shape and location of the estimated densities do not change much at different values of the maternal age. The results also show that the left tail of the distribution increases for high DDE dose with the credible intervals wider at high DDE values due to the relatively few observations in this region. It is observed in Figure 2.4 that

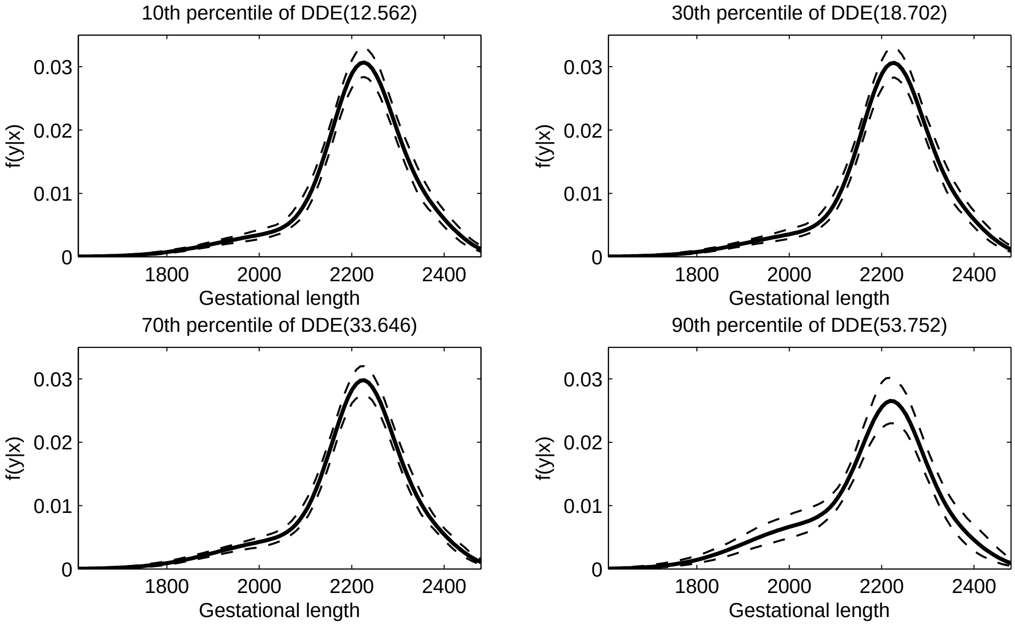

FIGURE 2.3: Estimated predictive densities (solid lines) for gestational age at delivery at preselected values of DDE with 99% pointwise credible intervals (dashed lines).

the conditional predictive mean of GAD had a slightly decreasing nonlinear trend over DDE level, while the maternal age was fixed at its mean. In using the GPPM for conditional density estimation and quantile regression estimation, the predictor-dependent partitioning is used as a tool for flexibly modeling of the conditional response distribution given the predictors through Theorem 2.3.1. However, in some cases, there may be interest in using the methodology for identifying clusters of subjects. Because the meaning of the clusters varies across the MCMC iterations, which is known as the label switching problem, there have been some contributions on post-processing approaches for clustering (Celeux et al., 2000; Stephens, 2000; Dahl, 2006; Lau and Green, 2007). We followed Lau and Green (2007) approach to estimate an optimal partition.

Figure 2.5 contains a symmetric heatmap presenting the pairwise marginal probabilities of being grouped with another subject in the given data. There were 13 clusters as a result of the obtained optimal partition, and some summary statistics within these clusters are arranged in Table 2.1. All the preterm births except one were grouped into 4 clusters. Most of the preterm

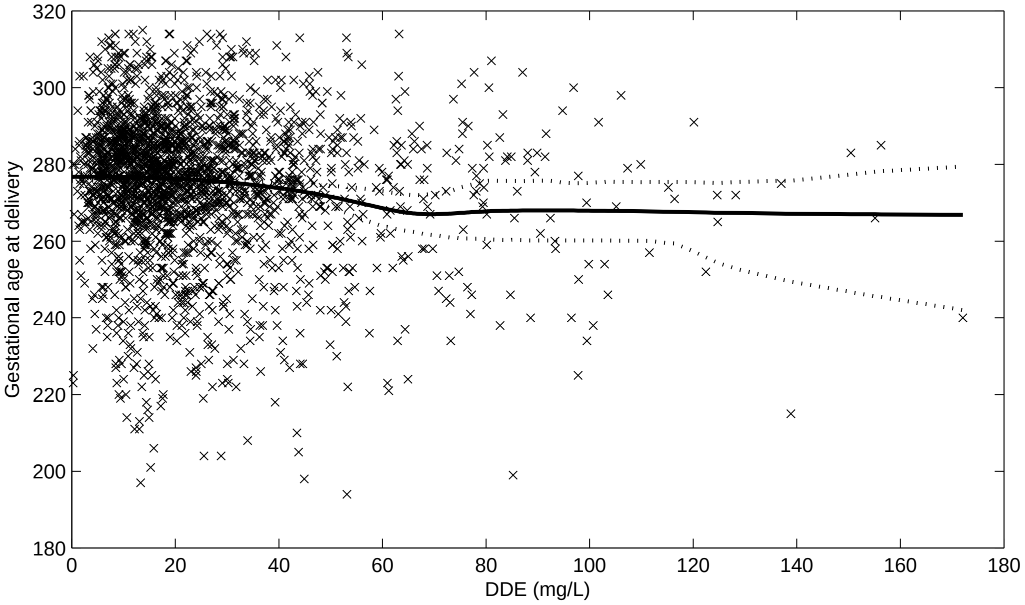

FIGURE 2.4: The conditional predictive mean of gestational age at delivery (solid line) with 99% pointwise credible intervals (dotted lines).

births were assigned to cluster 6, the mean DDE level of which was about the 80th percentile of observed DDE values. Preterm births in cluster 2 were characterized by both high DDE dose and old maternal age, while those in clusters 11 and 13 had extreme DDE levels beyond the 98th and 99th percentiles, respectively. It is observed that most of normal births in these 4 clusters had GAD values close to 37 weeks. Hence, the clustering result also strongly supports that preterm births were more likely to be observed with high DDE dose. Note that the order of clusters is arbitrary and that some of clusters have similar mean values of GAD, but they are separately grouped due to different predictor values.

Although the results of the analysis for conditional density estimation were similar to Dunson and Park (2008), the proposed computational algorithm was considerably less complex and simpler to implement. The kernel stick-breaking process (KSBP) proposed by Dunson and Park (2008) relied on a retrospective MCMC algorithm (Papaspiliopoulos and Roberts, 2007), which involved updating of random basis locations, stick-breaking weights, atoms and kernel parameters. In contrast, by using the GPPM proposed in the present paper, we bypass

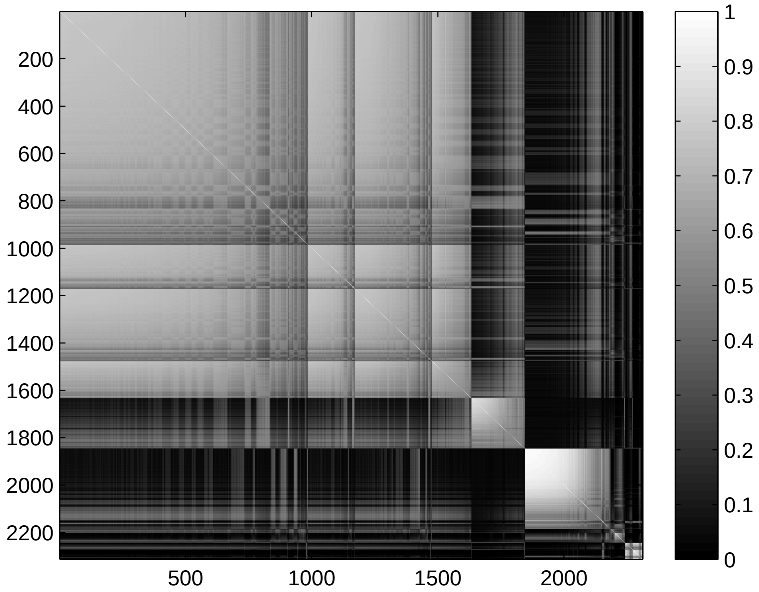

FIGURE 2.5: Pairwise marginal probabilities of being grouped with another subject in the CPP data.

the need to perform computation for the very many unknowns characterizing the collection of predictor-dependent mixture distributions. Instead through marginalization relying on the simple predictor-dependent urn scheme shown in Theorem 2.3.1, we obtain a simple and efficient Gibbs sampling algorithm. We found the mixing and convergence rates to be similar to those for the MCMC algorithm of the KSBP, but the computational time was substantially reduced, as fewer computations were needed at each step of the MCMC algorithm.

For predictive purposes, the KSBP may be more efficient in introducing only those clusters that are needed to flexibly characterize changes with predictors in the response distribution. However, in utilizing information in the predictor distribution, the GPPM may be particularly useful in semi-supervised learning settings, when there are missing predictors, and when interest focuses on inverse regression problems. Also, in many clustering applications, one would prefer to have subjects with very different predictor values but the same response allocated to different clusters.

TABLE 2.1: Summary statistics by clusters

		GAD[1]	DDE	AGE
Cluster	n[2]	mean (SD[3])	mean (SD[3])	mean (SD[3])
1	985 (1)	39.7 (1.21)	26.8 (14.49)	24.0 (5.72)
2	185 (0)	40.2 (1.04)	26.6 (14.05)	23.5 (5.57)
3	306 (0)	39.2 (1.10)	28.1 (13.78)	25.7 (6.19)
4	156 (0)	41.1 (1.21)	25.5 (13.80)	24.1 (4.85)
5	212 (0)	43.3 (0.78)	26.8 (14.89)	22.9 (5.37)
6	339 (309)	34.8 (1.94)	32.4 (16.55)	22.3 (5.09)
7	16 (0)	40.1 (0.91)	30.8 (18.13)	42.3 (1.53)
8	38 (30)	35.3 (2.05)	33.2 (14.93)	38.8 (2.69)
9	4 (0)	43.4 (0.88)	39.7 (14.39)	40.8 (0.50)
10	31 (0)	40.7 (1.53)	93.1 (9.62)	23.7 (5.20)
11	33 (19)	36.2 (2.38)	101.4 (13.95)	24.2 (6.65)
12	6 (0)	39.4 (1.03)	148.0 (13.88)	23.5 (4.51)
13	2 (2)	32.5 (2.53)	161.5 (23.48)	25.0 (1.41)

[1]in weeks, [2]preterm births in parenthesis, [3]SD=standard deviation

2.7 Discussion

There has been increasing interest in the use of partitioning to generate flexible classes of models and to identify interesting clusters of observations for further exploration. Much of the recent literature has relied on Dirichlet process-based clustering, an approach closely related to product partition models (PPMs). Our contribution is to develop a simple modification to PPMs to allow predictor dependent clustering, while bypassing the need for consideration of complex nonparametric Bayes methods for collections of predictor-dependent random probability measures. The resulting class of generalized PPMs (GPPMs) should be widely useful as a tool for generating new classes of models and for efficient computation in existing models, such as hierarchical mixtures-of-experts models.

Perhaps the most interesting and useful of our results is the proposed class of predictor-dependent urn schemes, which generalize the Blackwell and MacQueen (1973) Pólya urn scheme in a natural manner to include weights that depend on the distances between subjects predictor values. The distance metric is induced through a flexible nonparametric joint model for the predictors. Although this approach may be viewed as unnatural when the predictors are not random variables, the proposed class of predictor-dependent urn schemes are nonetheless useful and are defined conditionally on the predictor values. In this sense, the use of a joint distribution on the predictors in inducing the urn scheme can be viewed simply as a tool for proving that a coherent joint prior exists in cases in which the predictors are not random.

CHAPTER 3

BAYESIAN SEMIPARAMETRIC DENSITY REGRESSION WITH MEASUREMENT ERROR

3.1 Introduction

In many cases, a predictor X cannot be observed directly and one instead measures multiple surrogates $\mathbf{W} = (W_1, \ldots, W_q)'$. There is a rich literature on latent variable and measurement error models that allow inferences on the association between a predictor X and a response Y based on data collected for $\mathbf{W}$ and Y. In addition to assumptions of conditional independence of Y and $\mathbf{W}$ given X, which are necessary for identifiability, the literature in this area typically relies on a number of parametric assumptions. For example, most approaches assume the latent X is normally distributed and the models describing the conditional distributions of $\mathbf{W}$ and Y given X have a parametric form. The focus of this article is on using semiparametric Bayes methods to relax parametric assumptions to the extent possible given identifiability issues.

As a motivating application, we focus on data from the Longnecker et al. (2001) sub-study of the Collaborative Perinatal Project, which measured levels of two DDT (Dichlorodiphenyltrichloroethane) products (DDT isomer p,p'-DDT and persistent DDT metabolite p,p'-DDE) in maternal serum samples collected during pregnancy. Gestational age at delivery and birth weight data were also available. From a public health perspective, the main question of interest is how DDT exposure impacts the risk of adverse pregnancy outcomes, with “adverse” corresponding to the left tail of the distribution of gestational age at delivery and to the left or right tail of the birth weight distribution. As DDT exposure is only measured indirectly through its products in the serum, the level of exposure is a latent variable. To avoid sensitivity to arbitrarily chosen cutoffs, it is of interest to assess the effect of latent DDT exposure on the joint distribution of gestational age at delivery and birth weight. However, we find that the bivariate normal distribution provides a poor fit even after transformation. Hence, it is appealing to develop an approach that allows the joint distribution to change flexibly with exposure.

Although the vast majority of the literature on latent variable models has focused on normal linear structures, there has been an increasing focus on more flexible approaches. Attias (1999) used finite mixtures of Gaussians for latent variables in a factor analytic model, while Lee et al. (2008) generalized this approach to structural equation models using a Bayesian approach to model fitting and inference. Dunson et al. (2007) developed an alternative Bayes method, which allows for the incorporation of mean and variance constraints on the latent factors. To account for non-linearities within Gaussian latent variable models, Arminger and Muthén (1998) and Lee and Song (2003) incorporated quadratic terms and interactions, while Fahrmeir and Raach (2007) and Wang and Iyer (2007) used splines.

None of these approaches allow the conditional response distribution to vary nonparametrically according to latent predictors. When both Y and X are observed directly, this problem has been referred to as either density regression or conditional density estimation. For frequentist references, refer to Fan et al. (1996), Hyndman et al. (1996) and Fan and Yim (2004). Dunson et al. (2007) proposed an alternative approach for density regression, using predictor-dependent mixtures of normal linear regressions. Dunson and Park (2008) later developed a fully Bayes alternative based on kernel stick-breaking processes, while Park and Dunson (2007) instead relied on a generalized product partition model motivated by the formulations of Müller et al. (1996) and Quintana and Iglesias (2003). The order-based dependent Dirichlet process of Griffin and

Steel (2006) can also be used for flexible Bayes density regression.

In this article, our focus is on developing related methods for the case in which X is latent. Our proposed approach is based on a joint modeling strategy, which uses a Dirichlet process prior for the distribution of X and kernel stick-breaking process mixtures for the conditional distribution of Y given X. Our primary goal is to derive a simple but flexible mixture model for the conditional distribution of Y given X. In this context, Bayesian approaches have the advantage of allowing centering on a base parametric model, so that inferences can rely on the parametric model when appropriate, while adding flexibility in the presence of information in the data suggesting lack of fit.

Section 3.2 describes the general framework and discusses identifiability issues. Section 3.3 proposes a nonparametric Bayes formulation. Section 3.4 outlines an efficient Markov chain Monte Carlo algorithm (MCMC) for posterior computation. In Section 3.5 and Section 3.6, the proposed method is applied to a simulation example and reproductive epidemiologic data, respectively. Section 3.7 discusses the results.

3.2 Model Formulation

3.2.1 Proposed model

For subject i ($i = 1, \ldots, n$), the observed data consist of a $p \times 1$ response vector $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})$, a $q \times 1$ vector of error-prone measurements $\mathbf{w}_i = (w_{i1}, \ldots, w_{iq})'$ of a latent predictor x_i, and covariate vectors $\mathbf{v}_i = (v_{i1}, \ldots, v_{is})'$ and $\mathbf{z}_i = (z_{i1}, \ldots, z_{ir})'$. We focus on the case in which $\mathbf{y}_i \in \Re^p$ and $\mathbf{w}_i \in \Re^q$ and there is a single latent variable x_i, though extensions to accommodate mixed categorical and continuous measurements and multiple latent variables are straightforward.

We assume that the observed data likelihood can be expressed as

$$f(\mathbf{y}_i, \mathbf{w}_i \,|\, \mathbf{v}_i, \mathbf{z}_i) = \int f(\mathbf{y}_i \,|\, x_i, \mathbf{v}_i) f(\mathbf{w}_i \,|\, x_i, \mathbf{z}_i) f(x_i) dx_i, \tag{3.1}$$

where $f(\mathbf{y}\,|\,x, \mathbf{v})$ and $f(x)$ will be treated as unknown using Bayesian nonparametric methods, while $f(\mathbf{w}_i\,|\,x_i, \mathbf{z}_i)$ will be characterized using a simple normal linear measurement model as follows:

$$\mathbf{w}_i = \boldsymbol{\eta} + \boldsymbol{\Delta}\mathbf{z}_i + \boldsymbol{\lambda} x_i + \boldsymbol{\epsilon}_i, \tag{3.2}$$

where $\boldsymbol{\eta}$ is a $q \times 1$ intercept vector, $\boldsymbol{\Delta}$ is a $q \times s$ matrix of regression coefficients, $\boldsymbol{\lambda}$ is a $q \times 1$ vector of factor loadings, and $\boldsymbol{\epsilon}_i$ is a $q \times 1$ measurement error term with $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_w)$.

For the conditional likelihood $f(\mathbf{y}_i\,|\,x_i, \mathbf{v}_i)$, we propose to use a flexible mixture of normal linear measurement models as follows:

$$f(\mathbf{y}_i\,|\,x_i, \mathbf{v}_i) = \sum_{h=1}^{\infty} \omega_h(\mathbf{v}_i, x_i) N_p(\mathbf{y}_i; \boldsymbol{\mu}_h^* + \boldsymbol{\Psi}_h^* \mathbf{v}_i + \boldsymbol{\beta}_h^* x_i, \boldsymbol{\Sigma}_y), \tag{3.3}$$

where $\omega_h(\mathbf{v}_i, x_i)$ is a probability weight on the hth mixture component specific to predictor values including latent predictor value x_i, and $\boldsymbol{\theta}_h^* = \{\boldsymbol{\mu}_h^*, \boldsymbol{\Psi}_h^*, \boldsymbol{\beta}_h^*\}$ are parameters specific to component h, with $\boldsymbol{\mu}_h^*$ a $p \times 1$ intercept vector, $\boldsymbol{\Psi}_h^*$ a $p \times s$ matrix of regression coefficients, $\boldsymbol{\beta}_h^*$ a $p \times 1$ coefficient vector, and $\boldsymbol{\Sigma}_y$ a $p \times p$ covariance matrix. By mixing over normal linear latent factor models, we obtain a highly-flexible structure. To allow non-linear effects of the latent predictor x_i, it is necessary to allow the mixture weights to depend on x_i.

We complete the model with a specification for $f(x_i)$, with the specific form chosen depending on whether x_i is treated as discrete or continuous. Unknown smooth continuous densities can be accurately approximated by mixtures of Gaussian densities. However, we note that it is not possible to assess based on the observed data whether the distribution of the latent predictor is discrete or continuous without making unverifiable modeling assumptions. Hence, in order to simplify computation and interpretation, we focus on the case in which x_i is a discrete, with

$$x_i \sim G = \sum_{h=1}^{\infty} \pi_h \delta_{x_h^*}, \tag{3.4}$$

where π_h is the prior probability that $x_i = x_h^*$, so that individual i is allocated to the hth latent class, and δ_a denotes the measure that puts a unit point mass at a.

3.2.2 Identification

As for parametric latent variable models, identification is an important issue. Since the conditional likelihood functions in (3.2) and (3.3) are invariant to orthogonal transformations such as $\widetilde{\boldsymbol{\beta}}_h = \boldsymbol{\beta}^*_h \mathbf{p}'$, $\widetilde{x}_i = \mathbf{p} x_i$, $\widehat{\boldsymbol{\lambda}} = \boldsymbol{\lambda} \mathbf{q}'$, and $\widehat{x}_i = \mathbf{q} x_i$, where $\mathbf{p}$ and $\mathbf{q}$ are any orthogonal vectors (Lopez and West, 2004), it follows that constraints are needed for identifiability even in the parametric case.

Bayesian inferences can be conducted even in models that are non-identifiable from a frequentist perspective as long as informative priors are chosen. However, the resulting inferences may be very sensitive to the prior, even in large samples. Hence, it is appealing to incorporate identifiable constraints on parameters even in Bayesian models.

First, we assume the covariance matrix $\boldsymbol{\Sigma}_y$ of responses to be unstructured, while the covariance matrix $\boldsymbol{\Sigma}_w$ of the measurement error terms is forced to be diagonal. This is a standard assumption, which implies that dependence in the elements of $\mathbf{w}_i$ arises solely from shared dependence on $\mathbf{w}_i$. Additional constraints are made on $\boldsymbol{\eta}$ and $\boldsymbol{\lambda}$, with $\eta_1 = 0$, $\lambda_1 = 1$ and $\lambda_j > 0$, for $j = 2, \ldots, q$, to allow for unconstrained mean and variance of the latent variable, while avoiding sign ambiguity. We avoid constraining the mean and variance of the latent variable density, since this creates difficulties in nonparametric models.

3.3 Nonparametric Bayes Specification

Note that the framework proposed in Section 3.2 can be applied either from a frequentist perspective, using maximum likelihood estimation implemented with the EM algorithm, or from a Bayes perspective, using a nonparametric formulation implemented with MCMC. The Bayes approach has advantages, which result from avoiding the selection of a finite number of components with positive probabilities in (3.3) and (3.4). The frequentist method can be implemented for different numbers of components, using the BIC for dimensionality selection as is often done in finite mixture models. However, the BIC is not fully justified theoretically in this setting, and the resulting inferences and predictions would ignore uncertainty in estimating the number of components. In this section, we complete the model formulation in Section 3.2 by characterizing the conditional likelihood of $\mathbf{y}_i$ in (3.3) and the distribution of latent variable x_i in (3.4) using the nonparametric Bayes specifications.

First, a Dirichlet process (Ferguson, 1973; 1974) prior is used to allow the distribution G of latent variables to be unknown. It can be expressed in a hierarchical form as

$$x_i \sim G, \quad G \sim DP(\alpha G_0), \tag{3.5}$$

where $DP(\alpha G_0)$ denotes a Dirichlet process prior centered on base measure G_0 with precision α, and under the Sethuraman's (1994) stick-breaking representation of G,

$$G = \sum_{h=1}^{\infty} U_h \prod_{l<h} (1 - U_l) \delta_{x_h^*}, \quad U_h \overset{iid}{\sim} \text{Beta}(1, \alpha), \quad x_h^* \overset{iid}{\sim} G_0, \tag{3.6}$$

where $\mathbf{U} = \{U_h\}_{h=1}^{\infty}$ are stick-breaking weights and $\mathbf{x}^* = \{x_h^*\}_{h=1}^{\infty}$ are atoms, expression (3.4) is obtained with $\pi_h = U_h \prod_{l<h}(1 - U_l)$. In addition, letting S_i denote a latent variable indicating what latent class subject i belongs to, it follows that $x_i = x_{S_i}^*$.

On the other hand, a specification of the conditional likelihood function of $\mathbf{y}_i$ given $\mathbf{v}_i$ and x_i in (3.3) begins with considering the following hierarchical model:

$$\begin{aligned} \mathbf{y}_i | \mathbf{t}_i, \phi_i &\sim N_p(\mathbf{y}_i; \boldsymbol{\mu}_i + \boldsymbol{\Psi}_i \mathbf{v}_i + \boldsymbol{\beta}_i x_i, \boldsymbol{\Sigma}_y) \\ \phi_i | H_{\mathbf{t}_i} &\sim H_{\mathbf{t}_i} \\ \mathcal{H}_{\mathcal{T}} &\sim \mathcal{P}, \end{aligned} \tag{3.7}$$

where ϕ_i is a set of the parameters characterizing the conditional distribution of $\mathbf{y}_i$ given $\mathbf{t}_i$, with $\phi_i = \{\boldsymbol{\mu}_i, \boldsymbol{\Psi}_i, \boldsymbol{\beta}_i\}$ and $\mathbf{t}_i = \{\mathbf{v}_i, x_i\}$, $H_{\mathbf{t}_i}$ is an unknown distribution of ϕ_i which varies with predictor values $\mathbf{t}_i \in \mathcal{T}$, and $\mathcal{P}$ is the prior for the collection $\mathcal{H}_{\mathcal{T}} = \{H_{\mathbf{t}} : \mathbf{t} \in \mathcal{T}\}$. Although the most common choice for prior $\mathcal{P}$ in the nonparametric Bayes literature is a Dirichlet process with $H_{\mathbf{t}_i} \equiv H$, as in expression (3.5), it is not flexible enough to achieve our goal. The stick-breaking representation of the DP in (3.6) implies that mixing weights π_h are constant across predictors $\mathbf{t}_i$ and therefore together with the linear assumption for the model in (3.3), the conditional mean of $\mathbf{y}_i$ given $\mathbf{t}_i$ still remains linear. To express the conditional mean of $\mathbf{y}_i$ as a predictor-dependent mixture of linear models, we focus on generalizing the DP by incorporating predictor values into the stick-breaking structure of π_h as

$$\omega_h(\mathbf{t}) = V_h K(\mathbf{t}, \Gamma_h) \prod_{l<h} \big\{1 - V_l K_l(\mathbf{t}, \Gamma_l)\big\}, \tag{3.8}$$

where stick-breaking random variables $\mathbf{V} = \{V_h\}_{h=1}^{\infty}$ and random locations $\mathbf{\Gamma} = \{\Gamma_h\}_{h=1}^{\infty}$ independently and identically follow a beta distribution with parameters 1 and γ and a distribution F_Γ, respectively, and $K(\mathbf{x}, \Gamma_h) = \exp(-\psi_h ||\mathbf{x} - \Gamma_h||^2)$ is the Gaussian kernel function mapping to [0, 1], and letting $\boldsymbol{\theta}^* = \{\boldsymbol{\theta}_h^*\}_{h=1}^{\infty}$ denote an independent sample from a base measure H_0, with $\boldsymbol{\theta}_h^*$ assigned to the hth location Γ_h, the unknown distribution of $H_\mathbf{t}$ is formulated as

$$H_\mathbf{t} = \sum_{h=1}^{\infty} \omega_h(\mathbf{t}) \delta_{\boldsymbol{\theta}_h^*}, \tag{3.9}$$

which is a special case of the kernel stick-breaking processes (KSBPs) (Dunson and Park, 2008), with $\sum_{h=1}^{\infty} \omega_h(\mathbf{t}) = 1$. One of the appealing properties of expression (3.9) is that it works in a similar way to the DP, but with probability weights $\boldsymbol{\omega} = \{\omega_h(\mathbf{t})\}_{h=1}^{\infty}$ in (3.8) flexibly altered by the distances between predictor value $\mathbf{t}$ and locations $\mathbf{\Gamma}$, which range from 0 to 1, measured through the kernel function K. Hence, an atom $\boldsymbol{\theta}_h^*$ is assigned higher weight, as its location index h is lower and its corresponding location Γ_h is closer to predictor value $\mathbf{t}$, resulting in a sparse representation. Note that the DP is obtained as a special case with $K(\mathbf{t}, \Gamma) = 1$ for any ($\mathbf{t}$, Γ). Additional flexibility is allowed by letting the kernel precision ψ_h vary across locations, so that a few number of atoms are needed for sparse areas of $\mathcal{T}$.

Letting R_i denote a latent variable indicating what location subject i is assigned to, the model in (3.7) can be reexpressed as

$$\begin{aligned}\mathbf{y}_i|\,\mathbf{v}_i, x_i, R_i, \boldsymbol{\theta}, \boldsymbol{\Sigma}_y &\sim N_p(\mathbf{y}_i;\, \boldsymbol{\mu}^*_{R_i} + \boldsymbol{\Psi}^*_{R_i}\mathbf{v}_i + \boldsymbol{\beta}^*_{R_i}x_i, \boldsymbol{\Sigma}_y)\\ R_i|\mathbf{x}_i, \boldsymbol{\omega} &\sim \sum_{h=1}^{\infty}\omega_h(\mathbf{t}_i)\delta_h,\\ \boldsymbol{\theta}^*_h &\overset{iid}{\sim} H_0 \end{aligned} \qquad (3.10)$$

and therefore integrating out with respect to R_i results in the conditional distribution of $\mathbf{y}_i$ given $\mathbf{t}_i$ in (3.3).

On the other hand, we obtain another important mixture structure as a result of our specification. Letting $\mathbf{d}_i = (\mathbf{y}'_i, \mathbf{w}'_i)'$, $\boldsymbol{\mu}^*_i = \{(\boldsymbol{\mu}_{R_i} + \mathbf{B}_{R_i}\mathbf{v}_i)', (\boldsymbol{\eta} + \boldsymbol{\Delta}\mathbf{z}_i)'\}'$, $\boldsymbol{\lambda}^*_i = (\boldsymbol{\beta}'_{R_i}, \boldsymbol{\lambda}')'$, and $\mathbf{e}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_e)$, where $\boldsymbol{\Sigma}_e = \mathrm{diag}(\boldsymbol{\Sigma}_y, \boldsymbol{\Sigma}_w)$, two models in (3.2) and (3.3) conditional on R_i are vertically stacked into one regression model as

$$\mathbf{d}_i = \boldsymbol{\mu}^*_i + \boldsymbol{\lambda}^*_i x_i + \mathbf{e}_i,$$

and, integrating out x_i in (3.5) and (3.6) results in a Dirichlet process mixture (Lo, 1984; Escobar and West, 1995) of normal linear measurement models,

$$f(\mathbf{d}_i|\boldsymbol{\mu}^*_i, \boldsymbol{\lambda}^*_i, \mathbf{U}, \mathbf{x}^*, R_i) = \sum_{h=1}^{\infty}\pi_h N_{p+q}(\mathbf{d}_i; \boldsymbol{\mu}^*_i + \boldsymbol{\lambda}^*_i x^*_h, \boldsymbol{\Sigma}_e). \qquad (3.11)$$

The mixture models in (3.3) and (3.11) play an important role in posterior computation, which is discussed in the next section.

3.4 Posterior Computation

Dunson and Park (2008) proposed a conditional approach to posterior computation for KSBP mixture models, relying on a combined Markov chain Monte Carlo (MCMC) algorithm that uses retrospective sampling (Papaspiliopoulos and Roberts, 2007) and generalized Pólya urn sampling

(MacEachern, 1994; West et al., 1994) steps. However, since the slice sampler (Neal, 2003; Walker, 2007) is simpler to implement than the retrospective sampler, we propose a MCMC algorithm that relies on a slice sampler for both the KSBP and the DP priors.

3.4.1 Slice sampler

Suppose that we have the following mixture model for observed variable r:

$$f(r) = \sum_{h=1}^{\infty} p_h f(r|\theta_h), \tag{3.12}$$

where $f(\cdot|\theta)$ is a parametric density characterized by parameter θ and $\mathbf{p} = (p_1, \ldots, p_\infty)$ are probability weights, with $\sum_{h=1}^{\infty} p_h = 1$. We introduce a latent variable u in expression (3.12), such that the joint density of r and u is given by

$$\begin{aligned} f(r, u) &= \sum_{h=1}^{\infty} p_h U(u|0, p_h) f(r|\theta_h) \\ &= \sum_{h=1}^{\infty} 1(u < p_h) f(r|\theta_h), \end{aligned}$$

where $U(\cdot|a, b)$ is a uniform density ranging from a to b. It is obvious that the marginal density of r in (3.12) can be obtained by integrating the above joint density with respect to u. Appealing features of introducing latent variable u are that u is independent of r given weights $\mathbf{p}$ and that the conditional density of r given u is expressed as a finite mixture model only with finite weights such that $\{p_h; u < p_h\}$ (Walker, 2007). Letting T denote another latent variable indexing which mixture component variable r is sampled from, so that $r = \theta_T$, the complete data likelihood of an *iid* sample $(\mathbf{r}, \mathbf{u}, \mathbf{T}) = \{r_i, u_i, T_i\}_{i=1}^{n}$ is

$$L(\mathbf{r}, \mathbf{u}, \mathbf{T};\, \mathbf{p}, \boldsymbol{\theta}) = \prod_{i=1}^{n} 1(u_i < p_{T_i}) f(r_i|\theta_{T_i}), \tag{3.13}$$

and upon marginalizing out $\{u_i\}_{i=1}^n$, the complete data likelihood of $\{r_i, T_i\}_{i=1}^n$ is

$$L(\mathbf{r}, \mathbf{T}; \mathbf{p}, \boldsymbol{\theta}) = \prod_{i=1}^{n} p_{T_i} f(r_i|\theta_{T_i}). \tag{3.14}$$

The likelihood function in expression (3.13) is used to derive a full conditional distribution of the latent variables u and T, whereas that in expression (3.14) is used for component specific parameters $\boldsymbol{\theta}$ and mixture weights $\mathbf{p}$. Then, letting $q(\rho)$ be an appropriate prior density for parameter ρ, a slice sampler proceeds in the following steps:

S1. Sample u_i from $\mathrm{U}(u_i; 0, p_{T_i})$, for $i = 1, \ldots, n$.

S2. Sample θ_h, for $h = 1, \ldots, \infty$ from the conditional posterior distribution

$$(\theta_h|\mathbf{r}, \mathbf{T}) \propto q(\theta_h) \prod_{i:T_i=h} f(r_i|\theta_h).$$

S3. Sample $\mathbf{p}$ from the conditional posterior distribution

$$(\mathbf{p}|\mathbf{T}) \propto q(\mathbf{p}) \prod_{i=1}^{n} p_{T_i}.$$

S4. Sample $\mathbf{T}_i$, for $i = 1, \ldots, n$, from the conditional posterior distribution

$$\Pr(T_i = h|\mathbf{r}, u_i, \mathbf{p}) \propto 1(h \in A_i(u_i)) f(r_i|\theta_h),$$

where $A_i(u_i) = \{h; p_h > u_i\}$ is a finite index subset defined by sampling p_h, for $h = 1, \ldots, k$, with k being the smallest value satisfying

$$\sum_{h=1}^{k} p_h > (1 - u^*),$$

where $u^* = \min_i\{u_i\}$.

Note that the slice algorithm proposed in this article is different from Walker's algorithm in that our approach uses the likelihood function in expression (3.14) instead of that in expression (3.13) to sample mixture weights $\mathbf{p}$. It is observed that both algorithms performs well in terms of mixing and convergence rate, but our algorithm provides an analytically simple conditional posterior distribution for mixture weights in the KSBP.

3.4.2 Details of MCMC algorithm

For easy exposition of the posterior computation, we consider the following regression model:

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta}_i^* + \boldsymbol{\epsilon}_i^y, \tag{3.15}$$

where a $p \times p(r+2)$ matrix $\mathbf{X}_i = I_p \otimes (1, \mathbf{r}_i', x_i)$, with $\otimes$ denoting the Kronecker product, a $p(r+2) \times 1$ vector $\boldsymbol{\beta}_i^* = \{(\mu_{i,1}, \text{row}_1(\boldsymbol{\Psi}_i), \beta_{i,1}), \ldots, (\mu_{i,p}, \text{row}_p(\boldsymbol{\Psi}_i), \beta_{i,p})\}'$, with $\text{row}_i(A)$ denoting the ith row of matrix A, and a $p \times 1$ residual term $\boldsymbol{\epsilon}_i^y \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}_y)$. Note that the regression model in (3.15) corresponds to the normal linear measurement model in (3.7). For convenience in elicitation and computation, we assume the following conjugate priors for the parameters and latent variable to complete a Bayesian specification of the model: $H_0(\boldsymbol{\beta}_i^*) = N_{p(r+2)}(\boldsymbol{\beta}_i^*; \boldsymbol{\beta}_0, \boldsymbol{\Sigma}_\beta)$, $G_0(x_i) = N_1(x_i; \mu_x, \sigma_x^2)$, $\pi(\boldsymbol{\eta}) = N_q(\boldsymbol{\eta}; \boldsymbol{\eta}_0, \boldsymbol{\Sigma}_\eta)$, $\pi(\boldsymbol{\Delta}) = \prod_{j=1}^q N_s(\text{row}_j(\boldsymbol{\Delta}); \boldsymbol{\delta}_0, \boldsymbol{\Sigma}_\delta)$, $\pi(\boldsymbol{\lambda}) = \prod_{j=1}^q N^+(\lambda_j; \lambda_0, \sigma_\lambda^2)$, $\pi(\boldsymbol{\Sigma}_y) = \mathcal{W}\{\boldsymbol{\Sigma}_y; (\nu_0 \boldsymbol{\Sigma}_0)^{-1}, \nu_0\}$, $\pi(\tau_j) = \mathcal{G}(\tau_j; a_\tau, b_\tau)$, and $\pi(\alpha) = \mathcal{G}(\alpha; a_\alpha, b_\alpha)$, where τ_j, for $j = 1, \ldots, q$, are the inverse of the diagonal elements of $\boldsymbol{\Sigma}_w$ and $N^+(\cdot; a, b)$ denotes a normal distribution with mean a and variance b, but truncated below at zero, $\mathcal{W}(\cdot; \mathbf{C}, c)$ a Wishart distribution with mean $c\mathbf{C}$ and degrees of freedom c, and $\mathcal{G}(\cdot; a, b)$ a Gamma distribution with mean a/b. Note that only unconstrained elements of $\boldsymbol{\eta}$ and $\boldsymbol{\lambda}$ except for η_1 and λ_1 will be updated. In addition, since we don't know a priori how many classes we need for latent variable x, we update the DP precision α, which controls the number of unique values among $\{x_i\}_{i=1}^n$. Our MCMC algorithm then alternates the following steps:

1. Noting that expression (3.11) is equivalent to expression (3.12), we update the components of the DP mixture and the latent variable x_i using the slice sampler in Section 3.4.1 with

$r_i = \mathbf{d}_i$, $\mathbf{p} = \boldsymbol{\omega}$, $\theta_h = x^*_h$, $u_i = u_{x,i}$, and $T_i = S_i$. The full conditional posterior distributions of the relevant parameters can be expressed as follows:

$$
\begin{aligned}
x^*_h &\sim N_1(x^*_h; \widehat{x}^*_h, \widehat{V}_{x,h}),\\
U_h &\sim \text{Beta}\left(1 + m_h, \alpha + \sum_{l>h} m_l\right),
\end{aligned}
$$

where $\widehat{V}_{x,h} = (\sigma_x^{-2} + \sum_{i:S_i=h} \boldsymbol{\lambda}_i^{*\prime} \boldsymbol{\Sigma}_e^{-1} \boldsymbol{\lambda}_i^*)^{-1}$, $\widehat{x}^*_h = \widehat{V}_{x,h}\{\sigma_x^{-1}\mu_x + \sum_{i:S_i=h} \boldsymbol{\lambda}_i^{*\prime} \boldsymbol{\Sigma}_e^{-1}(\mathbf{d}_i - \boldsymbol{\mu}_i^*)\}$, and $m_h = \sum_{i=1}^n 1(S_i = h)$, for $h = 1, \ldots, \infty$. In addition, as in West (1992), α is updated by sequentially sampling from the full conditional distributions

$$
\begin{aligned}
&\eta_\alpha \sim \text{Beta}(\alpha_x + 1, n)\\
&\alpha \sim p\mathcal{G}(a_\alpha + k_x, b_\alpha - log(\eta_\alpha)) + (1 - p)\mathcal{G}(a_\alpha + k_x - 1, b_\alpha - log(\eta_\alpha)),
\end{aligned}
$$

where k_x is the number of unique values among $\{x_i\}_{i=1}^n$ and weight p is defined by

$$
\frac{p}{1-p} = \frac{(a_\alpha + k_x - 1)}{n(b_\alpha - log(\eta_\alpha))}.
$$

2. Also noting that expression (3.3) is equivalent to expression (3.12), we update the components of the KSBP mixture and $\boldsymbol{\beta}_i^*$ using the slice sampler by letting $r_i = \mathbf{y}_i$, $\mathbf{p} = \pi(\mathbf{t}_i)$, $\theta_h = \boldsymbol{\theta}_h^*$, $u_i = u_{y,i}$, and $T_i = R_i$. The full conditional posterior distributions of $\boldsymbol{\theta}_h^*$ is

$$
\boldsymbol{\theta}_h^* \sim N_{p(r+2)}(\boldsymbol{\theta}_h^*; \widehat{\boldsymbol{\theta}}_h, \widehat{V}_{\theta_h}),
$$

where $\widehat{V}_{\theta_h} = (\boldsymbol{\Sigma}_\beta^{-1} + \sum_{i:R_i=h} \mathbf{X}_i' \boldsymbol{\Sigma}_y^{-1} \mathbf{X}_i)^{-1}$ and $\widehat{\boldsymbol{\theta}}_h = \widehat{V}_{\theta_h}(\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\beta}_0 + \sum_{i:R_i=h} \mathbf{X}_i' \boldsymbol{\Sigma}_y^{-1} \mathbf{y}_i)$, and $\boldsymbol{\beta}_i^* = \boldsymbol{\theta}_{R_i}^*$. In Step *S3*, we use a data augmentation approach to update V_h, for $h = 1, \ldots, k$, as in Dunson and Park (2008). For each h, let A_{ih} and B_{ih} independently follow a Bernoulli distribution with probability V_h and $K(x_i, \Gamma_h)$, respectively, with

$R_i = \min\{h : A_{ih} = B_{ih} = 1\}$. Then, alternate between sampling (A_{ih}, B_{ih}) from their joint conditional distribution given R_i and updating V_h by sampling from the conditional posterior distribution

$$\text{Beta}\bigg(1 + \sum_{i:R_i \geq h} A_{ih}, \gamma + \sum_{i:R_i \geq h} (1 - A_{ih})\bigg).$$

The random location Γ_h and kernel precision ψ_h, for $h = 1, \ldots, k_y$, can be updated by a Metropolis-Hastings step or a Gibbs step if $F_\Gamma(\cdot) = \sum_{l=1}^{T} a_l \delta_{\Gamma_l^*}(\cdot)$, with $\Gamma^* = (\Gamma_1^*, \ldots, \Gamma_T^*)'$ a grid of potential locations.

3. Sample $\boldsymbol{\eta}$, $\boldsymbol{\Delta}$, $\boldsymbol{\lambda}$, $\boldsymbol{\Sigma}_y$, and $\boldsymbol{\Sigma}_w = diag(\tau_1^{-1}, \ldots, \tau_q^{-1})$ from

$$\begin{aligned}
\boldsymbol{\eta} &\sim N_q\bigg\{\boldsymbol{\eta}; \widehat{V}_\eta\bigg(\boldsymbol{\Sigma}_\eta^{-1}\boldsymbol{\eta}_0 + \sum_{i=1}^{n} \boldsymbol{\Sigma}_w^{-1}\{\mathbf{w}_i - (\boldsymbol{\Delta}\mathbf{z}_i + \boldsymbol{\lambda} x_i)\}\bigg), \widehat{V}_\eta\bigg\}, \\
\lambda_j &\sim N^+\bigg\{\lambda_j; \widehat{V}_\lambda\bigg(\sigma_\lambda^{-2}\lambda_0 + \sum_{i=1}^{n} \tau_j x_i\{w_{i,j} - (\eta_j + \text{row}_j(\boldsymbol{\Delta})\mathbf{z}_i)\}\bigg), \widehat{V}_\lambda\bigg\}, \\
\text{row}_j(\boldsymbol{\Delta})' &\sim N_s\bigg\{\text{row}_j(\boldsymbol{\Delta})'; \widehat{V}_\delta\bigg(\boldsymbol{\Sigma}_\delta^{-1}\boldsymbol{\delta}_0 + \sum_{i=1}^{n} \mathbf{z}_i\tau_j\{w_{i,j} - (\eta_j + \lambda_j x_i)\}\bigg), \widehat{V}_\delta\bigg\}, \\
\boldsymbol{\Sigma}_y^{-1} &\sim \mathcal{W}\bigg(\boldsymbol{\Sigma}_y^{-1}; \bigg\{\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_i^*)(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_i^*)' + \nu_0\boldsymbol{\Sigma}_0\bigg\}^{-1}, n + \nu_0\bigg), \\
\tau_j &\sim \mathcal{G}\bigg(\tau_j; a_\tau + \frac{n}{2}, b_\tau + \frac{1}{2}\sum_{i=1}^{n}\big\{w_{i,j} - (\eta_j + \text{row}_j(\boldsymbol{\Delta})\mathbf{z}_i + \lambda_j x_i)\big\}^2\bigg),
\end{aligned}$$

where $\widehat{V}_\eta = (\boldsymbol{\Sigma}_\eta^{-1} + n\boldsymbol{\Sigma}_w)^{-1}$, $\widehat{V}_\lambda = (\sigma_\lambda^{-2} + \sum_{i=1}^{n} \tau_j x_i^2)^{-1}$, and $\widehat{V}_\delta = (\boldsymbol{\Sigma}_\delta^{-1} + \sum_{i=1}^{n} \tau_j \mathbf{z}_i\mathbf{z}_i')^{-1}$.

3.5 Simulation Study

In this section, we examine the performance of the proposed method for conditional density estimation, using posterior samples from the MCMC algorithm. To implement the algorithm, we let $\boldsymbol{\beta}_0 = \mathbf{0}_{p(r+2)}$, $\boldsymbol{\Sigma}_\beta = \mathrm{I}_{p(r+2)\times p(r+2)}$, $\mu_x = 0$, $\sigma_x^2 = 1$, $\boldsymbol{\eta}_0 = \mathbf{0}_q$, $\boldsymbol{\Sigma}_\eta = \mathrm{I}_{q\times q}$, $\boldsymbol{\delta}_0 = \mathbf{0}_s$, $\boldsymbol{\Sigma}_\delta = \mathrm{I}_{s\times s}$, $\lambda_0 = 0$, $\sigma_\lambda^2 = 1$, $\nu_0 = p + 1$, $\boldsymbol{\Sigma}_0 = \mathrm{I}_{p\times p}$, $a_\tau = b_\tau = 0.1$, and $a_\alpha = b_\alpha = 2$, where $\mathbf{0}_a$ is a $a \times 1$ vector of zeros and I_b is an identity matrix of dimension b. The potential locations for $\boldsymbol{\Gamma}$ are

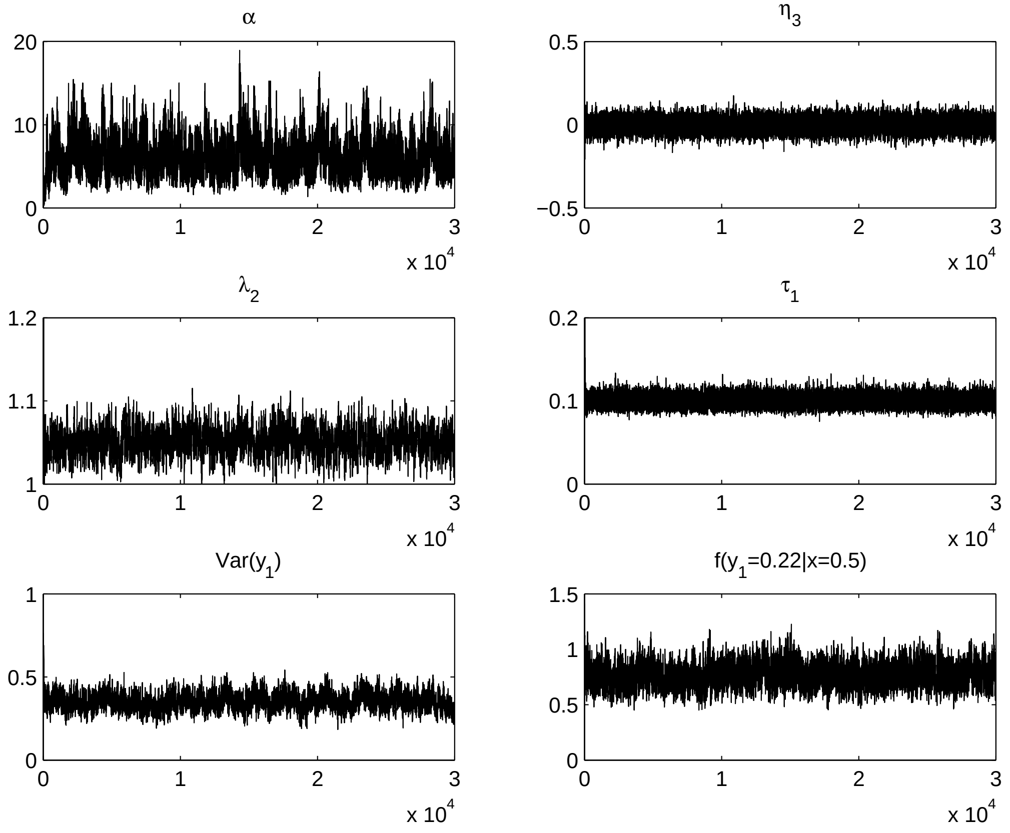

FIGURE 3.1: Trace plots of representative quantities.

taken from -3 to 3 with increment of 0.1, with $T = 61$ and $a_l = 1/T$. For more flexibility, we also let each kernel precision ψ_h be unknown with a log normal prior with mean 0 and variance 4. All observed data are normalized prior to analysis.

After conducting some descriptive and graphical analysis of the reproductive data handled in Section 3.6, we generated data of size 500 which are similarly distributed as the real data, as follows:

1. Latent variable x_i was sampled from $\mathrm{U}(x_i; 0, 1)$.

2. Response vector $\mathbf{y}_i$ with $p = 2$ was sampled from the mixture distribution

$$f(\mathbf{y}_i|x_i) = \frac{2(1-x_i)}{3} N\left\{\mathbf{y}_i; \begin{pmatrix} x_i \\ -x_i^3 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} 0.2^2 \right\}$$

$$+\frac{1}{3}N\left\{\mathbf{y}_i;\begin{pmatrix}0.5\\0\end{pmatrix},\begin{pmatrix}1&0.8\\0.8&1\end{pmatrix}0.4^2\right\}$$

$$+\frac{2x_i}{3}N\left\{\mathbf{y}_i;\begin{pmatrix}x_i\\0.1+2(x_i-0.5)^2\end{pmatrix},\begin{pmatrix}1&0.2\\0.2&1\end{pmatrix}0.2^2\right\}.$$

3. Error-prone predictor vector $\mathbf{w}_i$ with $q = 3$ was sampled from the linear regression model

$$f(\mathbf{w}_i|x_i) = N\left\{\mathbf{w}_i;\begin{pmatrix}0\\-2\\2\end{pmatrix}+\begin{pmatrix}1\\2\\0.4\end{pmatrix}x_i,\begin{pmatrix}0.1^2&0&0\\0&0.05^2&0\\0&0&0.2^2\end{pmatrix}\right\}.$$

Since other predictors $\mathbf{v}$ and $\mathbf{z}$ are not considered in the simulated data, we analyzed the data using the proposed semiparametric latent variable model without $\mathbf{v}$ and $\mathbf{z}$ and their corresponding parameters $\boldsymbol{\Psi}$ and $\boldsymbol{\Delta}$. The MCMC algorithm was run for 30,000 iterations with the first 10,000 discarded as a burn-in period. Figure 3.1 shows trace plots of the DP precision α, the intercept η_3, factor loading λ_2, precision τ_1, variance for y_1, and estimated conditional density at $y_1 = 0.22$ given $x = 0.5$. Based on examination of these plots, it can be said that convergence occurred quickly showing good mixing rate. In Figure 3.2, the left panel depicts the true marginal densities (dotted line), estimated predictive marginal densities (solid line) of y_1 , along with pointwise 99% credible intervals (dashed lines) at the 10th, 50th, and 90th percentiles of the distribution of x, while the right panel does the same for y_2. For both y_1 and y_2, the estimated densities are indistinguishable from the true densities. The pointwise credible intervals are relatively wider in the area in which the true density deviates from the normal density, implying that the KSBP with one atom at each location is still flexible enough to characterize changes in distribution occurring within a narrow area. With the same plot formatting as in Figure 3.2, the conditional predictive densities of y_1 given y_2 and x are plotted in Figure 3.3. The selected y_2 values are corresponding to $\bar{y}_2 - 2S_{y_2}$, $\bar{y}_2$, and $\bar{y}_2 + 2S_{y_2}$, where $\bar{y}_2$ and S_{y2} denotes the empirical mean and standard deviation of y_2. In all cases, 99% pointwise

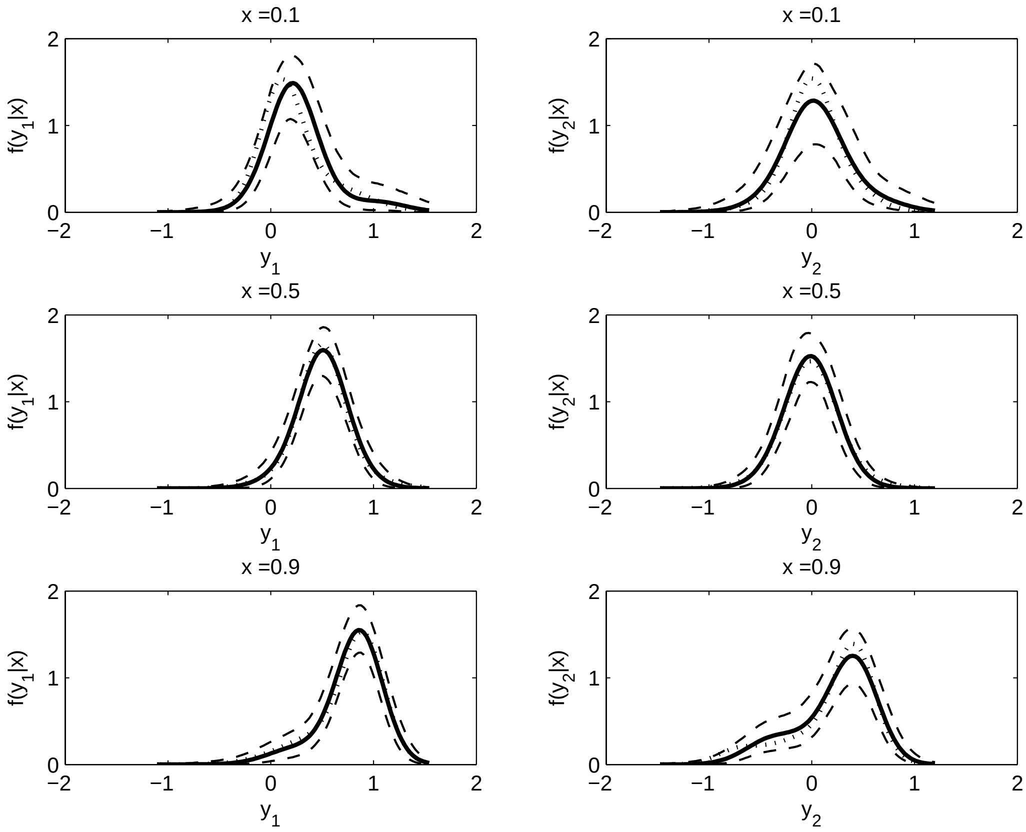

FIGURE 3.2: Estimated predictive conditional densities of y_1 (left) and y_2 (right) at the 10th, 50th and 90th percentiles of x: posterior means (solid lines), pointwise 99% credible intervals (dashed lines), and true values (dotted lines).

credible intervals completely embed the true conditional densities. Note that the intervals are wider, when y_2 and x are discordant, resulting in fewer observations. It was observed that as the sample size increases, estimated conditional densities get closer to the true ones with much narrower credible intervals (results not shown).

3.6 Application to Reproductive Epidemiology

3.6.1 Background

In this section, we return to the motivating reproductive study of DDT, briefly introduced in Section 3.1. DDT has been widely used for malaria vector control, but several adverse effects of

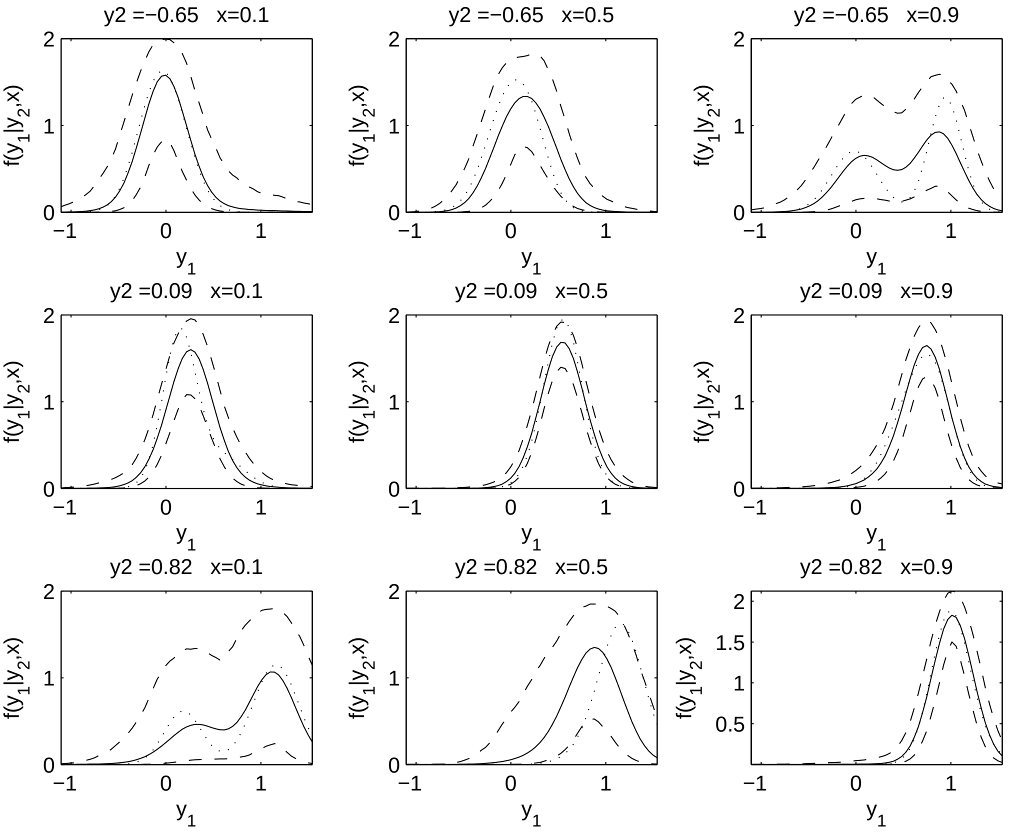

FIGURE 3.3: Estimated predictive conditional density of y_1 given y_2 at the 10th (left), 50th (middle) and 90th (right) percentiles of x. The dotted, solid, and dashed lines represent the true values, posterior means, and pointwise 99% credible intervals, respectively.

DDT have been reported not only in animals but also in human beings. Since it is not feasible to directly measure the level of DDT bioaccumulated in a human body through food consumption, it is common in the literature to measure p, p'-DDT and p, p'-DDE instead. Several animal studies have demonstrated that p, p'-DDT and p, p'-DDE have negative effects on reproductive factors, usually resulting in premature delivery and lower birth weight (Jusko et al., 2006).

Longnecker et al. (2001) examined the effects of DDT on preterm birth and small-for-gestational-age (SGA), using p, p'-DDE measured from serum samples of pregnant women enrolled in the Collaborative Perinatal Project (CPP) between 1959 and 1966. The authors showed that DDE had significant negative effects on both fetal outcomes. Dunson and Park (2008) analyzed the CPP data using a KSBP mixture model for density regression with a focus on the effect of DDE on preterm birth and drew a conclusion concordant to Longnecker et al. (2001)

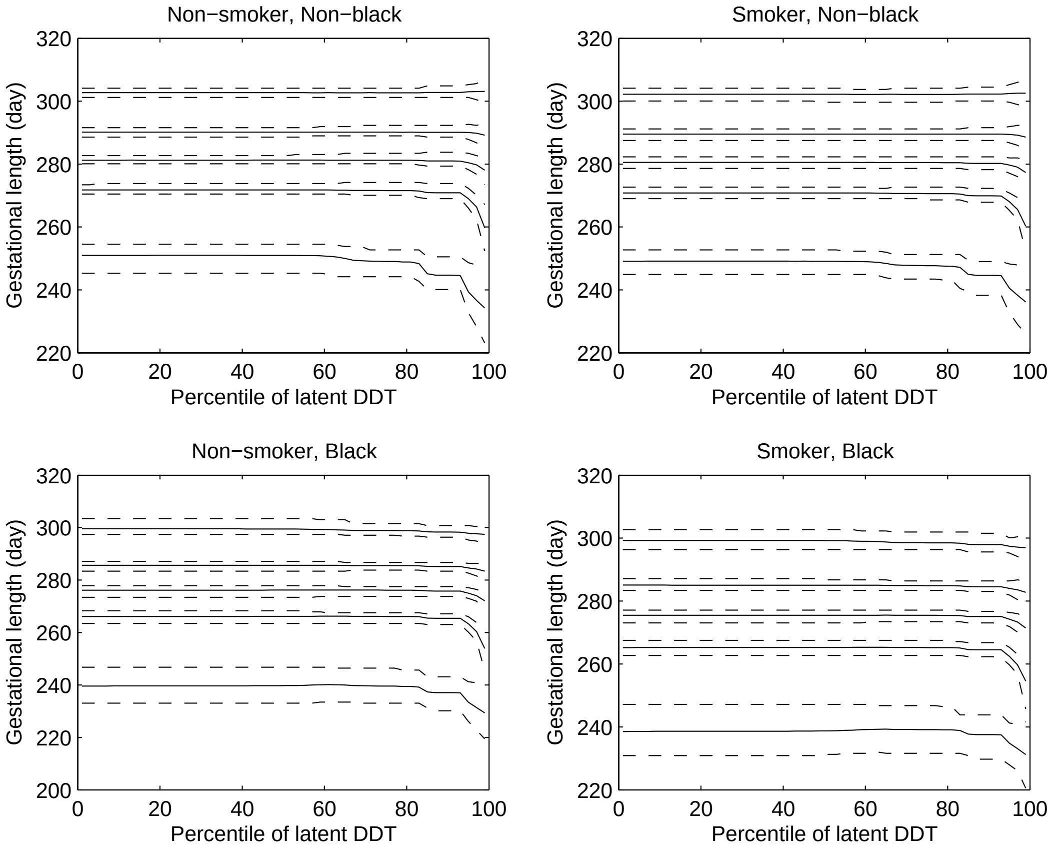

FIGURE 3.4: Plots of the 5th, 25th, 50th, 75th, and 95th percentiles of estimated conditional GAD density over percentiles of DDT by smoking status and ethnic group. The solid lines are the posterior means and the dashed lines are pointwise 99% credible intervals.

that high level concentration of DDE increased the risk of preterm birth.

3.6.2 Analysis and results

We analyzed the CPP data using the proposed semiparametric model with $\mathbf{y} = (\mathrm{BW},\ \mathrm{GAD})'$ being a response vector of birth weight (BW) and gestational age at delivery (GAD), $\mathbf{w} = (\mathrm{DDE}, \mathrm{DDT})'$ a error-prone predictor vector of maternal p, p'-DDE and p, p'-DDT, $\mathbf{v} = (\mathrm{TG}, \mathrm{CHOL}, \mathrm{RACE}, \mathrm{S}$ another predictor vector of serum triglycerides (TG), cholesterol (CHOL), infant ethnic origin (RACE), and mother's smoking status (SMOK), where RACE dichotomizes infant ethnicity into black and other groups and SMOK divides mother's smoking habit during pregnancy into current and non-current smoker. Since serum lipids have an effect on concentration of serum

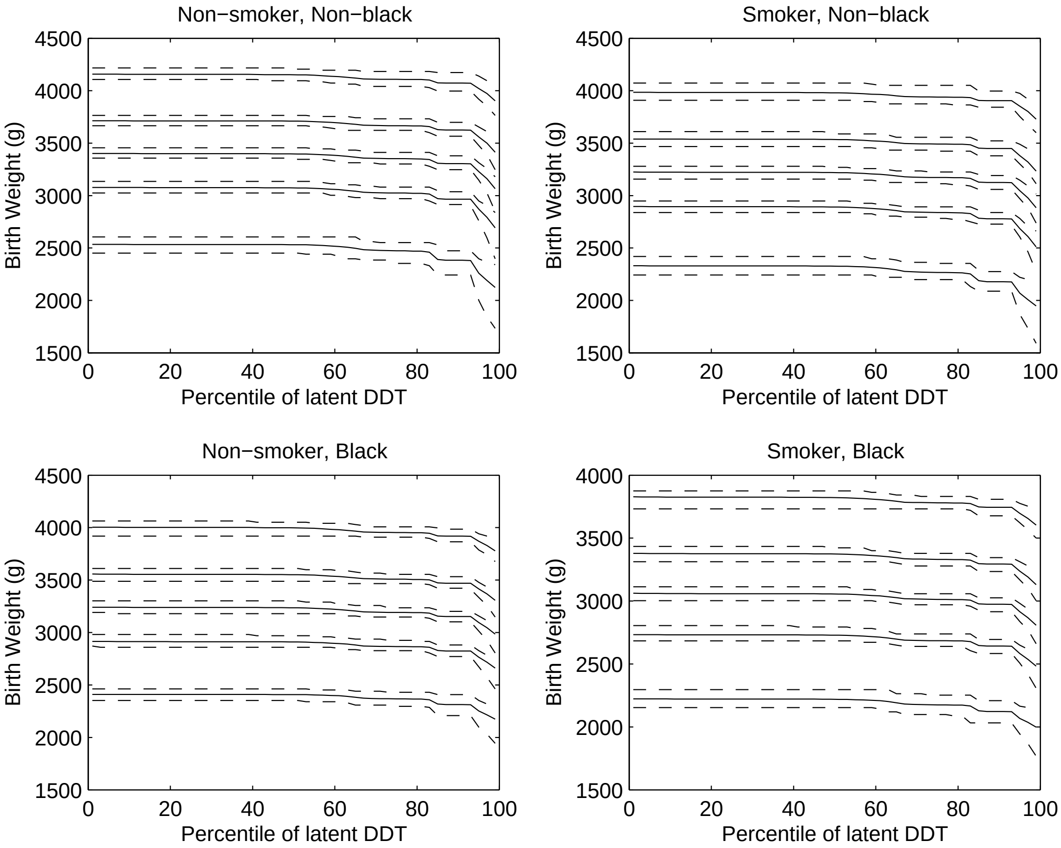

FIGURE 3.5: Plots of the 5th, 25th, 50th, 75th, and 95th percentiles of estimated conditional BW density over percentiles of DDT by smoking status and ethnic group. The solid lines are the posterior means and the dashed lines are 99% credible intervals.

DDE and DDT, triglycerides and cholesterol were also considered in the measurement model in (3.2), with $\mathbf{z} = (\mathrm{TG}, \mathrm{CHOL})'$. Data from the 2301 mothers and their children were used for analysis, after 79 subjects were excluded for biologically unrealistic values in GAD (> 45 weeks). The specification of the hyperparameters to implement the MCMC algorithm is the same as in the simulation example but with $p = 2$, $q = 2$, $r = 4$, and $s = 2$, and the chain was run for $30,000$ iterations with a burn-in period of $15,000$. It was observed that both mixing and convergence of the chain looked good (not shown).

Figures 3.4 and 3.5 show the 5th, 25th, 50th, 75th, and 95th percentiles of estimated density of GAD and BW, respectively, over percentiles of the empirical DDT distribution by mother's smoking status and infant ethnicity. Since different parameterizations and constraints for identifiability result in different values for the latent variable, making it difficult to interpret the meaning of such values, we make inferences at percentiles of an empirical distribution of the latent variable, which are invariant to different choices of parameterizations and identification constraints. From Figure 3.4, it is obvious that GAD is not normally distributed, showing left-skewness, and as the DDT level goes beyond its 60th percentile, the left tail of the GAD densities gets heavier for all ethnic groups and smoking status. In contrast, BW is approximately normally-distributed and the location of the BW densities moves to a lower value for higher DDT level, with no change in the shape of it (Figure 3.5). As in the simulation example, wide credible intervals for the estimated density at high DDT level are also due to sparse data. It is observed that all the percentiles of GAD and BW for smoking mothers and black babies are rather lower than those for non-smokers and babies in the other ethnic group, respectively, leading to a hypothetical question that percentiles of response densities are significantly different by smoking status and ethnic group. To answer the question, a Bayesian hypothesis test was conducted by comparing whether percentiles of estimated density for the non-smokers (the other ethnic group) are bigger than that for the smokers (the black group) over percentiles of DDT level, with the other continuous covariates fixed at their median values, at each MCMC sample. Based on the results, such differences by smoking status (not shown) and ethnic group (Figure 3.6) are highly significant for both BW and GAD, except at extremely low percentiles of responses and/or at extremely high percentiles of DDT exposure.

Finally, the posterior mean of conditional densities of birth weight given different gestational ages for black babies, whose mother smoked with the median DDT level, are profiled in Figure 3.7. Preterm births (solid line) result in a noticeable shift in the location of the conditional density of BW to the left with more variability, supporting the biological fact that babies with shorter gestation length tend to be smaller and weigh less, because they are delivered before fully maturing. The profile pattern is observed to be similar for other ethnic and smoking groups, with slight differences in locations.

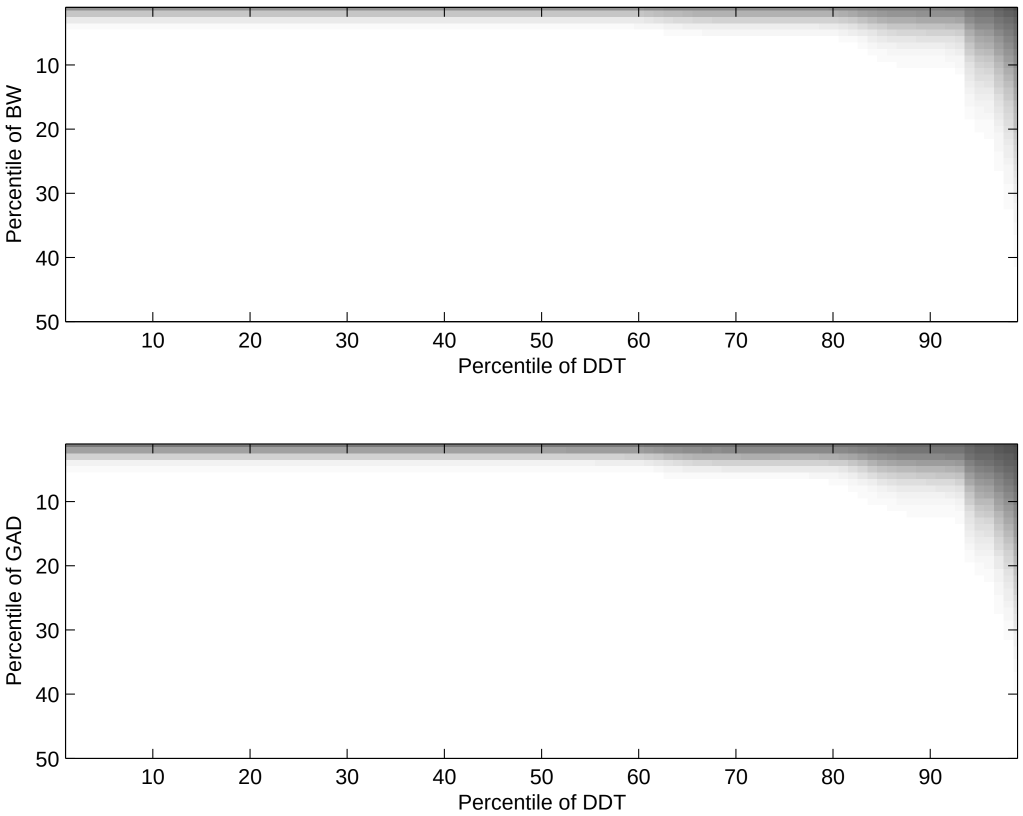

FIGURE 3.6: Heat plot of the marginal probability that the pth percentile of the conditional density of BW (top) and GAD (bottom) for the black group is smaller than that for the other group. Values range from 0 to 1, with 0 for black and 1 for white.

3.7 Discussion

This article has proposed a semiparametric Bayesian latent variable model for density regression, relying on a nonparametric mixture of normal linear measurement models. By taking nonparametric Bayes approaches to modeling of the unknown distribution of latent variable X and of the conditional distribution of Y given X, we successfully relax the two assumptions, linearity and normality, of a typical latent variable model. The proposed method is in part similar to Lee et al. (2008) in that the DP prior is used for the distribution of latent variables. However, our approach differs from the authors in that we avoid a truncation to the DP by implementing a slice sampling algorithm, and furthermore the main difference comes from our use of a flexible mixture model for the conditional distribution of Y given X. In fact, the proposed slice sampling algorithm can be thought of as a different version of Walker's (2007) algorithm,

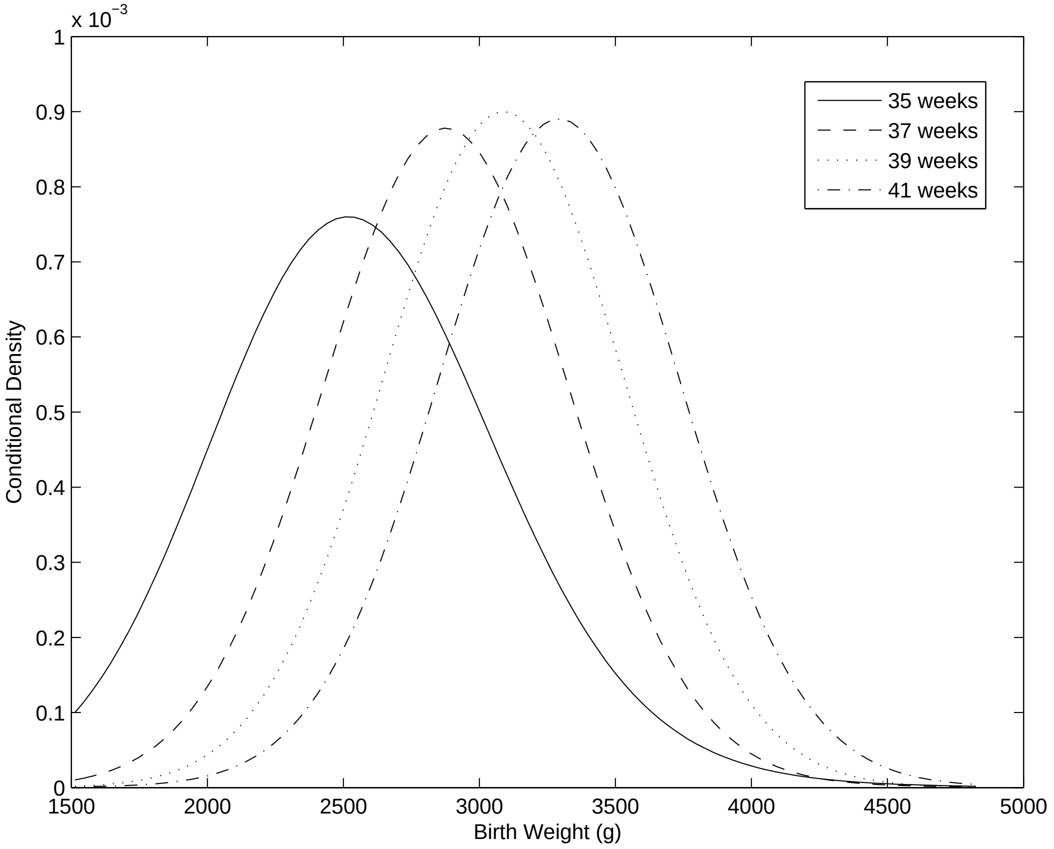

FIGURE 3.7: Profile of estimated conditional densities of BW for different values of GAD, with DDT level fixed at its median.

but provides more flexibility, so that it can be implemented for the KSBP prior with ease.

In cases in which there are more than one latent variables, it should be checked out whether the number of latent variables is smaller than the number of manifest variables, guaranteeing that the corresponding factor loading matrix is of full rank. Assuming a full-rank factor loading matrix, we recommend to constrain the factor loading matrix to be a block lower triangle matrix with diagonals being positive as in Lopez and West (2004). Although we have focused on density regression, the proposed method can be also used for other applications such as prediction or modeling of complex covariance structure.

CHAPTER 4

BAYESIAN SEMIPARAMETRIC DENSITY REGRESSION WITH INFINITE LATENT FACTOR MODELS

4.1 Introduction

There has been an increasing interest in developing approaches for predicting health responses for a patient utilizing not only patient demographics, clinical information and behavioral data but also a high-dimensional set of markers. These markers may consist of metabolite profiles, gene expression, gene sequence or other information. The availability of massive numbers of predictors for a patient, due to the development and streamlining of new biomedical tools, has generated considerable excitement about possibilities in improving patient care through personalized medicine. However, although the possibilities for enormous improvements in patient health are clear, the challenges faced in building accurate predictive models based on massive dimensional predictors are daunting. A fundamental challenge that arises is the well-known large p, small n problem in which the number of predictors typically exceeds the sample size by a substantial margin.

Although a number of strategies for dealing with this problem have been proposed, latent factor approaches (West, 2003; Caron et al., 2006, among many others) are particularly promising. Latent factor methods assume that the massive-dimensional predictors are measurements of a relatively small number of latent variables. In latent factor regression (West, 2003), one can incorporate latent factors underlying the predictors, with these latent factors used as predictors in a model for the health response instead of the measurement predictors. In gene expression applications, such latent factors have been referred to as meta-gene expression levels, and can be thought to have a biological interpretation as genes in a common pathway tend to co-express (Potti et al., 2006)

However, the use of latent variable models as exploratory tools has been criticized in that most of latent variable approaches assume that subjects are forced to have the fixed number of latent variables, which are linearly related to observed variables and normally distributed. In recent years, the literature has focused on relaxing the normality and linearity of latent variable models. Attias (1999) considered a finite mixture of Gaussians for latent factors in factor analysis. Lee et al. (2008) proposed a semiparametric Bayesian model in which latent variables are modeled as a infinite mixture of Gaussians induced by using the Dirichlet process (Ferguson, 1973; 1974) in structural equation models. Alternatively, Dunson et al. (2007) proposed a class of centered stick-breaking processes, which allows the distribution of latent variables to be unknown, while its mean and variance are constrained, as for parametric approaches. Some authors focused on non-linear latent variable models by adding quadratic or interaction terms to linear models (Arminger and Muthén, 1998; Lee and Song, 2003) or by using spline models (Fahrmeir and Raach, 2007;Wang and Iyer, 2007).

Although all these approaches are flexible in some sense, it is still assumed that the number of latent variables are fixed. The problem of allowing uncertainty in the number of latent variables is often considered as a model selection problem and there are a variety of model selection methods. For details, refer to Kass and Raftery (1995), Godsill (2001), and Lopez and West (2004). Alternatively, Ghahramani et al. (2007) accessed the problem in the nonparametric

Bayes perspective by proposing the Indian buffet process (IBP) for unbounded number of latent features.

This article focuses on proposing a semiparamtric Bayes method for conditional density estimation, relying on factor analytic models with the number of factors unknown. There has been increasing work in conditional density estimation (Fan et al., 1996; Hyndman et al., 1996; Fan and Yim, 2004; Dunson et al., 2007; Dunson and Park, 2008). However, it is in doubt whether these methods can produce reliable results accounting for many predictors effectively. Assuming that there are infinitely many factors in the population with a few number of them represented in data at hand, our method allows not only the number of latent factors to be unknown but also the number of factors to vary across subjects.

Section 4.2 discusses about latent factor models and their usages. We propose a nonparametric prior for infinite factors in Section 4.3 and outline an efficient MCMC algorithm in Section 4.4. In Section 4.5, the proposed method is illustrated with simulated examples. The results are discussed in Section 4.6.

4.2 Latent Factor Models

For subject $i = 1, \ldots, n$, let $\mathbf{z}_i = (y_i, \mathbf{x}_i')'$ be a $(p+1) \times 1$ vector of observed variables including a response y_i and p predictors $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$. Then, a latent factor model with k latent factors can be expressed as follows:

$$\begin{aligned} \mathbf{z}_i &= \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{f}_i + \boldsymbol{\epsilon}_i, \\ \mathbf{f}_i &\sim G \end{aligned} \tag{4.1}$$

where $\boldsymbol{\mu}$ is the $(p+1) \times 1$ intercept vector, $\boldsymbol{\Lambda} = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_k)$ is a $(p+1) \times k$ matrix of factor loadings with $\boldsymbol{\lambda}_h$ being a $(p+1) \times 1$ factor loading vector corresponding to the hth factor, $\mathbf{f}_i = (f_{i1}, \ldots, f_{ik})'$ is a $k \times 1$ vector of latent factors, and $\boldsymbol{\epsilon}_i$ is a $(p+1) \times 1$ residual vector with $\boldsymbol{\epsilon}_i \sim \mathrm{N}_{p+1}(\mathbf{0}_{p+1}, \boldsymbol{\Sigma}_\epsilon)$, where $\mathbf{0}_a$ is a $a \times 1$ vector of zeros and $\mathrm{N}_b(\mathbf{c}, \mathbf{B})$ denotes a b-variate normal distribution with mean $\mathbf{c}$ and covariance matrix $\mathbf{B}$. It is assumed that all $\mathbf{f}_i$ and $\boldsymbol{\epsilon}_i$ are mutually independent of each other. In the parametric approach, it is commonly assumed that G corresponds to a multivariate normal distribution, $\mathrm{N}_k(\mathbf{f}_0, \boldsymbol{\Sigma}_f)$, and upon marginalizing out the latent variables, the marginal distribution of $\mathbf{z}_i$ is given by

$$\mathbf{z}_i \sim \mathrm{N}_{(p+1)}(\mathbf{z}_i;\, \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z), \tag{4.2}$$

where $\boldsymbol{\mu}_z = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{f}_0$ and $\boldsymbol{\Sigma}_z = \boldsymbol{\Sigma}_\epsilon + \boldsymbol{\Lambda}\boldsymbol{\Sigma}_f\boldsymbol{\Lambda}'$. Since the model in (4.2) is not identifiable, some parameters are required to be constrained and the most commonly used choice is $\mathbf{f}_0 = \mathbf{0}_k$, $\boldsymbol{\Sigma}_f = \mathbf{I}_k$, and $\boldsymbol{\Sigma}_\epsilon = diag(\tau_1^{-1}, \ldots, \tau_{p+1}^{-1})$, with $\mathbf{I}_b$ being a $b \times b$ identity matrix, implying that $\mathbf{z}_i$ are uncorrelated given latent factors $\mathbf{f}_i$. In addition to these constraints, noting that the above model is invariant to a transformation of $\mathbf{f}_i$ and $\boldsymbol{\Lambda}$ to $\mathbf{f}_i^* = \mathbf{P}\mathbf{f}_i$ and $\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda}\mathbf{P}'$ with any orthogonal matrix $\mathbf{P}$, as known as *rotational ambiguity* (Anderson, 1971; Lopez and West, 2004), factor loadings $\boldsymbol{\Lambda}$ should be also appropriately constrained. Without loss of generality, it is further assumed that $\boldsymbol{\mu} = \mathbf{0}_k$ in the remaining sections.

In fact, the model in (4.1) provides a flexible modeling framework, so that it can be used under two different analytic purposes: confirmatory and exploratory analysis. The model specification that we have discussed so far implicitly assume that all subjects in the study must have the same number of factors that is known a priori and that a particular latent structure is specified through constraints on factor loadings, in which one may incorporate prior knowledge about relationships between observed and latent variables. These assumptions characterize the confirmatory use of the latent factor model. Latent factor models can be also used to explore the underlying factor structure in $\mathbf{z}_i$ by introducing a sequence of latent factors, until there is no common variability among $\mathbf{z}_i$. In this sense, it is reasonable to assume that there are infinitely many factors in the population, with a subject having only a few of them, and that the factors represented by subject i is not necessarily the same as those for subject j. To take these assumptions into account, the model in (4.1) is extended to

$$\mathbf{z}_i = \sum_{h=1}^{\infty} \boldsymbol{\lambda}_h f_{ih} + \boldsymbol{\epsilon}_i. \tag{4.3}$$

On the other hand, from (4.2) the conditional distribution of y_i given $\mathbf{x}_i$ is given by

$$y_i|\mathbf{x}_i \sim \mathrm{N}_p(y_i;\, \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_2),\, \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}),$$

where corresponding to $\mathbf{z}_i = (y_i, \mathbf{x}_i')'$, $\boldsymbol{\mu}_z$ and $\boldsymbol{\Sigma}_z$ are partitioned into

$$\boldsymbol{\mu}_z = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \boldsymbol{\Sigma}_z = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\mu}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

so that one may directly use the model for conditional density estimation, although its use is limited for non-normal cases. Noting that any unknown smooth density can be approximated by a mixture of Gaussians, we want to express the conditional distribution of y_i given $\mathbf{x}_i$ as a mixture of linear regression models upon integrating out latent variables.

4.3 A Nonparametric Prior for Infinite Factors

Noting that latent factors are all continuous, so that $\Pr(f_{ih} = 0) = 0$, it is implicitly assumed that all subjects have the same number of latent factors. To allow each subject to have different number of latent factors, we decompose a latent factor f_{ih} into two latent components S_{ih} and V_{ih}, as in Ghahramani et al. (2007), such that S_{ih} controls the occurrence of the latent variable and V_{ih} represents the value that the latent variable takes on. It is further assumed that independent priors are imposed on them, so that $G(f_{ih}) = G_S(S_{ih})G_V(V_{ih})$.

4.3.1 Proposed Formulation

In order to define a prior for $\mathbf{S}_i = (S_{i1}, \dots, S_{i\infty})$, it is assumed that the latent factors are ordered, so that factors with lower indices are more commonly represented in the population, with the first factor occurring the most, and the probabilities of occurrence of factors stochastically decrease in index h. To take it for account, we consider the following hierarchical model for $\mathbf{S}_i$:

$$S_{ih} |\, \pi_h \overset{iid}{\sim} \text{Bernoulli}(\pi_h)$$

$$\pi_h |\, \gamma_h \overset{ind}{\sim} \text{Beta}(1, \gamma_h),$$

where Bernoulli(a) denotes a Bernoulli distribution with the success probability of a and Beta(b,c) denotes a Beta distribution with mean $b/(b+c)$, and upon marginalizing over the prior of π_h, the marginal probability model for S_{ih} for $h = 1, \dots, \infty$ is

$$\Pr(S_{ih} = 1 |\, \gamma_h) = p_h = \frac{1}{1 + \gamma_h}, \tag{4.4}$$

which is dramatically decreasing in index h by the constraint of $\gamma_1 < \cdots < \gamma_\infty$, resulting in a sparse representation with a few latent factors.

On the other hand, we assign factor scores $\mathbf{V}_i = (V_1, \dots, V_\infty)$ a Dirichlet process (Ferguson, 1973; 1974) prior, which can be expressed as

$$\mathbf{V}_i \sim G_V, \quad G_V \sim DP(\alpha G_{V0}), \tag{4.5}$$

where $DP(\alpha G_{V0})$ denotes a Dirichlet process prior with base measure G_{V0} and precision α, and by the stick-breaking representation of the DP (Sethuraman, 1994), the unknown distribution

G_V can be expressed as

$$G_V = \sum_{h=1}^{\infty} U_h \prod_{l<h}(1-U_l)\delta_{\boldsymbol{\theta}_h}, \quad U_h \overset{iid}{\sim} \text{Beta}(1,\alpha), \quad \boldsymbol{\theta}_h \overset{iid}{\sim} G_{V0}.$$

Letting $C_i = h$ denote a latent variable indicating what latent class subject i belongs to, expression (4.5) can be reexpressed as

$$\mathbf{V}_i = \boldsymbol{\theta}_{C_i}, \quad C_i \sim \sum_{h=1}^{\infty} \pi_h, \quad \boldsymbol{\theta}_h \overset{iid}{\sim} G_{V0},$$

which is a useful form in posterior computation.

Although we allow the distribution of latent factor scores to be unknown, it is still assumed that the factors are uncorrelated and have the same mean and variance by specifying $G_{V0}(\mathbf{V}_i) = \prod_{l=1}^{\infty} G_{V0}(V_{il})$, so that $\text{E}(V_{il}) = \mu_V$, $\text{V}(V_{il}) = \sigma_V^2$, and $\text{Cov}(V_{il}, V_{im}) = 0$ for $l = 1, \ldots, \infty$ and $l \neq m$. Hence, marginalizing over the prior of $\mathbf{V}$ but given $\mathbf{S}_i$, we obtain a Dirichlet process mixture (Lo, 1984; Escobar and West, 1995):

$$f(\mathbf{z}_i|\boldsymbol{\Lambda}, \mathbf{S}_i, \boldsymbol{\theta}, \mathbf{U}) = \sum_{h=1}^{\infty} \pi_h N_{(p+1)}\left(\mathbf{z}_i; \sum_{l=1}^{\infty} \boldsymbol{\lambda}_l S_{il}\theta_{hl}, \boldsymbol{\Sigma}_\epsilon\right), \tag{4.6}$$

where $\boldsymbol{\pi} = \{\pi_h\}_{h=1}^{\infty}$, with $\pi_h = U_h \prod_{l<h}(1-U_l)$.

In addition, our prior specification results that the mean and covariance of $\mathbf{z}_i$ conditional on $\boldsymbol{\Lambda}$ are

$$\begin{aligned}\text{E}(\mathbf{z}_i|\boldsymbol{\Lambda}) &= \sum_{l=1}^{\infty} \boldsymbol{\lambda}_l \text{E}(S_{il}V_{il}) = \mu_V \sum_{l=1}^{\infty} \boldsymbol{\lambda}_l p_l, \\ \text{V}(\mathbf{z}_i|\boldsymbol{\Lambda}, \boldsymbol{\Sigma}_\epsilon) &= \boldsymbol{\Sigma}_\epsilon + \sum_{l=1}^{\infty} \boldsymbol{\lambda}_l \boldsymbol{\lambda}_l' \{p_l \sigma_V^2 + p_l(1-p_l)\mu_V\},\end{aligned}$$

where the second equation is obtained by repeatedly using $\text{V}(XY) = \text{E}\{\text{V}(XY|X)\}+\text{V}\{\text{E}(XY|X)\}$ for any random variables X and Y, and it implies that the marginal mean and covariance of $\mathbf{z}_i$ are finite, when 1) the factor loadings have independent priors but with the same finite mean and covariance, such that $\mathrm{E}(\boldsymbol{\lambda}_l) = \boldsymbol{\lambda}_0$ and $\mathrm{V}(\boldsymbol{\lambda}_l) = \boldsymbol{\Sigma}_\lambda$, and 2) infinite series $\mathbf{p} = \{p_1, \ldots, p_\infty\}$ is convergent.

4.3.2 Finite Truncations

In the nonparametric Bayes literature, finite truncations have been proposed for infinite processes such as the DP and more generally stick-breaking random measures (Ishwaran and James, 2001; 2002), resulting in more efficiency in computational algorithms, while producing almost the same inferential results as the original processes. Since the prior for $\mathbf{S}_i$ results in a sparse representation with a few dominant latent factors, it is appealing to focus on the following truncation approximation to (4.3):

$$\mathbf{z}_i = \sum_{h=1}^{T} \boldsymbol{\lambda}_h f_{ih} + \boldsymbol{\epsilon}_i. \tag{4.7}$$

Before investigating related properties of the truncated model, we let $S_i^T = \sum_{h=T}^{\infty} S_{ih}$ denote the number of represented latent factors, the index of which is greater than or equal to T, with $S_{i,-l}^T$ corresponding to S_i^T with S_{il} deleted. Then we obtain the following lemma.

Lemma 4.3.1. *Let $M_T(t)$ denote the moment generating function (MGF) of S_i^T and let $M_{T,-l}(t)$ denote the MGF of $S_{i,-l}^T$. Then, the MGF of S_i^T is given by*

$$M_T(t) = \prod_{h=T}^{\infty} \{(1 - p_h) + p_h \exp(t)\}$$

and the rth derivative of $M_T(t)$ is

$$M_T^{(r)}(t) = \begin{cases} \sum_{h=T}^{\infty} p_h \exp(t) M_{T,-h}(t) & \textit{for} \quad r = 1 \\ M_T^{(r-1)}(t) + \sum_{h=T}^{\infty} p_h \exp(t) \sum_{l=0}^{r-2} \binom{r-2}{l} M_{T,-h}^{(l+1)}(t) & \textit{for} \quad r > 1, \end{cases}$$

where the superscript in parenthesis indicates the order of the derivative.

The proof of Lemma 4.3.1 is straightforward, using the fact that the exponential function $\exp(t)$ has itself as its derivatives. From Lemma 4.3.1, the expected number of latent factors per subject is $\sum_{h=1}^{\infty} p_h$. In addition, Lemma 4.3.1 implies that if the first moment of S_i^T is finite, in other words, the positive sequence $\mathbf{p}_T = \{p_T, \ldots, p_\infty\}$ is convergent, then so is its rth moment. Based on Lemma 4.3.1, the following theorem holds.

Theorem 4.3.1. *Let $\Delta_i(T) = \sum_{h=T+1}^{\infty} \boldsymbol{\lambda}_h f_{ih}$. If $\{\boldsymbol{\lambda}_l\}_{l=1}^{\infty}$ has independent priors with $E(\boldsymbol{\lambda}_l) = \boldsymbol{\lambda}_0$ and $V(\boldsymbol{\lambda}_l) = \boldsymbol{\Sigma}_\lambda$ and $\mathbf{p}_T$ is convergent with $E(S_i^T)$ being in the order of $1/T$, then Δ_{T+1} converges in probability to zero.*

The proof is in Appendix B. Theorem 4.3.1 says that the truncated model in (4.7) can be used with a moderate value of T instead of (4.3).

4.3.3 A Special Case

Noting that $\mathbf{p}$ should be convergent for our model to hold and the increasing infinite sequence $\boldsymbol{\gamma} = \{\gamma_j\}_{j=1}^{\infty}$ control $\mathbf{p}$ in (4.4), it is an important part of our model specification how to model $\boldsymbol{\gamma}$, and we consider the following model:

$$\gamma_h = \exp\{\psi_0 + \psi_1(h-1)\}, \quad h = 1, \ldots, \infty,$$

which corresponds to a logistic regression model for factor occurrence, and a constraint of $\psi_1 > 0$ ensures that $\boldsymbol{\gamma}$ are strictly increasing in the index h. On the other hand, to increase the efficiency of the truncated model in (4.7) in approximating the original model, it is expected that the probabilities $\mathbf{p}$ of $\mathbf{S}$ quickly level off to zero, after introducing a sufficient number of latent factors enough to explain the shared variability among data. Note that hyperparameter ψ_1 characterizes how fast the probabilities $\mathbf{p}$ decreases in the index h, while ψ_0 does the probability of occurrence of the first factor. We treat ψ_0 as unknown with a prior truncated above at zero and ψ_1 as fixed at some reasonable value, say 3, so that the probability of introducing the first factor is at least 0.5 and $\mathbf{p}$ is convergent. This convergence is easy to show by using the fact that the reciprocals of squared integer numbers are a convergent series and that p_h is bounded above by h^{-2}, for $h = 1, \ldots, \infty$.

4.4 Posterior Computation

In this section, we propose a Monte Carlo Markov Chain (MCMC) algorithm for posterior computation with a focus on a truncated model with $T = 20$ in (4.7). While considering a truncation to factor occurrences for computational convenience, we avoid a truncation to a DP prior for factor scores by using the slice sampler approach of Park and Dunson (2007). The key idea of the slice sampler is to introduce a latent variable u_i to reduce an infinite sum in (4.6) to a finite sum given u_i. For more details, refer to Walker (2007) and Park and Dunson (2007).

To complete a Bayesian specification of the model, we choose the following conjugate priors : $G_{V0}(\mathbf{V}_i) = \prod_{h=1}^{\infty} N(V_{ih}; \mu_V, \sigma_V^2)$, $\pi(\boldsymbol{\Lambda}) = \prod_{h=1}^{\infty} N(\boldsymbol{\lambda}_h; \boldsymbol{\mu}_\lambda, \boldsymbol{\Sigma}_\lambda)$, $\pi(\psi_0) \propto 1(\psi_0 < 0)$, $\pi(\alpha) = \mathcal{G}(\alpha; a_\alpha, b_\alpha)$, and $\pi(\tau_j) = \mathcal{G}(\tau_j; a_\tau, b_\tau)$, for $j = 1, \ldots, p+1$, where $1(\cdot)$ is an indicator function and $\mathcal{G}(\cdot; a, b)$ denotes a Gamma distribution with mean a/b. Note that for additional flexibility, the DP precision is treated as unknown and updated. Then, parameters and latent factors of interest are updated within the following steps of a Gibbs sampling algorithm:

1. Update u_i, for $i = 1, \ldots, n$, by sampling from a uniform distribution within $[0, \pi_T]$,

2. Update $\boldsymbol{\theta}_h$, for $h = 1, \ldots, \infty$ by sampling from the multivariate normal distribution

$$\boldsymbol{\theta}_h \sim N_T(\boldsymbol{\theta}_h; \widehat{V}_{\theta_h}(\sigma_V^{-2}\mu_V \mathbf{1}_T + \sum_{i:C_i=h} \boldsymbol{\Delta}_i' \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{z}_i), \widehat{V}_{\theta_h}),$$

where $\widehat{V}_{\theta_h} = (\sigma_V^{-2}\mathbf{I}_T + \sum_{i:C_i=h} \boldsymbol{\Delta}_i' \boldsymbol{\Sigma}_\epsilon^{-1} \boldsymbol{\Delta}_i)$, $\boldsymbol{\Delta}_i = [S_{i1}\boldsymbol{\lambda}_1, \ldots, S_{iT}\boldsymbol{\lambda}_T]$ is a $(p+1) \times T$ matrix , $\mathbf{1}_a$ is a $a \times 1$ vector of ones, and $\mathbf{I}_b$ is an identity matrix of dimension b.

3. Update U_h, for $h = 1, \ldots, \infty$, by sampling from the beta distribution

$$U_h \sim \text{Beta}\bigg(1 + m_h, \, \alpha + \sum_{l>h} m_l\bigg),$$

where $m_h = \sum_{i=1}^n 1(C_i = h)$, and then update π_h by $\pi_h = U_h \prod_{l<h}(1 - U_l)$, for $h = 1, \ldots, \infty$.

4. Update C_i, for $i = 1, \ldots, n$, by sampling from the multinomial distribution

$$\Pr(C_i = h) \propto 1(h \in A_i(u_i)) N\left\{\mathbf{z}_i; \sum_{l=1}^{T} \boldsymbol{\lambda}_l (S_{il}\theta_{hl}), \boldsymbol{\Sigma}_\epsilon\right\}$$

   where $A_i(u_i) = \{h; \pi_h > u_i\}$ is a finite index subset defined by sampling π_h, for $h = 1, \ldots, k$, with k being the smallest value satisfying

$$\sum_{h=1}^{k} p_h > (1 - u^*),$$

   where $u^* = \min_i\{u_i\}$.

5. Update S_{il}, for $l = 1, \ldots, T$ by sampling from the binomial distribution

$$\Pr(S_{il} = j) \propto 1(j \in \{0, 1\}) N_{p+1}\left\{\mathbf{z}_i; \sum_{h=1, h\neq l}^{T} \boldsymbol{\lambda}_h (S_{ih}\theta_{C_i,h}) + \boldsymbol{\lambda}_l (j\theta_{C_i,l}), \boldsymbol{\Sigma}_\epsilon\right\}.$$

6. Update ψ_0 by a Metropolis-Hastings step and α as in West (1992).

7. Update the rest parameters as follows: for $l = 1, \ldots, T$ and $j = 1, \ldots, p + 1$,

$$\boldsymbol{\lambda}_l \sim N\left[\boldsymbol{\lambda}_l; \widehat{V}_{\lambda_l}\left\{\boldsymbol{\Sigma}_\lambda^{-1}\boldsymbol{\mu}_\lambda + \sum_{i=1}^{n} S_{il}\theta_{C_i,l}\boldsymbol{\Sigma}_\epsilon^{-2}\left(\mathbf{z}_i - \sum_{h=1}^{l-1} \boldsymbol{\lambda}_h S_{ih}\theta_{C_i,h}\right)\right\}, \widehat{V}_{\lambda_l}\right],$$

$$\tau_j \sim \mathcal{G}\left(\tau_j; a_\tau + \frac{n}{2}, b_\tau + \frac{1}{2}\sum_{i=1}^{n}\big\{z_{ij} - (\sum_{l=1}^{T} \lambda_{l,j} S_{il}\theta_{C_i,l)}\big\}^2\right),$$

   where $\widehat{V}_{\lambda_l} = \{\boldsymbol{\Sigma}_\lambda^{-1} + \sum_{i=1}^n (S_{il}\theta_{C_i,l})^2 \boldsymbol{\Sigma}_\epsilon^{-1}\}$.

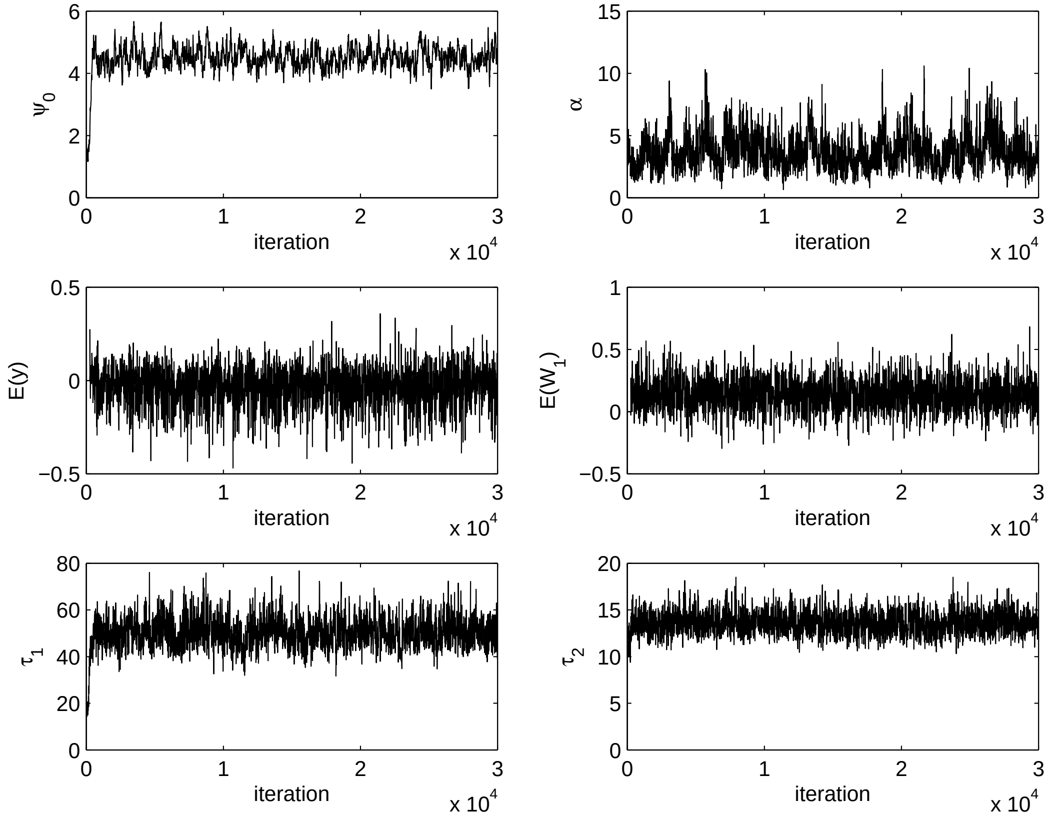

FIGURE 4.1: Trace plots of representative quantities in simulation case 1.

4.5 Simulation Examples

We consider two simulation examples to illustrate the proposed method, while assessing the computational performance of the proposed MCMC algorithm, with a focus on conditional density estimation and prediction. For the hyperparameters, we choose $\mu_V = 0$, $\sigma_V^2 = 1$, $\boldsymbol{\mu}_\lambda = \mathbf{0}_{p+1}$, $\boldsymbol{\Sigma}_\lambda = \mathbf{I}_{p+1}$, $a_\alpha = b_\alpha = 2$, and $a_\tau = b_\tau = 0.1$. For each simulation example, the Gibbs sampler was run for 30,000 iterations, discarding the first 10,000 iterations as burn-in. Since it is assumed that $\boldsymbol{\mu} = \mathbf{0}$, all observed data are centered before analysis is conducted.

In our first simulation, we consider the case in which there are two latent variables w_1 and

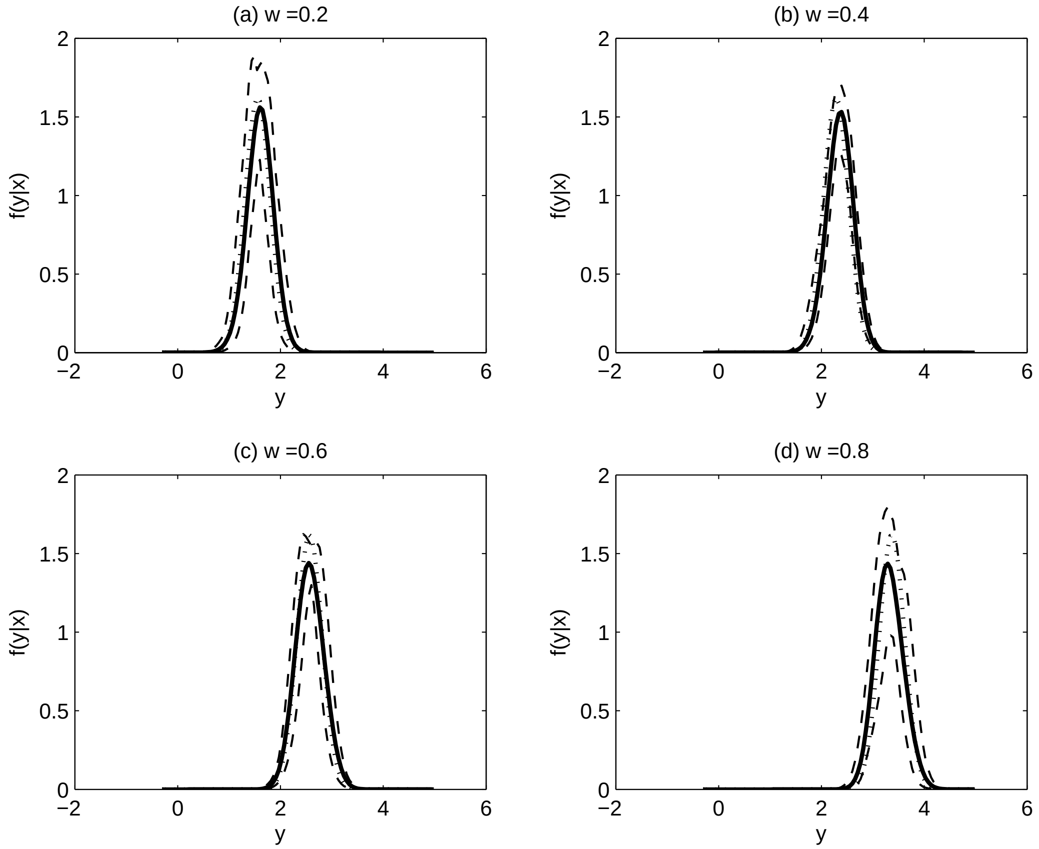

FIGURE 4.2: True conditional density of y given $\mathbf{x}$ (dotted lines), posterior mean estimates (solid lines) and 99% pointwise credible intervals (dashed lines) at the 20th, 40th, 60th, and 80th percentiles of the latent variable w in simulation case 1.

w_2 and observed data are simulated from the following model with $n = 500$ and $p = 120$:

$$\begin{aligned} w_j &\overset{iid}{\sim} \text{Uniform}(0,\ 1), \quad \text{for} \quad j = 1, 2, \\ f(y) &= N(y; 3w_1 + 2w_2, 0.1^2), \\ f(\mathbf{x}) &= N(\mathbf{x}; \boldsymbol{\beta}_1 w_1 + \boldsymbol{\beta}_2 w_2,\ 0.2^2\mathbf{I}_p) \end{aligned} \tag{4.8}$$

where $\boldsymbol{\beta}_1 = (1.2, 0.8, 1.4, \mathbf{0}_9')' \otimes \mathbf{1}_{10}$ and $\boldsymbol{\beta}_2 = (0, 0, 0, 1, 1.2, 0.8, \mathbf{0}_6')' \otimes \mathbf{1}_{10}$, with $\otimes$ denoting the Kronecker product, so that the half of the predictors have no effect on the distribution of y. Figure 4.1 depicts trace plots of selected quantities (ψ_0, α, $E(y)$, $E(x_1)$, $Var(y)$, $Var(x_1)$) showing rapid convergence and good mixing. Figure 4.2 shows the predictive density of a future observation at predictor values that are randomly sampled from the 20th, 40th, 60th and 80th

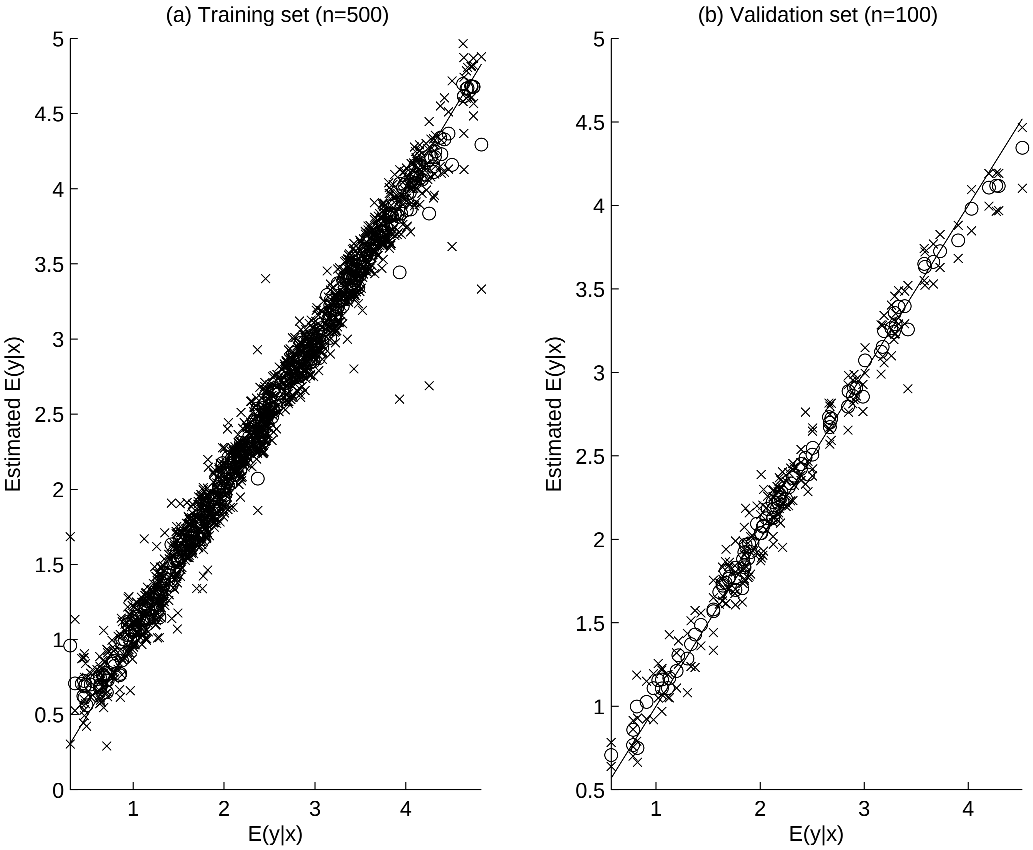

FIGURE 4.3: Estimated conditional mean response (plotted with o) and 99% pointwise credible intervals (plotted with *) along with a diagonal solid reference line in simulation case 1.

percentile of the distribution of x_1 with x_2 being fixed at its median. The true density function (dotted line) is entirely embedded by 99% pointwise credible intervals (dashed lines), indicating good performance of the proposed method for conditional density estimation. To assess the proposed method in terms of prediction, conditional mean responses were estimated not only at the predictor values of the training set (left panel) and but also at predictor values of a validation set of $n = 100$, randomly sampled from the model in (4.8) (right panel), presented in Figure 4. 3. For both sets, all estimated means (plotted with o) except a few well approximate the true means, which are completely contained with 99% credible intervals (plotted with *).

As our second simulation, data were simulated from the following model with $n = 500$ and

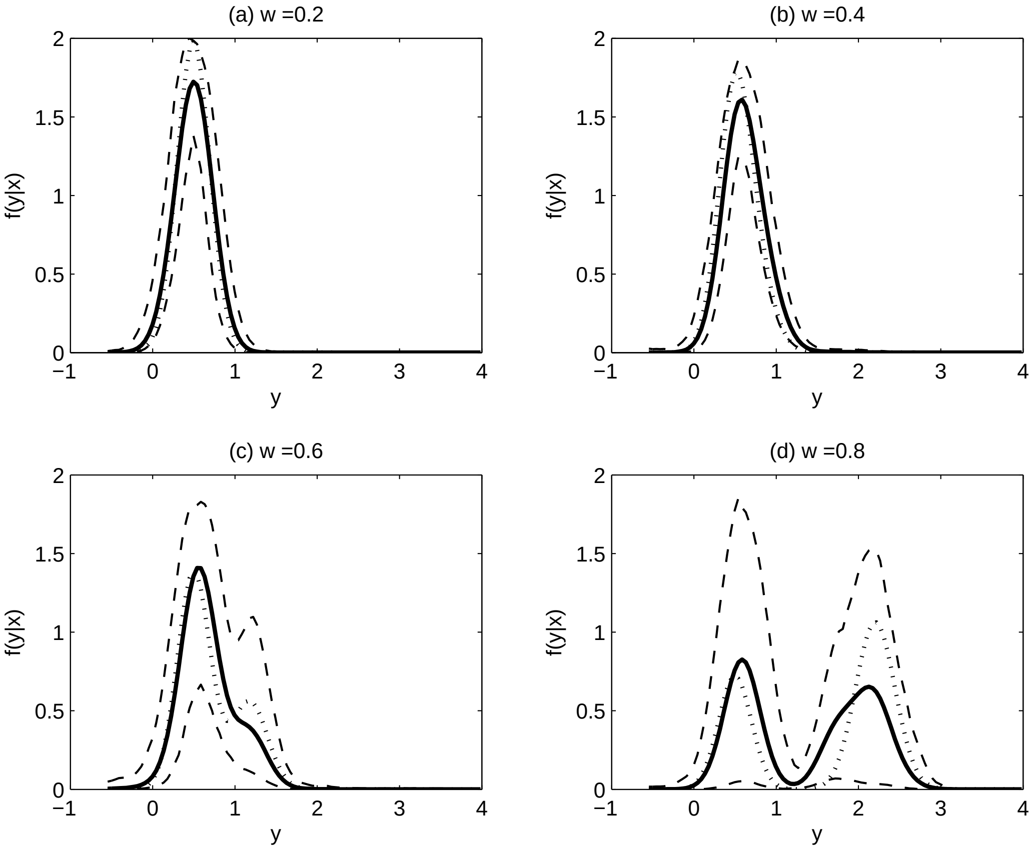

FIGURE 4.4: True conditional density of y given $\mathbf{x}$ (dotted lines), posterior mean estimates (solid lines) and 99% pointwise credible intervals (dashed lines) at the 20th, 40th, 60th, and 80th percentiles of the latent variable w in simulation case 2.

$p = 100$,

$$\begin{aligned} w &\overset{iid}{\sim} \text{Uniform}(0,\ 1) \\ f(y) &= (1-w^2)N(y; 0.5, 0.2^2) + w^2 N(y; (w+0.5)^3, 0.2^2), \\ f(\mathbf{x}) &= N(\mathbf{x}; \boldsymbol{\beta} w,\ 0.2^2\mathbf{I}_p) \end{aligned} \tag{4.9}$$

where $\boldsymbol{\beta} = (1.2, 0.8, 1.4, 1, 0.9, \mathbf{0}_5)' \otimes \mathbf{1}_{10}$ and response y is a non-linear function of latent variable w, while predictors $\mathbf{x}$ are linearly related to w. The results from the second simulation are plotted in Figures 4.4 and 4.5, that are respectively analogous to Figures 4.2 and 4.3, showing that our approach successfully characterizes even a non-linear relationship between responses and predictors. It was observed that the increase in either the sample size or the number of

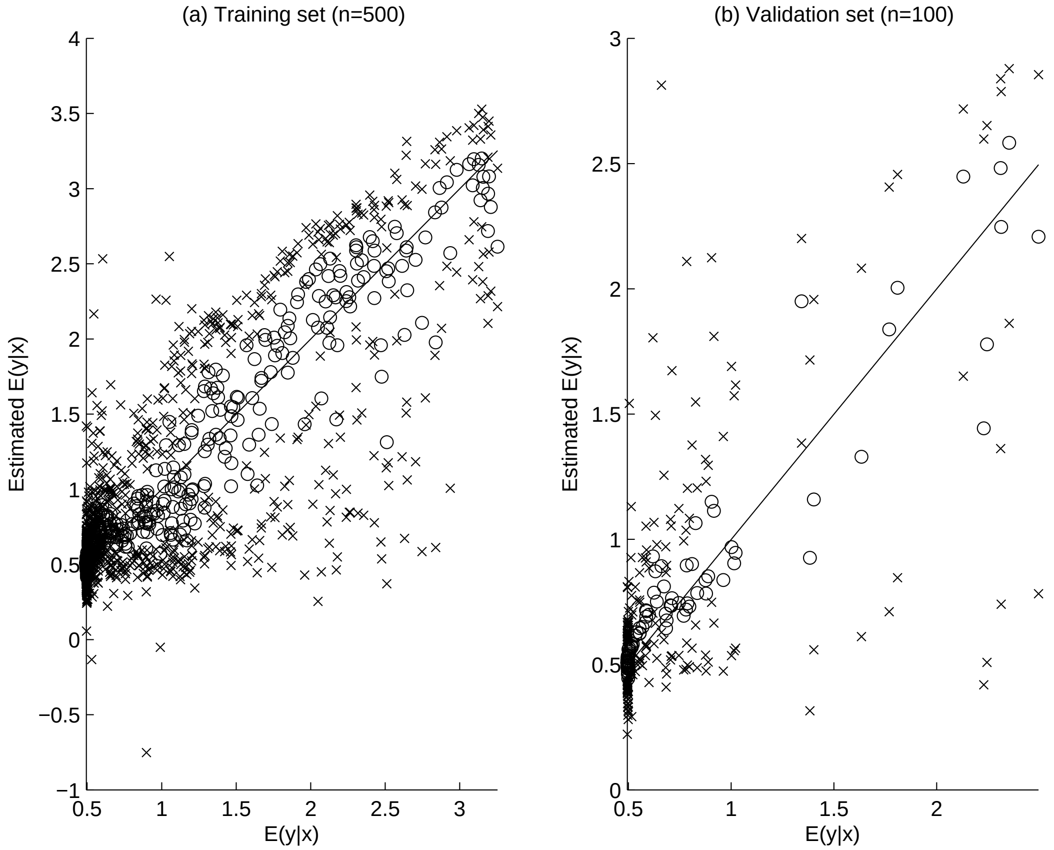

FIGURE 4.5: Estimated conditional mean response (plotted with o) and 99% pointwise credible intervals (plotted with *) along with a diagonal solid reference line in simulation case 2.

predictors related to the response improved posterior inferences, resulting in narrower credible intervals (results not shown).

4.6 Discussion

This article proposed a semiparametric Bayesian approach for conditional density estimation, using factor analysis models as a dimensionality-reduction technique. To treat the number of latent factors as unknown, we defined a prior for infinitely many factors by decomposing a factor into factor occurrence and score and then assigning independent priors to them, and upon marginalizing out over the prior we could obtain a sparse representation for the joint distribution of a response and predictors. In fact, the factor decomposition was first considered by Ghahramani et al. (2007), but there are unique characteristics of our approach. First, our prior model for factor occurrences orders latent factors by their frequency in the population, so that we can avoid possible problems caused by exchangeability of factors, such as the label switching problem (Celeux et al., 2000; Stephens, 2000). In addition, nonparametric modeling of factor scores can result in fewer latent factors represented in data, while providing a good characterization of the data.

Although the proposed method was illustrated only with simulation examples, it is expected that it can have applications to many exploratory studies for different inferences. It is also expected that with appropriate constraints on factor loadings, for example, block diagonal matrix as in Lopez and Mike (2004), the method can be used for other inferences rather than conditional density estimation and prediction.

CHAPTER 5

SUMMARY AND FUTURE RESEARCH

5.1 Summary

This dissertation has proposed semiparametric Bayesian methods for density regression and predictor-dependent clustering. Our focus was on modeling the conditional distribution of a response variable given predictors as an infinite mixture of regression models.

In Chapter 2, we derived a generalized product partition model (GPPM) by incorporating predictors in DP clustering, resulting in a generalized Plya urn scheme. In the application of conditional density estimation, the GPPM provides flexible semiparametric Bayes models while avoiding expensive computation of large numbers of unknowns characterizing priors for dependent collections of random probability measures. Another appealing feature of the GPPM is that for it is easy to implement by modifying Markov chain Monte Carlo (MCMC) algorithms commonly used for the Dirichlet process mixture.

Chapter 3 considered the problem of estimating conditional density estimation in cases, where predictors are not directly observed and multiple surrogates are instead. We proposed a flexible semiparametric Bayesian method that uses nonparametric Bayes approaches to modeling of the unknown distribution of latent variables and of the conditional distribution of responses given latent variables, resulting in relaxation of the normality and linearity. The proposed slice sampler has more flexibility in implementation, so that it showed its good performance not only for the DP prior but also for the KSBP prior. As in the data analysis, our method also allows us to see changes in the conditional distribution of a response given the other responses and latent variables.

Finally, we proposed a black-box type method for density estimation in cases, where there are many predictors but a small number of subjects. Factor analysis models were used to model both a response and predictors, with the number of latent factors unknown. Hence, the conditional distribution of a response given predictors can be derived from the joint distribution, which is obtained upon marginalizing over the prior of latent factors. It was shown that our truncated model produced good results in two simulation examples.

5.2 Future Research

In recent years, there has been an increasing attention to density regression and predictor-dependent clustering, but have not been much contributions in the literature. Hence, it would be of interest to find the theoretical properties of predictor-dependent collections of random probability measures and to develop much faster approaches to computation, so that these methods can be implemented routinely in practice for problems involving complex, large biological data. In this section, we discuss some possible extensions of the proposed methods and areas of future research.

First, as continuation of Chapter 2, we would generalize the use of the GPPM in analyses of more complex data. For example, Quintana et al. (2007) analyzed data consisting of sequences of indicators for loss of heterozygosity (LOH) with three nested levels of repetition: chromosomes for a given patient, regions within chromosomes, and single nucleotide polymorphisms nested within regions. A major draw-back of this approach is the use of fixed, pre-specified regions. We could use a simple location-specific model (e.g., Bernoulli with location-specific probability of LOH) but then cluster locations into regions using the GPPM. As we would be clustering based on both location and LOH probability, the result would be the identification of regions of high and low LOH, which would be quite interesting.

In the situation of Chapter 3, we would alternatively propose to define a prior of an uncountable collection of random probability measures by modifying the KSBP. In the KSBP, conditional probabilities of being assigned to a random location given predictor values are specified directly by sequential products of beta-distributed random variables and a kernel. Instead, the conditional probabilities can rely on the factorization, that is, the conditional probability of being assigned to the hth location given predictor values $\mathbf{x}$ is proportional to the product of the prior probability for the hth location and the likelihood of $\mathbf{x}$ given location h.

APPENDIX A

Proof of Theorem 2.1.1

The Pólya urn scheme in expression (2.5) can be reexpressed with a vector of unique values $\boldsymbol{\theta}^{(i)}$ and configuration $\mathbf{S}^{(i)}$:

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

Then, using expression (2.7), the joint distribution of $\boldsymbol{\phi}$ is

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

where $c_0 = \prod_{i=1}^{n}(\alpha+l-1)^{-1}$, $c(\mathbf{S}_h^{*(i)}) = \alpha(k_h^{(i)}-1)!$, and $1(\cdot)$ is an indicator function. By setting $\phi = (\gamma, \varphi)'$ and doing the same thing to obtain expression (2.11), we can obtain the joint distribution of $\boldsymbol{\varphi}$ and $\mathbf{X}$:

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

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

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

where $\mathbf{X}_m^{(i)} = \{\mathbf{x}_i | i \in \mathbf{S}_m^{*(i)}\}$ and $G_{0\gamma}^*(\gamma|\mathbf{X}_m^{(i)})$ is the posterior distribution of γ updated with the likelihood of $\mathbf{X}_m^{(i)}$. Therefore, the joint distribution of $\boldsymbol{\varphi}$ and $\mathbf{X}$ is simplified as

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

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

APPENDIX B

Proof of Theorem 4.3.1

By the Chebychev's inequality for any real number $\epsilon > 0$,

$$\begin{aligned}
\Pr(|\Delta_i(T)| \geq \epsilon) &\leq \frac{1}{\epsilon^2}\mathrm{V}(\Delta_i(T)) \\
&\leq \frac{1}{\epsilon^2}\mathrm{V}\bigg(\sum_{l=T+1}^{\infty} \boldsymbol{\lambda}_l S_{il} V_{il}\bigg) \\
&\leq \frac{1}{\epsilon^2}\bigg[\sum_{l=T+1}^{\infty} \mathrm{E}(\boldsymbol{\lambda}_l \boldsymbol{\lambda}_l')\{p_l \sigma_V^2 + p_l(1-p_l)\mu_V\} + \sum_{l=T+1}^{\infty} \mathrm{V}(\boldsymbol{\lambda}_l) p_l^2 \mu_V^2\bigg] \\
&< \frac{c}{\epsilon^2(T+1)},
\end{aligned}$$

where $c = \mathrm{E}(\boldsymbol{\lambda}_l \boldsymbol{\lambda}_l')(\sigma_V^2 + \mu_V) + \mathrm{V}(\boldsymbol{\lambda}_l)\mu_V^2$, and therefore $\lim_{T\to\infty} \Pr(|\Delta_i(T)| \geq \epsilon) = 0$.

REFERENCES

Ahmad, I. and Ran, I. (2004). Kernel contrasts: a data-based method of choosing smoothing parameters in nonparametric density estimation. *Nonparametric Statistics* **16**, 671–707.

Aldous, D. (1985). Exchangeability and related topics. In Hennequin, P. L., editor, École d'Été de Probabilités de Saint–Flour XII, volume 1117 of *Springer Lecture Notes in Mathematics*.

Antoniak, C. (1974). Mixtures of dirichlet processes with applications to nonparametric problems. *Annals of Statistics* **2**, 1152–1174.

Arminger, G. and Muthén, B. (1998). A bayesian approach to nonlinear latent variable models using gibbs sampler and the metropolis-hastings algorithm. *Psychometrika* **63**, 271–300.

Attias, H. (1999). Independent factor analysis. *Neural computation* **11**, 803–851.

Banfield, J. and Raftery, A. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics* **49**, 803–821.

Barry, D. and Hartigan, J. (1992). Product partition models for change point problems. *Annals of Statistics* **20**, 260–279.

Bashtannyk, D. and Hyndman, R. (2001). Bandwidth selection for kernel conditional density estimation. *Computational Statistics and Data Analysis* **36**, 279–298.

Bernardo, J. and Smith, A. (1994). *Bayesian theory.* Wiley, New York.

Binder, D. (1978). Bayesian cluster analysis. *Biometrika* **65**, 31–38.

Blackwell, D. and MacQueen, J. (1973). Ferguson distributions via pólya urn schemes.

Bollen, K. (1989). *Structural equations with latent variables.* Wiley, New York.

Bush, C. and MacEachern, S. (1996). A semiparametric bayesian model for randomized block designs. *Biometrika* **83**, 175–185.

Campbell, J., Fraley, C., Murtagh, F. and Raftery, A. (1997). Linear flaw detection in woven textiles using model based clustering. *Pattern Recognition Letters* **18**, 1539–1548.

Caron, F., Davy, M., Doucet, A., Duflos, E. and Vanheeghe, P. (2006). Bayesian inference for dynamic models with dirichlet process mixtures. In *International Conference on Information Fusion*, Florence, Italia.

Carvalho, A. and Tanner, M. (2005). Modeling nonlinear time series with local mixtures of generalized linear models. *Canadian Journal of Statistics* **33**, 97–113.

Carvalho, C., Chang, J., Lucas, J., Nevins, J., Wang, Q. and West, M. High-dimensional sparse factor modelling: Applications in gene expression genomics. *Journal of the American Statistical Association* to appear.

Celeux, G. (1998). *Bayesian inference for mixtures: the label-switching problem*, pages 227–232. Physica, Heidelberg.

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition* **28**, 781–793.

Celeux, G., Hurn, M. and Robert, C. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association* **95**, 957–970.

Crowley, E. (1997). Product partition models for normal means. *Journal of the American Statistical Association* **92**, 192–198.

Dahl, D. (2006). Model-based clustering for expression data via a dirichlet process mixture model. In Do, K.-A., Mueller, P. and Vannucci, M., editors, *Bayesian Inference for Gene Expression and Proteomics*. Cambridge University Press.

Dasgupta, A. and Raftery, A. (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association* **93**, 294–302.

Day, N. (1969). Estimating the components of a mixture of normal distributions. *Biometrika* **56**, 463–474.

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

Diebolt, J. and Robert, C. (1994). Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society, Series B* **56**, 363–375.

Dunson, D. (2007). Empirical bayes density regression. *Statistica Sinica* **17**, 481–504.

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

Dunson, D. and Peddada, S. (2008). Bayesian nonparametric inference on stochastic ordering. *Biometrika* to appear.

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

Dunson, D., Yang, M. and Baird, D. (2007). Semiparametric bayes hierarchical models with mean and variance constraints. Technical report, Duke University.

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.

Fahrmeir, L. and Raach, A. (2007). A bayesian semiparametric latent variable model for mixed responses. *Psychometrika* **72**, 327–346.

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. *Annals of Statistics* **1**, 209–230.

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

Fraley, C. and Raftery, A. (1998). How many clusters? which clustering method? answers via model-based cluster analysis. *Computer Journal* **41**, 578–588.

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

Freedman, D. (1999). On the bernstein-von mises theorem with infinite-dimensional parameters. *Annals of Statistics* **27**, 1119–1140.

Friedman, H. and Rubin, J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association* **62**, 1159–1178.

Ge, Y. and Jiang, W. (2006). On consistency of bayesian inference with mixtures of logistic regression. *Neural Computation* **18**, 224–243.

Gelfand, A., Kottas, A. and MacEachern, S. (2005). Bayesian nonparametric spatial modeling with dirichlet process mixing. *Journal of the American Statistical Association* **100**, 1021–1035.

Geweke, J. and Keane, M. (2007). Smoothly mixing regressions. *Journal of Econometrics* **138**, 252–290.

Ghahramani, Z., Griffiths, T. and Sollich, P. (2007). Bayesian nonparametric latent feature models. In *Baysian Statistics*, 8.

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

Ghosh, J. and Ramamoorthi, R. (2003). *Bayesian nonparametrics.* Springer Verlag, New York.

Godsill, S. (2001). On the relationship between markov chain monte carlo methods for model uncertainty. *Journal of Computational and Graphical Statistics* **10**, 230–241.

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

Griffiths, T. and Ghahramani, Z. (2005). Infinite latent feature models and the indian buffet process. Technical Report 2005–001, Gatsby Computational Neuroscience Unit.

Hall, P., Wolff, R. and Yao, Q. (1999). Methods for estimating a conditional distribution function. *Journal of the American Statistical Association* **94**, 154–163.

Hanson, T. (2006). Inference for mixtures of finite polya tree models. *Journal of the American Statistical Association* **101**, 1548–1565.

Hanson, T. and Johnson, W. (2002). Modeling regression error with a mixture of pólya trees. *Journal of the American Statistical Association* **97**, 1020–1033.

Hartigan, J. (1990). Partition models. *Communications in Statistics, Part A – Theory and Methods* **19**, 2745–2756.

Holmes, C., Denison, D., Ray, S. and Mallick, B. (2005). Bayesian prediction via partitioning. *Journal of Computational and Graphical Statistics* **14**, 811–830.

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

Hyndman, R. and Yao, Q. (1998). Nonparametric estimation and symmetry tests for conditional density functions. Technical Report 17/98, Department of Econometrics and Business Statistics, Monash University.

Ishwaran, H. and Jaems, L. (2003). Generalized weighted chinese restaurant processes for species sampling mixture models. *Statistica Sinica* **13**, 1211–1235.

Ishwaran, H. and James, L. (2001). Gibbs sampling methods for stick–breaking priors. *Journal of the American Statistical Association* **96**, 161–173.

Ishwaran, H. and Zarepour, M. (2000). Markov chain monte carlo in approximate dirichlet and beta two-parameter process hierarchical models. *Biometrika* **87**, 371–390.

Jain, S. and Neal, R. (2004). A split–merge markov chain monte carlo procedure for the dirichlet process mixture model. *Journal of Computational and Graphical Statistics* **13**, 158–182.

Jasra, A., Holmes, C. and Stephens, D. (2005). Markov chain monte carlo methods and the label switching problem in bayesian mixture modelling. *Statistical Science* **20**, 50–67.

Jiang, W. and Tanner, M. (1999). Hierarchical mixtures-of-experts for exponential family regression models: Approximation and maximum likelihood. *Annals of Statistics* **27**, 987–1011.

Jordan, M. and Jacob, R. (1994). Hierarchical mixures of experts and the em algorithm. *Neural Computation* **6**, 181–214.

Jusko, T., Koepsell, T., Baker, R., Greenfield, T., Willman, E., Charles, M., Teplin, S., Checkoway, H. and Hertz-Picciotto, I. (2006). Maternal ddt exposures in relation to fetal and 5-year growth. *Epidemiology* **17**, 692–700.

Kass, R. and Raftery, A. (1995). Bayes factor. *Journal of the American Statistical Association* **90**, 773–795.

Ker, A. and Ergün, A. (2005). Empirical bayes nonparametric kernel density estimation. *Statistics and Probability Letters* **75**, 315–324.

Kurihara, K., Welling, M. and Vlassis, N. (2006). Accelerated variational dirichlet mixture models. *Advances in Neural Information Processing Systems* **19**.

Lau, J. and Green, P. (2007). Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics* **16**, 526–558.

Lavine, M. (1992). Some aspects of pólya tree distributions for statistical modelling. *Annals of Statistics* **20**, 1161–1176.

Lavine, M. (1994). More aspects of pólya tree distributions for statistical modelling. *Annals of Statistics* **22**, 1222–1235.

Lee, S., Lu, B. and Song, X. (2008). Semiparametric bayesian analysis of structural equation models with fixed covariates. *Statistics in Medicine* **27**, 2341–2360.

Lee, S. and Song, X. (2003). Model comparison of nonlinear structural equation models with fixed covariates. *Psychometrika* **68**, 27–47.

Lijoi, A., Prünster, I. and Walker, S. (2005). On consistency of nonparametric normal mixtures for bayesian density estimation. *Journal of the American Statistical Association* **100**, 1292–1296.

Lo, A. (1984). On a class of bayesian nonparametric estimates i. density estimates. *Annals of Statistics* **12**, 351–357.

Longnecker, M., Klebanoff, M., Zhou, H. and Brock, J. (2001). Association between maternal serum concentration of the ddt metabolite dde and preterm and small-for-gestational-age babies at birth. *Lancet* **358**, 110–114.

Lopez, H. and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica* **14**, 41–67.

MacEachern, S. (1994). Estimating normal means with a conjugate style dirichlet process prior. *Communications 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. (2001). Decision theoretic aspects of dependent nonparametric processes. In E, G., editor, *Bayesian Methods with Applications to Science, Policy and Official Statistics*, pages 551–560. Creta: ISBA.

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.

Mukerjee, S., Feigelson, E., Babu, G., Murtagh, F., Fraley, C. and Raftery, A. (1998). Three types of gamma ray bursts. *Astrophysical Journal* **508**, 314–327.

Muliere, P. and Secchi, P. (1995). A note on a proper bayesian bootstrap. Technical Report 18, Universitá degli Studi di Pavia, Dipartamento di Economia Politica e Metodi Quantitativi.

Muliere, P. and Tardella, L. (1998). Approximating distributions of random functionals of ferguson-dirichlet priors. *Canadian Journal of Statistics-Revue Canadienne de Statistique* **26**, 283–297.

Müller, P., Erkanli, A. and West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika* **83**, 67–79.

Müller, P. and Quintana, F. (2004). Nonparametric bayesian data analysis. *Statistical Science* **19**, 95–110.

Murtagh, F. and Raftery, A. (1984). Fitting straight lines to point patterns. *Pattern Recognition* **17**, 479–483.

Naik, P., Shi, P. and Tsai, C. (2007). Extending the akaike information criterion to mixture regression models. *Journal of the American Statistical Association* **102**, 244–254.

Navarrete, C., Quintana, F. and M̈ller, P. (2008). Some issues on nonparametric bayesian modeling using species sampling models. *Statistical Modelling International Journal* to appear.

Neal, R. (2003). Slice sampling. *Annals of Statistics* **31**, 705–767.

Paddock, S., Ruggeri, F., Lavine, M. and West, M. (2003). Randomized polya tree models for nonparametric bayesian inference. *Statistica Sinica* **13**, 443–460.

Papaspiliopoulos, O. and Roberts, G. (2007). Retrospective markov chain monte carlo methods for dirichlet process hierarchical models. *Biometrika* to appear.

Park, J.-H. and Dunson, D. (2007). Bayesian generalized product partition model. *Biometrika* submitted.

Pitman, J. (1996). Some developments of the blackwell–macqueen urn scheme. In Ferguson, T., Shapley, L. and MacQueen, J., editors, *Statistics, Probability and Game Theory*, volume 30 of *IMS Lecture Notes–Monograph series.*

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

Potti, A., Sayan Mukherjee, R. P., Dressman, H., Bild, A., Koontz, J., Kratzke, R., Watson, M., Kelley, M., Ginsburg, G., West, M., Harpole, D. and Nevins, J. (2006). A genomic strategy to refine prognosis and therapeutic decision for adjuvant therapy in non-small cell lung carcinoma. *New England Journal of Medicine* **355**, 570–580.

Qin, L.-X. and Self, S. (2006). The clustering of regression models method with applications in gene expression data. *Biometrics* **62**, 526–533.

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, Series B* **65**, 557–574.

Quintana, F. and Newton, M. (2000). Computational aspects of nonparametric bayesian analysis with applications to the modeling of multiple binary sequences. *Journal of Computational and Graphical Statistics* **9**, 711–737.

Redner, R. and Walker, H. (1984). Mixture densities, maximum likelihood and the em algorithm. *SIAM Review* **26**, 195–239.

Richardson, S. and Green, P. (1997). On bayesian analysis of mixtures with unknown number of components with discussion. *Journal of the Royal Statistical Society, Series B* **59**, 731–792.

Rodrigues, A., Dunson, D. and Gelfand, A. (2007). Nonparametric functional data analysis through bayesian density estimation. *Biometrika* revision submitted.

Roeder, K. and Wasserman, L. (1997). Practical bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association* **92**, 894–902.

Rosenblatt, M. (1969). Conditional probability density and regression estimators. In Krishnaiah, P., editor, *Multivariate Analysis*, pages 25–31. Academic Press, New York.

Sánchez, B., Butdz-Jørgensen, E., Ryan, L. and Hu, H. (2005). Structural equation models: A review with applications to environmental epidemiology. *Journal of the American Statistical Association* **100**, 1443–1455.

Scheines, R., Hoijtink, H. and Boomsma, A. (1999). Bayesian estimation and testing of structural equation models. *Psychometrika* **64**, 37–52.

Scott, A. and Symons, M. (1971). Clustering methods based on likelihood ratio criteria. *Biometrics* **27**, 387–397.

Sethuraman, J. (1994). A constructive definition of dirichlet priors. *Statistica Sinica* **4**, 639–650.

Spiegelhalter, D., Best, N., Carling, B. and Van der linge, A. (2002). Bayesian measures of model complexity and fit with discussion. *Journal of the Royal Statistical Society, Series B* **64**, 585–616.

Stephens, M. (1997). Discussion on 'on bayesian analysis of mixtures with an unknown number of components' by s, richardson and p.j. green. *Journal of the Royal Statistical Society, Series B* **59**, 768–769.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society, Series B* **62**, 795–809.

Symons, M. (1981). Clustering criteria and multivariate normal mixtures. *Biometrics* **37**, 35–43.

Viele, K. and Tong, B. (2002). Modeling with mixtures of linear regressions. *Statistics and Computing* **12**, 315–330.

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

Walker, S., Lijoi, A. and Prünster, I. (2007). On rates of convergence for posterior distributions in infinite-dimensional models. *Annal of Statistics* **35**, 738–746.

Wang, H. and Iyer, H. (2007). Application of local linear embedding to nonlinear exploratory latent structure analysis. *Psychometrika* **72**, 199–225.

Ward, J. (1963). Hierarchical groupings to optimize an objective function. *Journal of the American Statistical Association* **58**, 234–244.

West, M. (1992). Modelling with mixtures. In Berger, J., Bernardo, J., Dawid, A. and Smith, A., editors, *Bayesian Statistics*, volume 4. Oxford University Press, Oxford.

West, M. (2003). Bayesian factor regression models in the large p, small n paradigm. In Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A. and West, M., editors, *Bayesian Statistics*, 7, pages 723–732. Oxford University Press, Oxford.

West, M., Müller, P. and Escobar, M. (1994). Hierarchical priors and mixture models, with applications in regression and density estimation. In Smith, A. and Freeman, P., editors, *A Tribute to D. V. Lindley*. John Wiley and Sons.

Yeung, K., Fraley, C., Murua, A., Raftery, A. and Ruzzo, L. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics* **17**, 977–987.