BAYESIAN NONPARAMETRIC METHODS FOR CONDITIONAL DISTRIBUTIONS

Suprateek Kundu

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.

Chapel Hill
2012

Approved by:

Dr. David B. Dunson
Dr. Pranab K. Sen
Dr. Michael R. Kosorok
Dr. Hongtu Zhu
Dr. Carolyn T. Halpern

Abstract

SUPRATEEK KUNDU: BAYESIAN NONPARAMETRIC METHODS FOR CONDITIONAL DISTRIBUTIONS
(Under the direction of Dr. David B. Dunson and Dr. Pranab K. Sen)

In the first paper, we propose a flexible class of priors for density estimation avoiding discrete mixtures, based on random nonlinear functions of a uniform latent variable with an additive residual. Although discrete mixture modeling has formed the backbone of the literature on Bayesian density estimation incorporating covariates, the use of discrete mixtures leads to some well known disadvantages. We propose an alternative class of priors based on random nonlinear functions of a uniform latent variable with an additive residual. The induced prior for the density is shown to have desirable properties including ease of centering on an initial guess for the density, posterior consistency and straightforward computation via Gibbs sampling.

In the second paper, we propose a Bayesian variable selection method involving non-parametric residuals, noting that the majority of literature has focused on the parametric counterpart. We generalize methods and asymptotic theory established for mixtures of g-priors to linear regression models with unknown residuals characterized by DP location mixture. We propose a mixture of semiparametric g-priors allowing for straightforward posterior computation via a stochastic search variable selection algorithm. In addition, Bayes factor and variable selection consistency is shown to result under a class of proper priors on g allowing the number of candidate predictors p to increase much faster than sample size n while making sparsity assumption on the true model size.

Our third paper is motivated by the fact that although there are standard algorithms for estimating minimum length credible intervals for scalars, there are no such methods for estimating minimum volume credible sets for vectors and functions. We propose a minimum volume covering ellipsoids (MVCE) approach for vector valued parameters, guaranteed to construct credible regions with probability $\geq 1 - \alpha$, while yielding highest posterior density regions under asymptotic normality. For one-dimensional random curves, our proposed approach starts with a MVCE region evaluated at finitely many knots, and then interpolates between the knots linearly or relying on Lipschitz continuity. For multivariate random surfaces, our approach uses Delaunay triangulations to approximate the credible region. Frequentist coverage properties and computational efficiency compared with frequentist alternatives are assessed through simulation studies.

This work is dedicated to the Peaceful Warriors.

Acknowledgments

I would like to thank my advisor Dr. David Dunson for his enormous support and guidance at every juncture of my dissertation. I would also like to thank my co-advisor Dr. Pranab K. Sen for his constant motivation and encouragement. A special thanks to Dr. Michael Kosorok for pointing me towards literature on minimum volume covering ellipsoids. I am grateful to Dr. Hongtu Zhu for his insights into my work on simultaneous credible regions. Last but not the least, thanks to Dr. Carolyn Halpern for her insights into potential applications of the proposed methodology.

My doctoral career has been a wonderful period of self discovery, unexpected friendships and unforgettable experiences. There have been quite a few people who have made innumerable contributions to my growth and success, and although it would never be possible to acknowledge everyone, I would like to give it my earnest try. I am grateful to my friends in Chapel Hill and beyond who have supported me through thick and thin. A special mention to Drs Rinku and Samarpan Majumder, Arpita Ghosh, Santanu Pramanik for their support and friendship. I also owe a lot to the talented and generous doctoral students in the department of Biostatistics at UNC, who have helped me time and again, and not just academically.

I am also thankful to the incredible YES+ ! group who are an amazingly fun and wonderful group of people. A special mention to Nirupama and Ananth Shankar

for their warm friendship and guidance. I am indebted to the Center for Peace and Mediation in Boone, NC for providing a nourishing place to grow and revitalize. Last but not the least, I am indebted to my father Mr. Bibek Kumar Kundu, my mother Mrs. Chandana Kundu and my sister Debashruti Kundu for their love and support and having an unwavering confidence in me, even when I had none in myself.

Preface

Bayesian nonparametrics is a rapidly expanding area in terms of methodological and theoretical developments, and is being successfully used in an increasing number of applications including, but not limited to, density estimation, density regression, survival analysis, hierarchical models and model validation, and more recently model selection techniques. These models are used to avoid critical dependence on parametric assumptions and to robustify parametric models.

The contribution of my dissertation is to develop very general nonparametric Bayes methods which can be used in a wide range of applications incorporating covariates, and which are shown to have appealing theoretical justifications. I have worked on three fundamental problems in statistical methodology and have proposed solutions based on a Bayesian nonparametrics paradigm. These problems include probability density estimation and density regression, variable selection in linear models with nonparametric residuals and constructing simultaneous credible regions for vectors and infinite dimensional functions, guaranteed to contain posterior probability of at least $1 - \alpha$.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Literature Review and Motivation

1.1.1 Latent Factor Models for Density Estimation

In the first paper of my dissertation, we propose a flexible class of priors for density estimation based on random nonlinear functions of a uniform latent variable with an additive residual. It is well known that nonparametric kernel mixture models are increasingly popular in density estimation, density regression and high dimensional data modeling. Kernel mixture models for density estimation have the form

$$f(y; G) = \int \mathcal{K}(y; \theta) G(d\theta), \tag{1.1}$$

where $G(\cdot)$ is a mixing distribution and $\mathcal{K}(\cdot)$ is a probability kernel. The majority of the nonparametric Bayesian development in this area relies on Dirichlet process (DP) priors (Ferguson, 1973; 1974) for G. A DP on $(\chi, \mathcal{A})$ with parameter α is essentially a stochastic process where, for any measurable paritition $(A_1, \ldots, A_k)$ of χ, the random vector $(P(A_1), \ldots, P(A_k))$ follows a Dirichlet distribution with parameters $(\alpha(A_1), \ldots, \alpha(A_k))$. The seminal paper by Ferguson (1973) shows that if P is a DP on $(\chi, \mathcal{A})$ with parameter α, and $X_1, \ldots, X_n$ are a sample from P, then the posterior

distribution of P given $X_1, \ldots, X_n$ is also a DP on $(\chi, \mathcal{A})$ with parameter $(\alpha + \sum_1^n \delta_x)$. Ghosal, Ghosh and Ramamoorthi (1999) provided general conditions in terms of L1 metric entropy to ensure strong posterior consistency and verified those conditions for Dirichlet process location mixtures of normal kernels under certain regularity conditions. Tokdar (2006) extended their result to the location-scale mixture case while encompassing a significantly larger class of 'true' densities. Sethuraman (1994), gave a constructive definition of DP of the form

$$P = \sum_{j=1}^{\infty} w_j \alpha_j, \quad \alpha_j \sim N(0, \tau^{-1}), \quad w_j = \nu_j \prod_{l<j} (1 - \nu_l), \quad \nu_l \sim Beta(1, m),$$

which enabled posterior computation involving DP kernel mixtures to become much more manageable. The weights w_j are called stick-breaking weights as they can be obtained by repeatedly breaking a stick of initial length 1 into proportions ν_h and $1 - \nu_h$, and continuing the proces with the fraction $1 - \nu_h$. Pati, Dunson and Tokdar (2011) has also shown large support and weak/strong posterior consistency for a broad class of generalized stick-breaking priors, which includes the Chung and Dunson (2009) probit stick-breaking prior as a special case. Walker (2007) proposed the slice sampler which greatly improved computational speed. His method relies on augmentation with uniform latent variables as follows

$$f_{w,\alpha}(y) = \sum_{j \in B_w(u)} N(y|\alpha_j), \quad B_w(u) = \{j : w_j > u\}.$$

Thus the development of DP kernel mixture approaches have addressed the primary two requirements for any prior distribution for nonparametric problems - (i) large support on the set of densities, and (ii) analytical tractability facilitating posterior computation. However, the current approaches have largely avoided addressing one other critical issue central to prior specification - interpretability. For example, we would like to center

our prior on a prior guess for the density.

These models have been generalized to density regression by defining dependence on the covariates x in various ways. If the covariates have a finite number of levels the Product of Dirichlet processes model introduced by Cifarelli and Regazzini (1978) allows the modelling of dependent distributions. Dependence is introduced through the use of a parametric regression model as the centering distribution of independent Dirichlet processes at each level of the covariates. Müller, Elkanli and West (1996) used a DP mixture of multivariate normals to jointly model the density of the response and predictors to induce a prior on $f(y|x)$. Their development boils down to a local linear regression of the form $\mathrm{E}(y|x,\theta)=\sum_{j=0}^{k} s_j(x)m_j(x)$, where $m_j(x)$ is the mean of the jth component distribution of y given x, which is linear due to the normality of the kernels. The regression weights $s_j(x)$ determine that components $m_j(x)$ will be more highly weighted in predicting y when the value of the density $f_j(x|\theta)$ is relatively large. In order to let the parameters of the DP vary over the predictor space $\mathcal{X}$, MacEachern (1999) defined dependent Dirichlet processes (DDP) by assigning stochastic processes on the components in Sethuraman's (1994) DP representation: $G_x = \sum_{i=1}^{\infty} p_i(x)\delta_{\theta_i(x)}$. This ensures that the parameters of the non-parametric prior are able to adapt across the parameter space, thus giving it more flexibility. De Iorio et al. (2004) proposed a fixed-p DDP in ANOVA models, while Griffin and Steel (2006) introduce dependence in nonparametric distributions by making the weights in the Sethuraman (1994) representation dependent on the covariates. They modeled each weight as a transformation of i.i.d. random variables and implemented the dependence by inducing an ordering of these random variables at each covariate value such that distributions for similar covariates values will be associated with similar orderings and, thus, be close. At any covariate value, the random distribution would be a so-called stick-breaking prior, and they focused on the special case where they assigned DP for the stick breaking prior.

Dunson, Pillai and Park (2007) instead used predictor-dependent convex combinations of DP components. More specifically, they introduced kernel stick breaking processes of the form

$$G_x = \sum_{h=1}^{\infty} U(x; V_h, \Gamma_h) \prod_{l<h} \Big(1 - U(x; V_l, \Gamma_l)\Big) G_h^*$$
$$U(x; V_h, \Gamma_h) = V_h K(x, \Gamma_h) \text{for all x } \in \mathcal{X}.$$

Here $V_h \sim Be(a_h, b_h)$, $K_h : \Re^p \times \Re^p \to [0, 1]$ is a bounded kernel function and G_h^* is the base measure located at Γ_h. Thus G_x is a predictor-dependent mixture over an infinite sequence of basis probability measures. Such a construction can encourage sparsity and borrowing of information across $\mathcal{X}$ through careful choice of hyper-parameters and kernels K_h.

There is also a rich literature on using mixture priors including the class of DP priors in hierarchical latent variable models to effectively characterize high dimensional multivariate distributions, with a focus on sparse covariance structure modelling. Bush and MacEachern (1996) proposed a Bayesian semiparametric model for randomised block experiments. Their model is a hierarchical model in which a Dirichlet process is inserted at the middle stage for the distribution of the block effects, thus allowing for an arbitrary distribution of block effects, and resulting in effective estimates of treatment contrasts, block effects and the distribution of block effects. Kleinman and Ibrahim (1998) put a DP on the random effects in a longitudinal random effects model. Brown and Ibrahim (2003) proposed a semiparametric model for joint modeling of longitudinal and survival data. Fokoue and Titterington (2003) and Fokoue (2005) incorporated a finite mixture of normal factor model in a mixture of factor analyzers (MFA), allowing for unknown number of mixture components and common factors. To elaborate, Factor Analysis (FA) is a well established probabilistic approach to unsupervised learning for

complex systems involving correlated variables in high-dimensional spaces. FA aims principally to reduce the dimensionality of the data by projecting high-dimensional vectors on to lower-dimensional spaces. However, because of its inherent linearity, the generic FA model is essentially unable to capture data complexity when the input space is nonhomogeneous. A finite Mixture of Factor Analysers (MFA) is a more flexible extension of the basic FA model that overcomes the above limitation by assigning a mixture distribution to the latent factors. The structure of the MFA model offers the potential to model the density of high-dimensional observations adequately while also allowing both clustering and local dimensionality reduction. Chen et al. (2009) and Carvalho et al. (2008) proposed nonparametric Bayes MFA where they allowed an uncertain number of factors by placing DP and Beta process priors respectively on the the number of factors. On the other hand, Dunson (2006) used dynamic mixtures of DPs to allow a latent variable distribution to change nonparametrically across groups. Lee, Lu, and Song (2008) placed a truncated DP on the distribution of the exogenous latent variables within a structural equation model (SEM). In order to ensure identifiability and interpretability in SEM, Yang and Dunson (2010) modelled the exogenous variable using centered Dirichlet processes (CDP) (Yang et al., 2010) in a latent class model and CDP mixtures in a latent trait model. To review, the SEM is specified using two components, (1) the measurement model, which relates the measurement variables to latent variables; and (2) the latent variable or structural model, which describes relationships among the latent variables, typically through a linear structural relations or LISREL model.

The above approaches relying on discrete mixture models, have a number of well known complications motivating alternative methods for modeling unknown densities, such as Polya trees (Mauldin et al, 1992; Lavine, 1992, 1994) and logistic Gaussian processes (LGP) (Lenk 1988, 1991; Tokdar 2007). The Polya tree generates random

probabilities G, such that for any partitioning subset B_ϵ ($\epsilon = (\epsilon_1, \ldots, \epsilon_m)$), $G(B_\epsilon) = \prod_{j=1,\epsilon_j=0}^{m} Y_{\epsilon_1,\ldots,\epsilon_{j-1}0} \prod_{j=1,\epsilon_j=1}^{m} (1-Y_{\epsilon_1,\ldots,\epsilon_{j-1}0})$, where $Y_{\epsilon 0} \sim Be(\alpha_{\epsilon 0}, \alpha_{\epsilon 1})$. Polya trees have appealing properties in terms of denseness, conjugacy and posterior consistency but have disadvantages in terms of favoring overly spiky densities. On the other hand LGP's are defined as $f_w(t) = \frac{e^{w(t)}}{\int e^{w(s)} ds}$, where $w(\cdot)$ is assigned a GP prior. The smoothness properties of the GP transfers on to the LGP, thus rendering it sound theoretical properties and control over the smoothness of the densities through the covariance kernel in the GP. However, posterior computation is a major hurdle. Recently, Jara and Hanson (2010) proposed dependent tail-free processes where they modeled the tail-free probabilities with LGP dependent on covariates. Their approach is shown to approximate the Polya tree marginally at each predictor value. An alternative was suggested by Tokdar, Zhu and Ghosh (2010) relying on LGP for density regression with dimensionality reduction. More specifically, they modeled the conditional density by equivalently modelling quantities such as $p(G_0(y)|\mathbf{F(z)})$ using LGP, where G_0 is any cumulative distribution function, $\mathbf{F}$ is a d-dimensional function having monotonically increasing components from $\Re$ to (-1,1) and $\mathbf{z}$ belongs to the d-dimensional central subspace of the predictor space $\mathcal{X}$.

In our first paper, we focus on a new approach for nonparametric density estimation and regression that induces a prior on the unknown density through placing a flexible prior on a nonlinear regression function θ in a latent factor model. The proposed class of models is related to Gaussian process latent variable models (GP-LVM) proposed in the machine learning literature (Lawrence, 2005; Silva and Gramacy, 2010), but our modeling details are different and the focus of this literature has been on nonlinear dimensionality reduction with no consideration of density estimation or associated properties. By using GP priors for θ, we obtain substantial control over the smoothness of the induced densities in a very different manner than that achieved by LGP-based

models. Unlike LGP-based models, the proposed model has conjugacy properties facilitating posterior computation. In addition, the method has appealing theoretical properties in terms of large support and posterior consistency.

Relative to some density estimation priors, the proposed latent factor approach is quite easy to generalize to more challenging settings involving multivariate densities, conditional density estimation, hierarchical modeling and other complexities. Although our primary focus in this article is to introduce the formulation, providing a basic intuition for how the model works, basic properties and computation, we also give a flavor of generalizations through a simple conditional density estimation example. In particular, we consider a model that induces a prior on the conditional density $f(y|x)$ through joint modeling of the response and predictors through separate nonparametric latent factor models containing the same latent variables. This formulation is completely flexible in the marginal densities, while making strong restrictions on the dependence to address the curse of dimensionality in a related manner to a copula model. An attractive feature of our model is that it naturally allows for incorporation of prior information on the marginal densities of response and predictors through the mean function of the GP.

1.1.2 BVS in Semiparametric Linear Models

BVS or Bayesian variable selection is very widely applied, with a rich literature on alternative priors and computational methods. For a recent review of Bayesian variable selection methods, refer to O'Hara and Sillanpää (2009). Most of the literature has focused on Gaussian linear regression models, with common methods including stochastic search variable selection (SSVS) (George and McCulloch, 1993; 1997), reversible jump MCMC (Green, 1995) and adaptive shrinkage (Tibshirani, 1996; Park and Casella, 2008; Yi and Xu, 2008). SSVS puts the prior on effect sizes as

$P(\beta_j|I_j) = (1 - I_j)N(0, \tau^2) + I_j N(0, g\tau^2)$, where I_j is the variable inclusion indicator and the first density is centered around 0 and has a small variance. The specification of parameters τ, g are data-dependent. Alternatively, SSVS could involve a point mass at 0 instead of $N(0, \tau^2)$ in the above formulation. Reversible jump MCMC is a flexible technique for model selection, which lets the Markov chain explore spaces of different dimensions. For variable selection, the positions (indices) of the selected variables are defined as $l_1, \ldots, l_{N_v}$, and the model is updated by randomly selecting variable j and then proposing either addition to ($N_v := N_v + 1$) or deletion from ($N_v := N_v - 1$) the model of the corresponding effect. The length of the parameter vector is therefore not fixed but varies during the estimation. The updating is done using a Metropolis-Hastings algorithm, but with the acceptance ratio adjusted for the change in dimension. The degree of sparseness can be controlled by setting the prior for N_v. On the other hand, shrinkage methods work by shrinking values of the effect sizes towards zero (or equivalently, concentrating the likelihood towards 0) if there is no evidence in the data for significant effects. Conversely, there should be practically no shrinkage for data-supported values of covariates that are non-zero. In practice, it is often the case that these adaptive shrinkage methods select a super set of the important set of predictors. The method is adaptive in the sense that the degree of sparseness is data driven, through the way it shrinks the covariates effects towards zero. The degree of sparseness of the model can be adjusted by changing the prior distribution of the precision of the effect sizes, with an exponential distribution amounting to Laplacian shrinkage. The degree of sparseness in Laplacian shrinkage is controlled by the scale parameter of the exponential prior, which when assigned a hyperprior renders the Bayesian Lasso (Park and Casella, 2008). Such methods can be applied directly for kernel or basis function selection in nonlinear regression with Gaussian residuals (Smith and Kohn, 1996) and can be adapted to accommodate generalized linear models with outcomes in

the exponential family (Raftery and Richardson 1993; Meyer and Laud 2002).

It is well known that Bayesian variable selection can be sensitive to the prior, and there is an increasingly rich literature showing asymptotic properties providing support for carefully-chosen priors, such as mixtures of g-priors (Zellner and Siow, 1980; Liang et. al., 2008), with such priors also having appealing computational properties. This literature is essentially entirely focused on Gaussian linear regression models, and the emphasis of this article is on developing methods that generalize this work to semiparametric regression models having unknown residual distributions.

To set the stage, first consider the well-studied problem of comparison of linear models of the following type:

$$\begin{aligned} M_1 : Y^n &= \alpha 1_n + X_{\gamma_1}\beta_{\gamma_1} + \epsilon_1, \quad \epsilon_1 \sim N(0, \tau^{-1} I_n), \\ M_2 : Y^n &= \alpha 1_n + X_{\gamma_2}\beta_{\gamma_2} + \epsilon_2, \quad \epsilon_2 \sim N(0, \tau^{-1} I_n), \end{aligned} \tag{1.2}$$

where Y^n is n×1 vector of responses, α is the common intercept, X_{γ_j} is a n×p_j design matrix (j=1,2) excluding the column of intercepts, and ϵ_j's are Gaussian residuals, j=1,2. The models may or may not be nested, and the number of candidate predictors is p. Among numerous model selection criteria available for such comparisons, the Bayes factor (Kass and Raftery, 1995) has received substantial attention as the most widely accepted Bayesian measure of the weight of evidence in the data in favor of one model over another. The Bayes factor for comparing M_1 versus M_2 based on a sample Y^n is defined as $\mathrm{BF}_{12}^n = \frac{L(Y^n|M_1)}{L(Y^n|M_2)}$, the ratio of marginal likelihoods under M_1 and M_2. Assuming one of the models under comparison is true, Bayes factor consistency refers to the phenomenon where $\mathrm{BF}_{12}^n \xrightarrow{P} \infty$ as $n \to \infty$ under M_1 and $\mathrm{BF}_{12}^n \xrightarrow{P} 0$ as $n \to \infty$ under M_2. A stronger form of consistency is also possible when the convergence happens almost surely. When comparing the true model pairwise to each model in a list, Bayes

factor consistency typically implies that the posterior probability on the true model goes to one.

Although priors most commonly used in practice assume *a priori* independence in the elements of the coefficient vectors (β_1 and β_2), priors that have been shown to result in Bayes factor consistency typically incorporate dependence. Examples include the intrinsic prior (Berger and Pericchi, 1996; Moreno, 1997; Moreno, Bertolino and Racugno, 1998) which builds a prior for the alternative model with varying degrees of concentration around the null, and Zellner's g-prior (Zellner, 1986) specified by $\beta_j \sim N(0, g\tau^{-1}(X_j'X_j)^{-1})$, j=1,2. The intrinsic priors have proven to behave very well for multiple testing problems (Casella and Moreno, 2006). Moreno and Girón (2005) showed consistency for the intrinsic Bayes procedure for nested linear models, while Casella (2009) extended consistency for the intrinsic Bayes procedure to non-nested linear models. Zellner's g-prior allows for a convenient correlation structure and can control for the amount of prior information relative to the sample through only one hyperparameter g. Among others, Fernández et al. (2001) investigated Bayes factor consistency under various choices of fixed g, which was allowed to depend on the sample size and/or the number of candidate predictors. In order to resolve difficulties associated with a fixed choice of g, such as Bartlett's paradox (Bartlett, 1957; Jeffreys 1961) and information paradox (Zellner 1986; Berger and Pericchi 2001), Zellner and Siow (1980) placed an inverse-gamma prior on g, while Liang et. al. (2008) extended the idea of Strawderman (1971) to the regression context by proposing hyper-g and hyper-g/n priors on g, under which they established Bayes factor consistency. To review, Bartlett's paradox refers to the fact that in the limiting case when $g \to \infty$ while (n, p_γ) are fixed, the Bayes factor for comparing M_γ to the null model will go to 0. That is, large spread of the prior induced by the non-informative choice of g has the unintended consequence of forcing the Bayes factor to favor the null model, the smallest model, regardless of

the information in the data. On the other hand, the information paradox refers to the fact that the Bayes factor in favor of M_γ goes to a constant for a fixed choice of g as the coefficient of determination of M_γ goes to 1 (i.e. when there is overwhelming evidence in favor of M_γ,), keeping (n, p_γ) fixed. This is against conventional wisdom, as one expects the Bayes factor in favor of M_γ to go to ∞ as evidence against the null model accumulates. Coming back to the afore-mentioned approaches, they entail specifying improper priors on common model parameters and proper priors on model parameters unique to any one model, which results in a prior specification for the more complex model depending upon the simpler model. To avoid such pitfalls, Guo and Speckman (2009) adopted the idea of Marin and Robert (2007) and placed mixtures of $g-$priors on all the elements of both β_1 and β_2, which leads to tractable Bayes factors as well as Bayes factor consistency.

There has also been a growing interest in model selection procedures for normal linear models when the number of candidate predictors (p) increase with sample size (n). Such increases occur in a wide variety of applications, such as in nonparametric regression when the number of candidate kernels or basis functions depends on n. Shao (1997) analyzed the consistency of several frequentist and Bayesian approximation criteria for model selection in normal linear models with increasing model dimensions, assuming the true model to be the submodel minimizing the average squared prediction error. Moreno et. al. (2010) examined consistency of Bayes factors and the BIC under intrinsic priors for nested normal linear models, when the dimension of the parameter space increases with the sample size. Jiang (2007) considered Bayesian variable selection in generalized linear models in $p > n$ settings and provided conditions to obtain near optimal rates of convergence in estimating the conditional predictive distribution, but did not consider asymptotic properties in selecting the important predictors.

To our knowledge, this area has entirely focused on parametric models with a

particular focus on normal linear regression. Such a parametric assumption on the residual error is rather stringent and may not hold in practice, thus invalidating the earlier assumption of the true model belonging to the class of models under comparison and potentially leading to inconsistent Bayes factors. Simulations illustrate that when residuals are generated from a bimodal distribution, Bayesian variable selection under a Gaussian linear regression model tends to have poor performance. With this motivation, our focus is on developing Bayes variable selection methods that do not require Gaussian residuals and that can be shown to be consistent.

There is a limited literature on variable selection in Bayesian regression models having unknown residual distributions. Kuo and Mallick (1997) consider an accelerated failure time model for time-to-event data containing a linear regression component and a mixture of Dirichlet processes for the residual density. To perform variable selection, they add indicator variables to the regression function and implement an MCMC algorithm. Also, in the survival analysis setting, Dunson and Herring (2005) proposed a Bayesian approach for selecting predictors in a semiparametric hazards model that allows uncertainty in whether predictors enter in a multiplicative or additive manner. More specifically, to accommodate this uncertainty, they placed a model selection prior on the coefficients in an additive-multiplicative hazards model. This prior assigned positive probability, not only to the model that has both additive and multiplicative effects for each predictor, but also to sub-models corresponding to no association, to only additive effects, and to only proportional effects, and further, they constrained the additive component of the model to ensure non-negative hazards. Kim, Tadesse and Vannucci (2006) instead define a Bayesian variable selection approach, which uses a Dirichlet process to define clusters in the data, while updating the variable inclusion indicators using a Metropolis scheme. They introduced a latent binary vector to identify discriminating variables and used Dirichlet process mixture models to define

the cluster structure. They updated the variable selection index using a Metropolis algorithm and obtained inference on the cluster structure via a split-merge Markov chain Monte Carlo technique. Mostofi and Behboodian (2007) model a symmetric and unimodal residual density using a Dirichlet process scale mixtures of uniforms, while conducting Bayesian variable selection by selecting the highest posterior probability model. Chung and Dunson (2009) modeled the conditional response density given predictors using a flexible probit stick-breaking mixture of Gaussian linear models, allowing variable selection via a Bayesian stochastic search method. More specifically, they introduced the probit stick-breaking process (PSBP) as a prior for an uncountable collection of predictor-dependent random probability measures and proposed a PSBP mixture (PSBPM) of normal regressions for modeling the conditional distributions. They incorporated a global variable selection structure so as to discard unimportant predictors, while allowing estimation of posterior inclusion probabilities. Further, they did local variable selection while relying on the conditional distribution estimates at different predictor points.

These articles focused on defining methodology and computational algorithms, but without study of theoretical properties, such as consistency. In fact, to our knowledge, there has been no previous work on consistent Bayesian variable selection in semiparametric models, though there is recent work on consistent non-parametric Bayesian model selection (Ghosal, Lember and van der Vaart, 2008 among others). Ghosal, Lember and van der Vaart (2008) considered nonparametric Bayesian estimation of a probability density based on a random sample of size n from this density using a hierarchical prior. They presented a general theorem on the rate of contraction of the resulting posterior distribution as $n \to \infty$ which gives conditions under which the rate of contraction is the one attached to the model that best approximates the true density of the observations. This shows that, for instance, the posterior distribution can adapt

to the smoothness of the underlying density. They also studied the posterior distribution of the model index, and found that under the same conditions the posterior distribution gives negligible weight to models that are bigger than the optimal one, and thus selects the optimal model or smaller models that also approximate the true density well. However, it is not straightforward to apply such theory directly to the problem of variable selection in semiparametric linear models.

With this motivation, we define a practical, useful and general methodology for Bayesian variable selection in semiparametric linear models, while providing basic theoretical support by showing Bayes factor and variable selection consistency. We accomplish this by generalizing the methods and asymptotic theory for mixtures of g-priors to linear regression models with unknown residuals characterized via Dirichlet process (DP) location mixture of Gaussians. We propose a new class of mixtures of semiparametric g-priors under a family of proper priors for g, which results in consistent Bayesian variable selection even when there are many more candidate predictors (p) than samples (n) as long as the prior assigns probability zero to models having greater than or equal to n predictors. Additionally, posterior computation for the proposed method is straightforward via a SSVS algorithm.

1.1.3 Bayesian credible regions for vectors and functions

In the Bayesian paradigm, posterior uncertainty is commonly summarized using credible regions containing $1 - \alpha$ posterior probability, and it is appealing to obtain the minimum volume region. When the posterior density is unimodal, the minimum volume credible set corresponds to a highest posterior density (HPD) region. For scalar parameters, HPD intervals can be estimated easily based on draws from the posterior obtained using Markov chain Monte Carlo (MCMC) (Chen and Shao, 1999). However, such methods are only appropriate for scalars and there are no general use methods

for estimating HPD regions for vectors or functions. Current state-of-the-art methods for Bayesian estimation of credible regions focus on using either large sample elliptical regions that can be justified by asymptotic normality of the posterior or rectangular regions that inflate the size of scalar regions for each parameter using multiplicity adjustments. Given the routine use of credible sets in practice, there is a clear need for improved methods for efficiently estimating minimum volume credible sets for vectors and functions based on MCMC output.

To adjust for conservatism while maintaining a rectangular region, Crainiceanu et al. (2007) assumed approximate posterior normality and proposed inflated hyper-rectangular regions as

$$\left[\hat{\theta}_j - c_{1-\alpha}\sqrt{\widehat{var}(\theta_j)}, \hat{\theta}_j + c_{1-\alpha}\sqrt{\widehat{var}(\theta_j)}; j = 1, \ldots, p\right],$$

where $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)'$ is the posterior mean of the parameter of interest $\theta = (\theta_1, \ldots, \theta_p)'$, and $c_{1-\alpha}$ is the $1-\alpha$ percentile of $\max_{j=1,\ldots,p} \left|\frac{\theta_j^s - \hat{\theta}_j}{\sqrt{\widehat{var}(\theta_j)}}\right|$ with θ^s corresponding to the sth MCMC draw from the posterior.

Unfortunately, restricting consideration to rectangular regions ignores the true topology of the joint posterior and can lead to $100(1-\alpha)\%$ credible regions that can have dramatically larger volume than necessary. As an alternative for estimating convex credible regions based on MCMC samples, we initially considered convex hull peeling in which one starts with the convex hull of the MCMC samples and peels off outer layers by discarding outer points until obtaining a region containing $100(1-\alpha)\%$ of the samples. Although such an approach is promising, even fast algorithms for calculating the convex hull of a set of points have substantial computational burden. For example, the worst case complexity of the quick-hull algorithm (Barber et. al, 1996) is $O(nf_r/r)$ for $p > 3$, where r is the number of processed points and f_r is the maximum number of faces for r vertices. The number of processed points is much larger than p, implying that

$f_r = O(r^{\lfloor p/2 \rfloor}/\lfloor p/2 \rfloor)$ (Klee, 1966) increases rapidly as p increases, so that computation becomes unmanageable quickly. Hence, we do not consider such an approach further.

In developing an alternative, we use the equivalence between HPD regions, minimum volume (MV) sets and density level sets of the posterior to ascertain the subset of MCMC samples falling within the HPD region. MV sets (Polonik, 1995, 1997) summarize regions with a pre-assigned probability content where the mass is most concentrated, and are closely related to density level sets (Tsybakov, 1997; Ben-David and Lindenbaum, 1997; Cuevas and Rodriguez-Casal, 2003; Steinwart et al., 2005). The main difference is that the latter requires the specification of a density level of interest instead of the probability mass to be enclosed. The equivalence between MV sets and density level sets was established by Nunez-Garcia et al. (2003) under some reasonable conditions.

Assuming that the posterior $g(\theta|Y^n) \propto h(\theta, Y^n) = \pi(\theta)L(Y^n|\theta)$ ($L(\cdot)$ being the likelihood) is known up to a normalizing constant, we can exploit the aforementioned equivalence to assert that any posterior MCMC sample θ^j, j=1,. . . ,J, satisfying

$$h(\theta^j, Y^n) > \lambda, \quad P\{\theta : h(\theta, Y^n) > \lambda | Y^n\} = 1 - \alpha, \tag{1.3}$$

will belong to the $100(1-\alpha)\%$ HPD credible region. An estimate for λ can be obtained from the MCMC samples as $\hat{\lambda}$ such that $\frac{\#\left\{\theta^j : h(\theta^j, Y^n) \geq \hat{\lambda}\right\}}{J} \geq 1 - \alpha$. To allow for additional parameters ψ in the model without requiring a known analytic form for the marginal $h(\theta, Y^n) = \int L(Y^n|\theta, \psi)\pi(\theta|\psi)\pi(\psi)d\psi$, we can rely on an approximation, with a simple choice corresponding to the Monte Carlo approximation $\hat{h}(\theta, Y^n) = 1/H \sum_{h=1}^{H} L(Y^n|\theta, \psi^h)\pi(\theta|\psi^h)$, where ψ^hs ($h = 1, \ldots, H$) are realizations from the prior $\pi(\psi)$. Such an approximation is not possible under improper priors for ψ, and may be inefficient when the prior is very diffuse relative to the likelihood, but there is a rich literature proposing alternatives.

We propose a method for constructing elliptical credible regions which enclose the $100(1-\alpha)\%$ HPD set defined as the collection of posterior MCMC samples satisfying (1.3), with the estimated density level $\hat{\lambda}$. In the presence of nuisance parameters, we replace $h(\theta, Y^n)$ with $\hat{h}(\theta, Y^n)$. In cases when such an approximation is poor, the HPD set may contain samples outside of the $100(1-\alpha)\%$ HPD region, thereby potentially yielding slightly inflated finite dimensional credible regions. Elliptical regions provide a convenient approximation, with the exact credible region having an elliptical form under multivariate normality of the posterior. Bernstein von Mises theorems guarantee asymptotic normality for sufficiently regular parametric models, with similar results arising in certain semi- or nonparametric models (Castillo, 2012; Rivoirard and Rousseau, 2009). Our approach utilizes minimum volume covering ellipsoids (MVCE) (Rousseeuw 1985), which have been implemented in a variety of application areas (Knorr et al., 2001; Kumar and Orlin, 2008). Typically approximate algorithms are used (Khachiyan, 1996; Kumar and Yildirim, 2005; Sun and Freund, 2004), as exact computation is often not feasible.

Another major focus of our paper is constructing credible regions for functions. Although there is a rich literature on Bayesian semi- and nonparametric methods, with recent theoretical results on properties of credible regions for functions (Knapik, van der Vaart and van Zanten, 2011), there is a surprising lack of methods for calculating such regions in practice. Most commonly one reports pointwise intervals for the function at different locations or for functionals of interest. However, this clearly provides an inadequate characterization of uncertainty in the function as a whole. Simultaneous credible regions have useful applications in hypothesis testing for random functions. For example, we might conclude that a curve is well approximated by a parametric function if the parametric curve falls entirely within the estimated credible region. In addition, there is often interest in assessing whether the region includes a flat line

or surface corresponding to a null hypothesis under consideration. In nonparametric regression, such a null hypothesis may correspond to there being no effect of a predictor or predictors of interest. In other settings, the null corresponds to there being no difference in group-specific surfaces.

For one-dimensional curves, we propose two distinct approaches for computing credible bands by computing credible limits at a grid of points or knots and subsequently interpolating between these limits using linear interpolation and a method based on Lipschitz continuity. The proposed methods are a simple add on to existing MCMC algorithms and are not computationally expensive for moderate number of knots. In interpolating, the hope is that the posterior will assign small probability to the set of *aberrant* curves, which are contained within the credible bands at the knots but fall outside the bands somewhere between the knots. If this is not the case, the posterior probability contained in the estimated credible set may be less than $1-\alpha$. When the support of the posterior corresponds to Lipschitz continuous functions, we show that one can use a modified type of linear interpolation that appropriately inflates the width of the intervals between knots to ensure that the posterior places probability zero on the set of aberrant curves. For functions over higher dimensional surfaces, we construct piecewise hyperplanar credible surfaces by generalizing the linear interpolation approach through Delaunay triangulations.

Chapter 2

Latent Factor Models for Density Estimation

2.1 Model Specification

Initially suppose $y_i \in \Re$ are iid draws from an unknown density $f \in \mathcal{F}$, where $\mathcal{F}$ is the set of densities on $\Re$ with respect to Lesbesgue measure. We propose to induce a prior $f \sim \Pi$ through

$$\begin{aligned} y_i &= \mu(x_i) + \epsilon_i, \quad \epsilon_i \sim \Gamma_\sigma, \\ \mu \sim \Pi^*, \quad \sigma &\sim \nu, \quad x_i \sim \text{Uniform}(0,1), \end{aligned} \tag{2.1}$$

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

$$f(y; \mu, \sigma) = f_{\mu,\sigma}(y) = \int_0^1 \Gamma_\sigma(y - \mu(x))dx. \tag{2.2}$$

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

$$f(y) = \int_0^\infty \int_\Theta \int_0^1 \Gamma_\sigma(y - \mu(x))dx\Pi^*(d\mu)\nu(d\sigma). \tag{2.3}$$

Hence, a prior $f \sim \Pi$ is induced through assigning independent priors to μ and σ in expression (2.2). When the prior on μ is a Gaussian process and the error distribution is $N(0, \sigma^2)$ (denoted as $\Gamma_\sigma = \phi_\sigma$), we refer to $f_{\mu,\sigma}$ as a Gaussian process transfer (GPT) model and the induced prior $f \sim \Pi$ as a GPT prior.

The GPT prior does not have the kernel mixture form (1.1). There will be no clustering of subjects or label switching issues. Instead, the prior $f \sim \Pi$ is induced through adding a Gaussian residual to a Gaussian process regression model in a uniform latent variable. This is a simple structure aiding computation and interpretability. One can control the smoothness of the density through the covariance in the GP prior for the regression function μ and the size of the scale parameter σ. In limiting cases, one can obtain realizations of μ concentrated close to a flat line, leading to a normal density as a special case. In addition, by making σ small and choosing the GP covariance to generate a very bumpy μ, one can obtain arbitrarily bumpy densities. In practice, by choosing hyperpriors for key covariance parameters, we obtain a data adaptive approach that often outperforms discrete kernel mixtures. The performance of discrete kernel mixtures relies on the ability to accurately approximate the density with few components, and DP mixtures tend to heavily favor a small number of dominate kernels. This tendency can sometimes lead to relatively poor estimation, as illustrated in section 2.7.

2.2 Prior Specification

Prior elicitation is an important aspect of Bayesian modeling, with the prior playing a particularly important role in Bayesian nonparametric models involving infinitely

many parameters. Most of the Bayesian nonparametrics literature relies on default priors, which do not reflect available prior knowledge in a particular application area, but are chosen to lead to good performance in terms of posterior behavior in a wide variety of applications. However, as in parametric models, well chosen informative priors that utilize information, such as historical data on the variables under study, can substantially improve the performance in small to moderate samples. In DP mixtures, such prior information is typically incorporated through choice of hyperparameters in the base measure, while maintaining conjugacy for ease in computation. For example, in Gaussian kernel mixtures, a normal-inverse gamma base measure would be chosen having parameters representing prior knowledge. This is appropriate when prior knowledge implies that the density follows a t distribution, but when one has prior information that the density follows a more complex form (as in our premature delivery application) then elicitation is substantially more difficult. Obtaining a base measure that leads to a particular elicited density is a deconvolution problem, which can be difficult to solve for non-atomic base measures. In addition, posterior computation under the resulting complex and non-conjugate base measure may be challenging. An advantage of the GPT is that the prior for the density can be centered on an arbitrary choice easily through the prior mean in the GP prior for μ.

To elaborate, Theorem 3 (section 2.3) ensures that $\mu \approx \tilde{\mu} = \tilde{F}^{-1}$ implies that $f_{\mu,\sigma} \approx f_{\tilde{\mu},\sigma} = \tilde{f}$, with $\tilde{f} = \frac{d}{dy}\tilde{F}$ and $\sigma \approx 0$. In terms of application, this translates to incorporating a prior guess $\tilde{f}$ for the density through the corresponding mean function $\tilde{\mu} = \tilde{F}^{-1}$ of the GP, and letting the prior for σ to have mode near 0. Such a mean function can be constructed by obtaining frequentist kernel estimates of the concerned density using some external data, and then converting it into an inverse cdf on a grid of points in [0,1] (using a linear approximation). Thus, the characteristics of the entire density is captured through $\tilde{F}^{-1}$ as the mean function of the GP, and we let the data

influence the deviation of the posterior from the prior guess.

These ideas are demonstrated in Figure A.1, where we use some earlier data on gestational age at delivery to construct prior densities. We choose a Ga(25,1) prior for the residual precision, and different sets of hyper-parameters for the covariance kernel of the GP. The frequentist kernel estimates were obtained by the bandwidth selection method of Sheather and Jones (1991), using a Gaussian kernel ('kernel' function in R). It is evident that the smoothness as well as the degree of deviation of the prior from the frequentist estimate can be controlled through the hyper-parameters in the covariance kernel of the GP.

2.3 Theoretical Properties

To further justify the proposed prior, we show large support and posterior consistency properties. Large support is an important property in that it ensures that our prior can generate densities that are arbitrarily close to any true density f_0 in a large class, a defining property for a nonparametric Bayesian procedure and a necessary condition to allow the posterior to concentrate in small neighborhoods of the truth. Instead of focusing narrowly on GPT priors, we provide broad theoretical results for priors in the general class of expression (3.1).

Before proceeding, it is necessary to define some notation and concepts. We denote the Kullback-Leibler (KL) divergence of $f_{\mu,\sigma}$ from f_0 as $KL(f_{\mu,\sigma}, f_0)$ and an $\epsilon-$sized KL neighborhood around f_0 as $KL_\epsilon(f_0) = \{f_{\mu,\sigma} : KL(f_{\mu,\sigma}, f_0) < \epsilon\}$. The sup-norm distance is denoted by $||.||_\infty$. We note that, to generate $y_i \sim f_0$, one can draw $x_i \sim \text{Uniform}(0,1)$ and let $y_i = F_0^{-1}(x_i)$, F_0^{-1} being the inverse cdf on $\Re$ (assuming F_0^{-1} exists). This is equivalent to drawing samples from the limiting distribution in model (3.1) as $\sigma \to 0$ with $\mu_0 = F_0^{-1}$. For our development, we will assume that the true (data generating) density can be expressed as $f_0 = \frac{d}{dy}F_0 = \lim_{\sigma\to 0}\frac{d}{dy}F_{0,\sigma}$, where $F_{0,\sigma}$ is

the cdf of $f_{\mu_0,\sigma}$. More precisely,

(A1): The true (data generating) density can be represented as the limiting case

$$f_0(y) = \lim_{\sigma \to 0} \int_0^1 \Gamma_\sigma(y - \mu_0(x))dx \in (0, \infty), \tag{2.4}$$

where $\mu_0 = F_0^{-1}$, for all $y \in \Re$ and assuming Γ_σ is chosen so that such the limit exists. Since the limiting distribution of model (3.1) can be used to generate samples from any arbitrary distribution, *(A1)* includes the class of all strictly positive and finite densities for which convergence in distribution implies convergence of the corresponding density functions. Further in the proofs, we will also repeatedly use the fact that for μ close to μ_0 and $\sigma \in \Re^+$, $\log \frac{f_0}{f_{\mu,\sigma}} < \infty$ for f_0 defined in $(A1)$ and Γ_σ = Gaussian or Laplace. For notational convenience, we will often use μ and $\mu(x)$ interchangeably in the sequel.

Theorem 1 *Let Γ_σ be normal or Laplace having scale parameter σ with f_0 being the corresponding density in $\mathcal{F}$ defined as in* (A1). *If* $\Pi^* \otimes \nu\Big((\mu, \sigma) : ||\mu - \mu_0||_\infty < \eta_1, \sigma \in (0, \eta_2)\Big) > 0$ *for arbitrarily small* $(\eta_1, \eta_2) > 0$, *then* $\Pi(KL_\epsilon(f_0)) > 0$ *for all* $\epsilon(\eta_1, \eta_2) > 0$.

Theorem 1 allows us to verify that the induced prior on the density f assigns positive probability to KL neighborhoods of any strictly positive and finite true density f_0. From Schwartz (1965), if the true density f_0 is in the KL support of the prior for f, the posterior distribution for f will concentrate asymptotically in arbitrarily small weak neighborhoods of f_0. Theorem 1 requires the prior $\mu \sim \Pi^*$ to place positive probability in sup-norm neighborhoods of the inverse cdf F_0^{-1}. Although one can verify this condition for certain choices of Π^*, such as appropriately chosen Gaussian process priors, it is nonetheless somewhat stringent. We show in Theorem 2 that this condition can be relaxed to only require that the prior $\mu \sim \Pi^*$ assigns positive probability to L-1 neighborhoods of any element μ_0 of Θ. It is well known that positive sup-norm support automatically guarantees positive L-1 support but the converse is not true. Let us

denote an ϵ_1-sized L-1 neighborhood around μ_0 as $N_{\epsilon_1}(\mu_0) = \left\{ f_{\mu,\sigma} : \int |f_{\mu,\sigma} - f_0| < \epsilon_1 \right\}$.

Theorem 2 *Let $\Gamma_\sigma = \phi_\sigma$ and f_0 be the corresponding density in $\mathcal{F}$ defined in (*A1*). If $\Pi^* \otimes \nu \{(\mu, \sigma) : \mu \in N_{\eta_1}(\mu_0), \sigma \in (0, \eta_2)\} > 0$ for arbitrarily small $\eta_1, \eta_2 > 0$, then $\Pi(KL_\epsilon(f_0)) > 0$ for all $\epsilon(\eta_1, \eta_2) > 0$.*

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

Theorem 3 *Suppose $\tilde{f} = \lim_{\sigma \to 0} \int_0^1 \phi_\sigma(y - \tilde{\mu}(x))dx$, where $\tilde{\mu} = \tilde{F}^{-1}$, the inverse cdf corresponding to $\tilde{f}$. Then for $\mu \in N_{\epsilon_1}(\tilde{\mu})$ and $\sigma \in (\epsilon_2,\ \epsilon_2^*)$, we have $f_{\mu,\sigma} \in N_{\frac{\epsilon_1}{\epsilon_2}}\left(\tilde{f}\right)$ for arbitrarily small $\epsilon_1, \epsilon_2, \epsilon_2^*$ such that $0 < \epsilon_1 < \epsilon_2 < \epsilon_2^*$.*

Although Theorems 1-2 lead to weak posterior consistency, small weak neighborhoods around f_0 are topologically too large and may include densities that are quite different from f_0 in shape and other characteristics. Hence, it is appealing to establish a strong posterior consistency result in which the posterior probability allocated to arbitrarily small L-1 neighborhoods of f_0 increases towards one exponentially fast with increasing sample size. Focusing on the GPT prior described above, we show in Theorem 4 that strong posterior consistency holds under some conditions on the prior. Notably, for a GPT prior satisfying (*A2*) and a tail condition on prior ν, we obtain L-1 posterior consistency for all strictly positive and finite true densities $f_0 \in \mathcal{F}$ with the weak regularity condition (*A1*).

(A2): Suppose $\mu \sim GP(m, c)$ such that the mean function m$(\cdot)$ is continuously differentiable with $\sup_x m(x) < \infty$, and the covariance function c$(\cdot, \cdot)$ has continuous fourth derivatives.

Theorem 4 *Suppose (*A2*) holds and define f_0 as in (*A1*) where $\Gamma_\sigma = \phi_\sigma$. Further let $\nu(\sigma \in (0, L_n)) < d_1 \exp(-d_2 n)$ with $d_1, d_2 > 0$ for large n and $\lim_{n\to\infty} L_n = 0$. Then, f_0 is in the KL support of Π implies that the posterior is strongly consistent at f_0.*

The above assumptions can be verified for many popular GP covariance functions, both stationary and nonstationary. Some such examples can be found in Choi et. al. (2004). The proof of the above Theorem relies on Theorem 2 of Ghosal, Ghosh and Ramamoorthi (1999), which is listed as **Theorem 5** in Appendix A.

2.4 Single Factor Density Regression

As a simple and parsimonious single factor model that generalizes the model of Section 2.1 to include predictors $z_i = (z_{i1}, \ldots, z_{ip})'$ of a response y_i, we let

$$\begin{aligned}
y_i &= \mu^Y(x_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_Y^2), \\
z_{ik} &= \mu^{Z_k}(x_i) + \epsilon_{ik}^*, \quad \epsilon_{ik}^* \sim N(0, \sigma_{Z_k}^2), k = 1, 2, \ldots, p, \\
x_i &\sim \text{Uniform}(0, 1), \\
\mu^Y &\sim \Pi^Y, \quad \mu^{Z_k} \sim \Pi^{Z_k}, k = 1, 2, \ldots, p, \\
\sigma^Y &\sim \nu, \quad \sigma_{Z_k} \sim \nu,
\end{aligned} \tag{2.5}$$

where $\mu^Y, \mu^{Z_k} \in \Theta$ are unknown $[0, 1] \to \Re$ functions, ϵ's are independent errors and x_i is the latent variable. For simplicity, we assume the same prior ν on the precision of the measurement errors in each component model, though this assumption is trivial to relax. Expression (2.5) is a multivariate generalization of the univariate density

estimation model (3.1). Marginally each of the variables is assigned exactly the prior in (3.1) and to allow dependence we incorporate the same latent factor x_i in each of the models.

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

2.5 Posterior Computation

For simplicity, we focus on the single predictor density regression case when outlining an MCMC algorithm for posterior computation. Let $\mathrm{Y}_{n\times 1}$ and $\mathrm{Z}_{N\times 1}$ denote the vector of observations and covariates, respectively. We are interested in prediction of $y_{n+1},\ldots,y_N$ based on $z_{n+1},\ldots,z_N$. Let μ_Y^n $(n\times 1)$ and μ_Z^N $(N\times 1)$ denote the (unobserved) realizations of the GP μ^Y and μ^Z at the latent variable values $\mathbf{x}=(x_1,\ldots,x_n,x_{n+1},\ldots,x_N)'$. From the GP prior, we have $\mu_Y^n \sim \mathrm{N}_n(m_Y^n,\mathbf{K}_Y^n)$ and $\mu_Z^N \sim \mathrm{N}_N(m_Z^N,\mathbf{K}_Z^N)$. Let $\mathrm{N}(A|B)$ denote the conditional normal distribution. The covariance kernels are squared exponential with $\mathbf{K}_Y(x,x')=\frac{1}{\phi_Y}\exp\left\{-C_Y(x-x')^2\right\}$ and $\mathbf{K}_Z(x,x')=\frac{1}{\phi_Z}\exp\left\{-C_Z(x-x')^2\right\}$. We specify conjugate gamma priors: $\sigma_Y^{-2}\sim Ga(a_\sigma,b_\sigma)$, $\sigma_Z^{-2}\sim Ga(aa_\sigma,bb_\sigma)$, $\phi_Y\sim Ga(a_\phi,b_\phi)$ and $\phi_Z\sim Ga(aa_\phi,bb_\phi)$. For updating the latent variables $\mathbf{x}$, we adopt the griddy Gibbs approach using a set of

evenly distributed grid points $g_1^*, g_2^*, \ldots, g_G^* \in (0,1)$. Let $\mu_Y^n(-i)$ include all elements of μ_Y^n except $\mu^Y(x_i)$, and similarly for $\mu_Z^N(-i)$. The Gibbs sampling algorithm alternates between the following steps.

Step1: Update σ_Y^2 and σ_Z^2 using $\pi(\sigma_Y^{-2}|-) \sim$Ga(a$_\sigma$+n/2, b$_\sigma + \frac{1}{2}\sum_{i=1}^n (y_i - \mu^Y(x_i))^2$) and $\pi(\sigma_Z^{-2}|-) \sim$Ga(aa$_\sigma$+N/2, bb$_\sigma + \frac{1}{2}\sum_{i=1}^N (z_i - \mu^Z(x_i))^2$) respectively.

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

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

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

Step3: Update μ_Y^n and μ_Z^N using $\pi(\mu_Y^n|-) = N_n\Big((\mathbf{D}_Y^{-1}+(\mathbf{K}_Y^n)^{-1})^{-1}(\mathbf{D}_Y^{-1}Y+(\mathbf{K}_Y^n)^{-1}m_Y^n), (\mathbf{D}_Y^{-1}+(\mathbf{K}_Y^n)^{-1})^{-1}\Big)$ and $\pi(\mu_Z^N|-) = N_N\Big((\mathbf{D}_Z^{-1}+(\mathbf{K}_Z^N)^{-1})^{-1}(\mathbf{D}_Z^{-1}Z+(\mathbf{K}_Z^N)^{-1}m_Z^N), (\mathbf{D}_Z^{-1}+(\mathbf{K}_Z^N)^{-1})^{-1}\Big)$, where $\mathbf{D}_Y$=$\sigma_Y^2 I_n$ and $\mathbf{D}_Z$=$\sigma_Z^2 I_N$.

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

Step5: Update ϕ_Y and ϕ_Z using $\pi(\phi_Y|-) \sim$Ga(a$_\phi$+$\frac{n}{2}$, b$_\phi$+$\frac{1}{2}(\mu_Y^n - m_Y^n)'(\mathbf{K}_Y^n)^{-1}(\mu_Y^n - m_Y^n)$) and $\pi(\phi_Z|-) \sim$Ga(aa$_\phi + \frac{N}{2}$, bb$_\phi + \frac{1}{2}(\mu_Z^N - m_Z^N)'(\mathbf{K}_Z^N)^{-1}(\mu_Z^N - m_Z^N)$) respectively.

Step6: Update C$_Y$ and C$_Z$ using Metropolis random walk for log(C$_Y$) and log(C$_Z$).

For prediction of y_k based on z_k, k = $n+1, \ldots, N$, we use $\pi(y_k|-) = N(y_k; \mu^Y(x_k), \sigma_Y^2)$, while the conditional density estimate is calculated as $\hat{f}(y|z) = \frac{\frac{1}{G}\sum_{k=1}^G \phi_{\sigma_Y}(y-\mu^Y(g_k^*))\phi_{\sigma_Z}(z-\mu^Z(g_k^*))}{\frac{1}{G}\sum_{k=1}^G \phi_{\sigma_Z}(z-\mu^Z(g_k^*))}$.

2.6 Simulation Study

To assess the performance of the GPT approach in density estimation as well as density regression, we conducted several simulation studies. We chose the mean

function for the GP as m(x)=2sin(x)+cos(x) and utilized the squared exponential covariance kernel. For computational purposes, we worked with the standardized data and then transformed it back in the final step. The hyperparameters for the gamma priors were chosen to be one throughout. Although we used 75 grid points for the griddy Gibbs approach, the number of points could be as low as 60. The number of iterations used was 10000 with a burn in of 1000. The convergence for the main quantities such as μ was rapid with good mixing. All results are reported over 5 replicates.

2.6.1 Univariate Density Estimation

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

From table A.1, we see that even when the truth is generated from a finite mixture

of Gaussians, the GPT tends to do better or at least as well as the DP mixture of Gaussians. Mixtures of Polya trees have somewhat worse performance and result in overly spiky looking estimates.

2.6.2 Single Factor Density Regression

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

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

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

percentiles of the predictor distribution.

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

2.7 Epidemiological Application

2.7.1 Study Background

DDT is a cheap and popular alternative for reducing the transmission of malaria, but has been shown to have negative effects on public health. In order to study the association between the DDT metabolite DDE and preterm delivery, Longnecker et al. (2001) measured DDE in mother's serum in the third trimester of pregnancy and also recorded gestational age at delivery (GAD) as well as age. They did logistic regression with response as dichotomized GAD (preterm or normal depending on a cut-off of 37 weeks of completed gestation) and explanatory variables as categorized DDE based on empirical quantiles. Their results showed a significant dose-response relationship which had important public health implications. Dunson et al. (2008) analyzed the data using kernel stick-breaking processes, and showed an increasing bump in the left tail of

the GAD density with increasing DDE.

2.7.2 Analysis and Results

We used the GPT to analyze the dose response relationship in a subset of 182 women of advanced maternal age ($\geq$ 35 yrs) in the above dataset. We examined the conditional distribution of GAD at 10th, 60th, 90th and 99th percentile of DDE. Further, we looked at the dose response relationship between preterm birth and DDE, by examining the left tail of GAD over varying doses of DDE. We used normalized data for analysis and converted it back in the final step. Using the prior specification approach of section 2.2, we were able to incorporate prior information on the marginal density of GAD (using an external data) through the mean function of the GP. Note that prior on σ^{-2} for GAD was chosen as Ga(25,1). Given the limited sample size and the complexity of the data we are trying to model, we adjusted other hyperparameter settings to reflect our prior belief about the data. The starting value for the length-scale parameter in the covariance kernel in the Metropolis random walk was chosen to be 25, so as to have smooth Gaussian process prior. Instead of working with DDE, we used log(DDE) which resembled a Gaussian distribution, with a 0 mean function for the predictor component and Ga(1,1) prior for the corresponding residual precision.

Figure A.3 shows the conditional distribution curves for GPT along with 90% credible intervals. Although we focused on a small subsample of 182 women of advanced maternal age, the GPT results for the conditional density are remarkably similar to the ones reported in Dunson et al. (2008), which suggests that there is no systematic difference for women of advanced maternal age. The conditional densities show an increasing bump in the left tail with increasing DDE, suggesting increased risk of preterm birth at higher doses. This is further supported by dose-response curves

for P(GAD<T) in Figure A.5, with different choices for cut-off T. Although the dose-response curve is mostly flat for T=33 weeks, the relationship becomes more significant as cut-off increases, with the dose-response tapering off at T=40 weeks. This suggests that increased risk of preterm birth at higher DDE dosage is attributable to premature deliveries between 33 and 37 weeks. Trace plots of f($y|z$) for different DDE percentiles (not shown) exhibit excellent rates of convergence and mixing. For comparison, Figure A.4 shows the density estimates from the DP mixture of Gaussians which has a tendency to overly favor multimodal densities, which is as expected given our simulation study results. These results were obtained using DP package in R (and the data driven hyperparameter values therein), which utilizes algorithm 8 of Neal (2000) with m=1.

2.8 Discussion

In this paper, we propose a latent factor model for density estimation. This novel method provides us with a flexible non-discrete mixture alternative to be used in a variety of situations including density estimation, density regression, hierarchical latent variable models and even mixed models. We provide theoretical theoretical justifications for GPT and demonstrate it's usefulness as a building block for more complex models involving covariates. Building on our work, Pati, Bhattacharya and Dunson recently showed minimax optimal rates of posterior contraction for Bayesian density estimation from non-linear latent variable models, also obtaining initial results on contraction rates in conditional density estimation. The close relationship between non-linear latent variable models for densities and non-linear mean regression models facilitates not only posterior computation but also derivations of theoretical properties, such as contraction rates, which have proven difficult to study for discrete mixtures beyond simple settings.

Chapter 3

Bayes Variable Selection in Semiparametric Linear Models

Having proposed an elegant solution based on non-mixture alternatives to the important problem of Bayesian density estimation and density regression, we now turn our attention to another fundamental problem in statistics, namely variable selection. Variable selection involves selecting an important and potentially parsimonious subset of predictors which significantly affects the outcome. As stated in the introduction, majority of the literature has focused on linear regression models involving Gaussian residuals. Our objective is to propose a consistent Bayes variable selection method in linear regression models having unknown residuals, which is expected to perform well in a wide variety of situations due to greater flexibility of the proposed model.

3.1 Model Formulation

In this section, we propose a new class of priors for Bayesian variable selection in linear regression models with an unknown residual density characterized via a Dirichlet

process (DP) location mixture of Gaussians. In particular, let

$$\begin{aligned} y_i &= \mathbf{x}'_{\gamma,i}\beta_\gamma + \epsilon_i, \quad \epsilon_i \sim f, \quad i = 1, \ldots, n, \\ f(\cdot) &= \int N(\cdot; \alpha, \tau^{-1}) dP(\alpha), \quad P \sim DP(mP_0), \quad P_0 = N(0, \tau^{-1}), \end{aligned} \tag{3.1}$$

where $\mathbf{x}_{\gamma,i}$ is the ith row of $\mathbf{X}_\gamma$ and does not include an intercept as we do not restrict f to have zero mean, and f is a density with respect to Lebesgue measure on $\Re$. We address uncertainty in subset selection by placing a prior on γ, while the prior on β_γ characterizes prior knowledge of the size of the coefficients for the selected predictors.

The DP mixture prior on the density f induces clustering of the n subjects into k groups/subclusters, where k is random and each group has a distinct intercept in the linear regression model. Let A denote an $n \times k$ allocation matrix, with $A_{ij} = 1$ if the ith subject is allocated to the jth cluster and 0 otherwise. The jth column of A then sums to n_j, the number of subjects allocated to subcluster j, with $\sum_{j=1}^{k} n_j = n$. Following Kyung, Gill and Casella (2009), conditionally on the allocation matrix A, (3.1) can be represented as a linear model with random intercepts

$$Y^n = A\eta + X_\gamma \beta_\gamma + \epsilon, \quad \eta \sim N(0, \tau^{-1} I_k), \quad \epsilon \sim N(0, \tau^{-1} I_n), \tag{3.2}$$

where A is random with a certain prior probability given by the coefficients in the summation of the likelihood expression (3.7).

In keeping with the mixtures of g-priors literature, we would like the prior on the regression coefficients to retain the essential elements of Zellner's g-prior (Zellner, 1986), but at the same time to be suitably adapted to reflect the semi-parametric nature of the model - more specifically, the clustering of responses by the DP kernel mixture prior. To this effect, we propose a mixture of semi-parametric g-priors which is constructed to scale the covariance matrix in Zellner's g-prior to reflect the clustering phenomenon

as follows:

$$\pi(\beta_\gamma) = N(0, g\tau^{-1}(X'_\gamma \Sigma_A^{-1} X_\gamma)^{-1}), \quad \Sigma_A = I + AA', \quad g \sim \pi(g). \tag{3.3}$$

Prior (3.3) inherits the advantages of the traditional mixtures of g-priors including computational efficiency in computing marginal likelihoods (conditional on A) and robustness to mis-specification of g. In addition, the prior can be interpreted as having arisen from the analysis of a conceptual sample generated using a scaled design matrix $\Sigma_A^{-1/2} X_\gamma$, reflecting the clustering phenomenon due to the DP kernel mixture prior. Moreover, the proposed prior leads to Bayes factor and variable selection consistency in semi-parametric linear models (3.1), as highlighted in the sequel.

Note that since $(X'_\gamma \Sigma_A^{-1} X_\gamma)^{-1} \geq (X'_\gamma X_\gamma)^{-1}$, the prior variance of Y conditional on (g, τ) is higher for the semi-parametric $g-$prior as compared to the traditional $g-$prior for any allocation matrix A. To assess the influence of A on the prior for β_γ, we did simulations which revealed that for fixed $(n,\ p)$, $\text{var}(\beta_{\gamma_l})$ increases but the $\text{cov}(\beta_{\gamma_l}, \beta_{\gamma_{l'}})$ decreases as the number of underlying subclusters in the data increase $(l', l = 1, \ldots, p, l' \neq l)$. This suggests that as the number of groups in A increase, the components of β_γ are likely to be more dispersed with decreasing association between each other.

3.2 Bayes Factor in Semiparametric Linear Models

Throughout the rest of the paper, we will assume that the data $\text{Y}^n=(Y_1, \ldots, Y_n)'$ are generated from the true model $\mathcal{M}_T : Y^n = X_{\gamma_1}\beta_{\gamma_1} + \epsilon$, with ϵ_i i.i.d. from the true residual density f_0, which is a density on $\Re$ with respect to Lesbesgue measure. For modeling purposes, we put a DP location mixture of Gaussians prior on the unknown f_0. For pairwise comparison, we evaluate the evidence in favor of $\mathcal{M}_1$ compared to $\mathcal{M}_2$

using the Bayes factor, where

$$\begin{aligned}
\mathcal{M}_1 &: Y^n = X_{\gamma_1}\beta_{\gamma_1} + \epsilon_1, \quad \epsilon_{1i} \sim f \\
\mathcal{M}_2 &: Y^n = X_{\gamma_2}\beta_{\gamma_2} + \epsilon_2, \quad \epsilon_{2i} \sim f \\
f(\cdot) &= \int N(\cdot; \alpha, \tau^{-1}) dP(\alpha), \quad P \sim DP(mP_0), \quad P_0 = N(0, \tau^{-1}) \\
\beta_{\gamma_j} &\sim \pi(\beta_{\gamma_j}), j = 1, 2, \quad \pi(\tau^{-1}) \propto 1/\tau^{-1}, \quad g \sim \pi(g),
\end{aligned} \tag{3.4}$$

where $\gamma_j \in \Gamma$ indexes models of dimension p_j and $\pi(\beta_{\gamma_j})$ is defined in (3.3), $j = 1, 2$. Our prior specification philosophy is similar to the one adopted by Guo and Speckman (2009) for normal linear models, in that we assign proper priors on all elements of both $\beta_{\gamma_1}, \beta_{\gamma_2}$ conditional on (g, τ^{-1}), and an improper prior on τ^{-1} for a more objective assessment. However unlike Guo and Speckman (2009), our focus is on Bayesian variable selection in semi-parametric linear models.

Note that the conditional likelihood of the response after marginalizing out η in (3.2) is $L(Y^n|A, \beta_\gamma, \tau^{-1}) = N(X_\gamma\beta_\gamma, \tau^{-1}\Sigma_A)$ (Kyung et. al., 2009). Thus conditional on A and under the DPM of Gaussians prior on f, $\mathcal{M}_j$ in (3.4) reduces to the normal linear model:

$$\Sigma_A^{-1/2} Y^n = Z_A = \tilde{X}_{A,\gamma_j}\beta_{\gamma_j} + \epsilon, \quad \epsilon \sim N(0, \tau^{-1} I_n), \quad \pi(\beta_{\gamma_j}) = N(0, g\tau^{-1}(\tilde{X}'_{A,\gamma_j}\tilde{X}_{A,\gamma_j})^{-1}), \tag{3.5}$$

where $\tilde{X}_{A,\gamma_j} = \Sigma_A^{-1/2} X_{\gamma_j}$. Under a mixture of semi-parametric g-priors, we can directly use expression (17) in Guo and Speckman (2009) to obtain (conditional on A) for $j = 1, 2$

$$L(Z_A|\mathcal{M}_j) \equiv L(Y^n|A, \mathcal{M}_j) \propto (Z'_A Z_A)^{-n/2} \int_0^\infty (1+g)^{-p_j/2} \left[1 - \frac{g}{1+g}\frac{Z'_A \tilde{H}_{A,j} Z_A}{Z'_A Z_A}\right]^{-n/2} \pi(dg), \tag{3.6}$$

where $\tilde{H}_{A,j} = \tilde{X}_{A,\gamma_j}(\tilde{X}'_{A,\gamma_j}\tilde{X}_{A,\gamma_j})^{-1}\tilde{X}'_{A,\gamma_j}$ is the equivalent of a hat matrix in normal linear regression.

Also, marginalizing over all possible subcluster allocations for a given sample size

n, the following marginal likelihood can be obtained under a DP prior on f (Kyung et. al., 2009):

$$L(Y^n|\mathcal{M}_j) = \frac{\Gamma(m)}{\Gamma(m+n)} \sum_{k=1}^{n} m^k \sum_{A \in \mathcal{A}_k} \prod_{i=1}^{k} \Gamma(n_i) L(Y^n|A, \mathcal{M}_j) = \sum_{A_l \in \mathcal{C}_n} w_l L(Y^n|A_l, \mathcal{M}_j) \quad (3.7)$$

where $\mathcal{A}_k$ is the collection of all possible $n \times k$ matrices corresponding to different allocations of n subjects into k subclusters, $\mathcal{C}_n$ is the collection of all possible allocation matrices for a sample size n with $\sum_{A_l \in \mathcal{C}_n} w_l = 1$. In the limiting case as $n \to \infty$, we have $\mathcal{C}_\infty$ as the class of limiting allocation matrices. Further using (3.6), the Bayes factor in favor of $\mathcal{M}_2$ conditional on the allocation matrix A is given by

$$BF^n_{21,A} = \frac{L(Z_A|\mathcal{M}_2)}{L(Z_A|\mathcal{M}_1)} = \frac{\int_0^\infty (1+g)^{-p_2/2} \left[1 - \frac{g}{1+g}\tilde{R}^2_{A,2}\right]^{-n/2} \pi(dg)}{\int_0^\infty (1+g)^{-p_1/2} \left[1 - \frac{g}{1+g}\tilde{R}^2_{A,1}\right]^{-n/2} \pi(dg)}, \quad (3.8)$$

where $\tilde{R}^2_{A,j} = Z'_A \tilde{H}_{A,j} Z_A / Z'_A Z_A$, $(j = 1, 2)$. Finally using (3.7), the unconditional Bayes factor in favor of $\mathcal{M}_2$ marginalizing out A is

$$BF^n_{21} = \frac{L(Y^n|\mathcal{M}_2)}{L(Y^n|\mathcal{M}_1)} = \frac{\sum_{A_l \in \mathcal{C}_n} w_l L(Z_{A_l}|\mathcal{M}_2)}{\sum_{A_l \in \mathcal{C}_n} w_l L(Z_{A_l}|\mathcal{M}_1)}. \quad (3.9)$$

3.3 Posterior Computation

We propose an MCMC algorithm for posterior computation for (3.1), which combines a stochastic search variable selection algorithm (George and McCulloch, 1997) for variable selection with recently proposed methods for efficient computation in DP mixture models. In particular, we utilize the slice sampler of Walker (2007) incorporating the modification of Yau et al.(2011). Using Sethuraman's (1994) stick-breaking

representation, let

$$P = \sum_{j=1}^{\infty} \pi_j \alpha_j, \quad \alpha_j \sim N(0, \tau^{-1}), \quad \pi_j = \nu_j \prod_{l<j} (1 - \nu_l), \quad \nu_l \sim Beta(1, m). \quad (3.10)$$

The slice sampler of Walker (2007) relies on augmentation with uniform latent variables, which allows us to move from an infinite summation for P in (3.10) to a finite sum given the uniform latent variable. In particular,

$$f_{\pi,\alpha}(y|u) = \sum_{j \in B_\pi(u)} N(y|\alpha_j), \quad B_\pi(u) = \{j : \pi_j > u\} \text{ is a finite set}, \quad u \sim U(0, 1).$$

For the DP precision parameter, we specify the hyperprior $m \sim Ga(a_m, b_m)$ for greater flexibility. We specify a $Ga(a_\tau, b_\tau)$ prior on τ and $Be(a_1, b_1)$ prior on the marginal inclusion probabilities $\Pr(\gamma_l = 1)$, $l = 1, \ldots, p$. We outline the posterior computation steps in Appendix B.

3.4 Asymptotic Properties

In this section we establish asymptotic properties for the proposed approach using γ_1 to index the true model $\mathcal{M}_1$ and γ_2 to index an arbitrary model $\mathcal{M}_2$ being compared to γ_1, with $\mathcal{M}_1 \subset \mathcal{M}_2$ denoting nesting of $\mathcal{M}_1$ in $\mathcal{M}_2$. Before proceeding, we introduce some regularity conditions essential for the development of asymptotic theory.

$(A1')$ $\lim_{n\to\infty} \frac{\beta'_{\gamma_1}(X'_{\gamma_1}X_{\gamma_1})\beta_{\gamma_1}}{n} \to b_1 > 0$.

$(A2')$ For $\mathcal{M}_1 \not\subseteq \mathcal{M}_2$, $\lim_{n\to\infty} \frac{\beta'_{\gamma_1}X'_{\gamma_1}H_2X_{\gamma_1}\beta_{\gamma_1}}{n} \to b_2 \in [0, b_1)$, with $H_2 = X_{\gamma_2}(X'_{\gamma_2}X_{\gamma_2})^{-1}X'_{\gamma_2}$.

(A1) For $p_1 = O(n^{a_1}), 0 \le a_1 < 1$, $\lim_{n\to\infty} \frac{\beta'_{\gamma_1}(X'_{\gamma_1}\Sigma_A^{-1}X_{\gamma_1})\beta_{\gamma_1}}{n} \to b_{A,1} > 0$.

(A2) For $\mathcal{M}_1 \not\subseteq \mathcal{M}_2$, $\lim_{n\to\infty} \frac{\beta'_{\gamma_1}\tilde{X}'_{A,\gamma_1}\tilde{H}_{A,2}\tilde{X}_{A,\gamma_1}\beta_{\gamma_1}}{n} \to b_{A,2}$, where $b_{A,2} \in [0, b_{A,1})$ for fixed p_1, p_2, and $b_{A,2} \in (0, b_{A,1})$ for $p_j = O(n^{a_j})$ $(j = 1, 2, 0 \le a_1 < a_2 < 1)$.

$(A1) - (A2)$ depend on the allocation matrix A, which is an $n \times k$ binary matrix that

for large n tends to have $k << n$, and be very sparse containing mostly zeros with sparsity increasing with column index. We also assume the following for the class of proper priors $\pi(g)$ on g:

(A3): There exists a constant $k \geq 0$ such that $\int_{a_n}^{c_0 a_n} \pi(dg) \approx n^{-k}$ for any constant $c_0 > 1$ and any sequence $a_n \approx n$. Here $a_n \approx b_n$ implies that $\lim_{n\to\infty} a_n/b_n > 0$.

(A4): There exists a constant k_u such that $k - (p_2 - p_1)/2 < k_u \leq k$ and $\int_0^\infty (1+g)^{k_u}\pi(dg) \approx 1$.

We note that *(A1′)*, *(A2′)* are the standard assumptions for establishing Bayes factor consistency in normal linear models, on which our assumptions *(A1)*, *(A2)* are based. We develop the asymptotic theory for semiparametric linear models (3.4) based on assumptions *(A1)*–*(A4)*. We note that *(A1)* is a stronger assumption compared to *(A1′)*, since *(A1)* implies *(A1′)* as $\Sigma_A^{-1} = I_n - A(I_k + A'A)^{-1}A'$. Further, in the extreme case when $A = I_n$, we have $\frac{X'_{\gamma_1}\Sigma_A^{-1}X_{\gamma_1}}{n} = \frac{1}{2}\frac{X'_{\gamma_1}X_{\gamma_1}}{n}$, so that *(A1′)* implies *(A1)*. Again when $A = 1_n$, $X'_{\gamma_1}\Sigma_A^{-1}X_{\gamma_1} \approx X'_{\gamma_1}X_{\gamma_1} - n\bar{X}'_{\gamma_1}\bar{X}_{\gamma_1}$ for large n, for $\bar{X}_{\gamma_1} = 1'_n X_{\gamma_1}/n$. Hence $\frac{\beta'_{\gamma_1}(X'_{\gamma_1}\Sigma_A^{-1}X'_{\gamma_1})\beta_{\gamma_1}}{n} \approx \frac{\beta'_{\gamma_1}(X^{c'}_{\gamma_1}X^c_{\gamma_1})\beta_{\gamma_1}}{n}$, where $\mathrm{X}^c_{\gamma_1}$ is the centered design matrix. When $\lim_{n\to\infty}\frac{\beta'_{\gamma_1}(X^{c'}_{\gamma_1}X^c_{\gamma_1})\beta_{\gamma_1}}{n} > 0$, *(A1′)* $\Rightarrow$ *(A1)*.

Assumption *(A2)* can be interpreted as a positive 'limiting distance' between the two models corresponding to design matrices X_{γ_1} and X_{γ_2} in (3.2) conditional on A, after marginalizing out η, i.e. $\Delta_{21,A} = \lim_{n\to\infty}\frac{\beta'_{\gamma_1}\tilde{X}'_{A,\gamma_1}(I_n - \tilde{H}_{A,2})\tilde{X}_{A,\gamma_1}\beta_{\gamma_1}}{n\tau^{-1}} = \frac{b_{A,1}-b_{A,2}}{\tau^{-1}} \in (0,\infty)$. Such a 'limiting distance' ($\Delta_{21,A}$) can be considered as a natural extension of the definition of distance between two normal linear models in Casella et. al. (2009) and Moreno et. al. (2010) to models with random intercept as in (3.2).

Assumptions *(A3)*, *(A4)* define a class of proper priors for g described in Guo and Speckman (2009). This class includes hyper-g ($\frac{a-2}{2}(1+g)^{-a/2}$) and hyper-$g/n$ ($\frac{a-2}{2n}(1+g/n)^{-a/2}$) priors with $2 < a \leq 4$ (Liang et. al. 2008), Zellner-Siow priors (Zellner and Siow, 1980) as well as beta-prime priors (Maruyama and George, 2008).

It is clear that these assumptions on $\pi(g)$ are satisfied by quite a few standard priors are hence are quite reasonable.

The following lemma gives the limits of quantities such as $\tilde{R}^2_{A,j} = Z'_A \tilde{H}_{A,j} Z_A / Z'_A Z_A$ $(j = 1, 2)$, which would be useful for establishing asymptotic properties. The proof follows directly using Lemmas 1, 2 of Guo and Speckman (2009) and from (3.5) which essentially states that under the DP mixture of Gaussians prior on f for $\mathcal{M}_j$ in (3.4) and conditional on allocation matrix A, $Z_A = \Sigma_A^{-1/2} Y^n \sim N(\tilde{X}_{A,\gamma_j} \beta_{\gamma_j}, \tau^{-1} I_n)$, $j = 1, 2$.

Lemma 1 *Let assumptions (A1), (A2) hold.*

(i) If $\mathcal{M}_1 \subset \mathcal{M}_2$, conditional on A, $\tilde{R}^2_{A,1} \overset{a.s.}{\to} \frac{b_{A,1}}{\tau^{-1}+b_{A,1}}$, $\tilde{R}^2_{A,2} \overset{a.s.}{\to} \frac{b_{A,1}}{\tau^{-1}+b_{A,1}}$, under $\mathcal{M}_1$.

(ii) If $\mathcal{M}_1 \not\subseteq \mathcal{M}_2$, conditional on A, $\tilde{R}^2_{A,1} \overset{a.s.}{\to} \frac{b_{A,1}}{\tau^{-1}+b_{A,1}}$, $\tilde{R}^2_{A,2} \overset{a.s.}{\to} \frac{b_{A,2}}{\tau^{-1}+b_{A,1}}$, under $\mathcal{M}_1$.

As shown by the following result, the proposed approach leads to Bayes factor consistency when comparing fixed dimensional models as well as models growing at the rate $O(n^t), 0 < t < 1$, when the truth is sparse.

Theorem 6 *Let assumptions (A1), (A2) hold .*

(I) Suppose p_1 and p_2 are fixed. If $\mathcal{M}_1 \subset \mathcal{M}_2$, then under $\mathcal{M}_1$ and assumptions (A3), (A4), $BF^n_{21} \overset{P}{\to} 0$ as $n \to \infty$ and if $p_2 - p_1 > 2 + 2(k - k_u)$, $BF^n_{21} \overset{a.s.}{\to} 0$ as $n \to \infty$. Further, if $\mathcal{M}_1 \not\subseteq \mathcal{M}_2$, then under $\mathcal{M}_1$ and assumption (A3), $BF^n_{21} \overset{a.s.}{\to} 0$ as $n \to \infty$.

(II) Suppose p_j is growing at the rate $O(n^{a_j})$, j=1,2, with $0 \le a_1 < a_2 < 1$. Then under $\mathcal{M}_1$ and assumption (A3), $BF^n_{21} \overset{a.s.}{\to} 0$ as $n \to \infty$.

REMARK 1. Although we do not present the proof here, Theorem 6 can be modified to accommodate the case of improper priors on g (i.e. $\pi(g) \propto \frac{1}{1+g}$). In such a case, assumptions *(A3), (A4)* are excluded and we require $p_2 - p_1 \ge 3$ for a.s. convergence in (I) for $\mathcal{M}_1 \subset \mathcal{M}_2$.

The next result establishes model selection consistency for the proposed approach, even in cases when the cardinality of the model space increases with n. In particular, we consider cases when the number of candidate predictors p_n (abusing the notation

slightly) is growing at the rate $O(n^a), a > 0$, but the prior on the model space assigns zero probability to models growing at a rate equal to or faster than n. When $a \geq 1$, the prior support consists of models constructed using $O(n^t)$ $(0 \leq t < 1)$ sized subsets of p_n candidate predictors.

To elaborate, let the support of the prior on the model space be $\mathcal{M} = \mathcal{M}_F \bigcup \mathcal{M}_I$, where $\mathcal{M}_F$ is the set of all models γ such that there exists a sample size $n_0 < \infty$ for which $\gamma_j = 0$ for all $j > p_{n_0}$, and $\mathcal{M}_I$ is the set of all models with dimensions growing at a rate strictly less than n, $\mathcal{M}_I = \{\gamma : \sum_{j=1}^{p_n} \gamma_j = O(n^t), 0 < t < \min(a, 1)\}$. Letting $p_0 = \max\{j : \gamma \in \mathcal{M}_F, \gamma_j = 1\}$, we can discard predictors having a higher index than p_0 for all $\gamma \in \mathcal{M}_F$ and treat $\mathcal{M}_F$ as finite dimensional having $2^{p_0} - 1$ elements (excluding the null model). Let γ_{jl} denote the lth model having dimension p_j, with $l = 1, \ldots, \binom{p_0}{p_j}$ when $\gamma_{jl} \in \mathcal{M}_F$ and $l = 1, \ldots, \binom{p_n}{p_j}$ when $\gamma_{jl} \in \mathcal{M}_I$. Consider the following sequence of priors which assigns greater penalty to models with increasing dimensions, thus encouraging sparsity: $\left\{\pi^n(\gamma_{jl}) \propto 2^{-p_j} I[\gamma_{jl} \in \mathcal{M}_F] + \binom{p_n}{p_j}^{-1} I[\gamma_{j_l} \in \mathcal{M}_I]\right\}$. When the truth is sparse such that $\mathcal{M}_1 \in \mathcal{M}_F$, we have the following result.

Theorem 7 *Suppose assumptions (A1)-(A4) hold. For fixed p and under $\mathcal{M}_1$, $P(\mathcal{M}_1|Y^n) \xrightarrow{P} 1$ for $\{\pi(\mathcal{M}_\gamma) : \pi(\mathcal{M}_1) > 0\}$. When $p_n = O(n^a)$ $(a > 0)$ and $\mathcal{M}_1 \in \mathcal{M}_F$, $P(\mathcal{M}_1|Y^n) \xrightarrow{P} 1$ under $\mathcal{M}_1$, for $\pi^n(\gamma_{j_l}) \propto 2^{-p_j} I[\gamma_{j_l} \in \mathcal{M}_F] + \binom{p_n}{p_j}^{-1} I[\gamma_{j_l} \in \mathcal{M}_I]$ such that $\pi^n(\mathcal{M}_F \cup \mathcal{M}_I) = 1$.*

3.5 Simulation Study

We present the results of two simulation studies comparing our method (SLM) with the normal linear model (NLM) having $\beta_\gamma \sim N(0, g\tau^{-1}(X'_\gamma \Sigma^{-1}_{A=1_n} X_\gamma)^{-1})$ (designed to assign comparable prior information when the residual is Gaussian), the lasso (Tibshirani, 1996) and elastic net (Zou and Hastie, 2005), as well as robust variable selection methods including an MM-type regression estimator (Yohai, 1987; Koller and Stahel,

2011), and a median regression model with SSVS for variable selection (technical report by Reed, Dunson and Yu, 2010). The data is generated as follows:

$$\begin{aligned} \textbf{Case I}: \quad y_i &= \mathbf{x}_i\beta_T + \epsilon_i, \quad \epsilon_i \sim 0.5N(2.5,1) + 0.5N(-2.5,1), \\ \textbf{Case II}: \quad y_i &= 1 + \mathbf{x}_i\beta_T + \epsilon_i, \quad \epsilon_i \sim N(0,1), \quad i = 1,\ldots,n, \end{aligned}$$

where $\mathbf{x}_i$ is a ten dimensional predictor (p=10), with $x_{ij}, j = 1,\ldots,10$ generated independently from U(-1,1), and $\beta_T = (3, 2, -1, 0, 1.5, 1, 0, -4, -1.5, 0)$.

We used $Ga(0.1, 1)$ prior on the DP precision parameter and $Be(0.1, 1)$ prior on $P(\gamma_j = 1)$, j=1,...,p, which corresponds to a weakly informative prior favoring parsimony. For the griddy Gibbs approach, we chose 1,000 equally spaced quantiles from $Be(1, 1)$ prior for $\frac{g}{1+g}$ (corresponding to $a = 4$ in the hyper-g prior). For both SLM and NLM, we made 50,000 runs with a burn in of 5,000. We implemented the lasso (L1) and elastic net (EL) using the GLMNET package in R with default settings, while the MM-type estimator (LMR) was implemented using 'lmrob' function in 'robustbase' package in R and the median regression with SSVS (QR) was implemented using function 'SSVSquantreg' in 'MCMCpack' package in R, with a $Be(0.1, 1)$ prior on the prior inclusion probability for predictors. All results are summarized across 20 replicates. The computation time for SLM per iteration was marginally slower than NLM. The mixing for the fixed effects was good under both the methods. The results for SLM do not appear to be sensitive to the hyper-parameters in $\pi(m)$, but are mildly sensitive to hyper-parameters in $\pi(g)$ for $n = 100$.

We study the marginal inclusion probabilities (MIP) under SLM and NLM over varying sample sizes in Figures B.1 and B.2. These plots suggest a faster rate of increase of the MIP for the important predictors under SLM as compared to NLM when the true residuals are non-Gaussian, and a very similar rate of increase under

both methods when the true residuals are Gaussian (thus justifying the prior choice for NLM). In contrast, the exclusion probabilities for the unimportant predictors converge to one slowly under both the methods, reflecting the well known tendency for slower accumulation of evidence in favor of the true null.

Tables B.1 and B.2 present some summaries for $n = 100$ for Case I. The MIPs in Table B.1 suggests correct variable selection decision by SLM, but poor performance by NLM which fails to exclude any of the unimportant predictors under median probability model. Further, L1, EL and QR seem to favor an overly complex model by choosing a superset of important predictors. In terms of estimation of the fixed effects, SLM has the highest degree of accuracy as reflected by the smallest mean square error ($\frac{||\hat{\beta}-\beta_T||_2}{p}$) in Table B.2. In addition, the replicate average mean square error for out of sample prediction for a test sample size of 25 (Table B.2) is smallest under the SLM, followed by lasso and elastic net. NLM is seen to be clearly inadequate for prediction purposes as indicated by the enormously high out of sample predictive MSE. Thus in conclusion, when the true residual is non-Gaussian, the SLM has the best performance compared to competitors, whereas NLM completely fails as an out of sample prediction tool.

3.6 Application to Diabetes Data

The prevalence of diabetes in the United States is expected to more than double to 48 million people by 2050 (Mokdad et. al., 2001). Previous medical studies have suggested that Diabetes Mellitus type II (DM II) or adult onset diabetes could be associated with high levels of total cholesterol (Brunham et. al., 2007) and obesity (often characterized by BMI and waist to hip ratio) (Schmidt et. al., 1992), as well as hypertension (indicated by a high systolic or diastolic blood pressure or both) which is twice as prevalent in diabetics compared to non-diabetic individuals (Epstein and Sowers, 1992).

We develop a comprehensive variable selection strategy for indicators of DM II in African-Americans based on data obtained from Department of Biostatistics, Vanderbilt University website. Our primary focus is to discover important indicators of DM II by modeling the continuous outcome glycosylated hemoglobin ($> 7mg/dL$ indicates a positive diagnosis of diabetes) based on predictors such as total cholesterol (TC), stabilized glucose (SG), high density lipoprotein (HDL), age, gender, body mass index (BMI) indicator (overweight and obese with normal as baseline), systolic and diastolic blood pressure (SBP and DBP), waist to hip ratio (WHR) and postprandial time indicator (PPT) (1/0 depending on whether the blood was drawn within 2 hours of a meal). In addition to total cholesterol, obesity and hypertension, we note that lower levels of HDL have been known to be associated with insulin resistance syndrome (often considered a precursor of DM II with a conversion rate around 30%). We also expect PPT to be a significant indicator as blood sugar levels are high up to 2 hours after a meal.

After excluding the records containing missing values, the data consisted of 365 subjects which was split into multiple training and test samples of sizes 330 and 35 respectively. The replicate averaged fixed effects estimates (multiplied by 100) for the SLM, NLM, L1, EL, LMR and QR are presented in Table B.3, and the marginal inclusion probabilities (MIP) for the SLM, NLM and QR are summarized in Table B.3. We also evaluate the out of sample predictive performance for each training-test split using predictive MSE in Table B.5, and additionally provide the mean coverage (COV) and width (CIW) of 95% pointwise credible intervals for the predicted responses under SLM and NLM. The same values of hyper-parameters were used as in previous section. For each replicate, we randomized the initial starting points and made 100,000 runs for SLM (burn in = 20,000) and 50,000 runs for NLM (burn in = 5,000).

It is interesting to note from Table B.4 that the variable selection decisions under

SLM (using median probability model) are quite different compared to the NLM. In particular, while both the models successfully identify total cholesterol, stabilized glucose and postprandial time as important predictors, it is only the SLM which identifies systolic hypertension (MIP = 0.72), HDL (MIP = 0.64) and waist to hip ratio (MIP = 0.93) as important indicators, compared to NLM which assigns MIP = 0.14, 0.39 and 0.13 to SBP, HDL and WHR respectively. Age is identified as an important predictor under NLM (MIP = 0.67), but not under SLM (MIP=0.43). For both the methods, the MIPs for BMI (overweight and obese) were low, which could potentially be attributed to adjusting for the other obesity factors such as waist to hip ratio. From Tables B.3 and B.4, we also see that the lasso, elastic net and the MM-type estimator select an overly complex model by excluding minimal number of predictors, while the quantile regression with SSVS fails to include several important predictors and selects a highly parsimonious and inadequate model.

Variable selection in this application is clearly influenced by the assumptions on the residual density, with the nonparametric residual density providing a more realistic characterization that should lead to a more accurate selection of the important predictors. Figure B.3 shows an estimate of the residual density obtained from the SLM analysis, suggesting a unimodal right skewed density with a heavy right tail. The SLM results suggest that a mixture of two Gaussians provides an adequate characterization of this density. The computation time for SLM is only marginally slower than NLM, and in addition SLM exhibits good mixing for most of the fixed effects (Table B.6). These results are robust to SSVS starting points, and consistency in the results across training-test splits also indirectly suggests adequate computational efficiency of SSVS.

In terms of out of sample predictive MSE (Table B.5), the relative performance between SLM, NLM, L1 and EL vary across training-test splits so that none of the

models can be said to dominate the others, while LMR and QR produce relatively inferior prediction results. Overall, the NLM has narrower 95% pointwise credible intervals compared to SLM, often resulting in poorer coverage for out of sample predictions. In conclusion, SLM succeeds in choosing the most reasonable model for DM II, consistent with previous medical evidence, and compares favorably with other competitors for prediction purposes.

3.7 Discussion

We develop mixtures of semi-parametric g-priors for linear models with non-parametric residuals characterized by DP mixtures of Gaussians. The proposed method addresses the often encountered issue of non-Gaussianity of residuals in variable selection settings, and has attractive asymptotic justifications such as Bayes factor and variable selection consistency involving fixed p as well as $p > n$ (under some restrictions on the model space). Further, the method is essentially no more difficult to implement than SSVS for normal linear models and can lead to substantially different conclusions, as illustrated in the diabetes application. The general topic of semi- and nonparametric Bayesian model selection is understudied and we hope that this work stimulates additional research of this type in broader model classes, such as for generalized linear models and nonparametric regression.

Chapter 4

Bayesian Credible Regions for Vectors and Functions

The previous chapter addressed an important problem of developing a consistent variable selection methodology for linear regression models with unknown residuals. Another fundamental problem in the Bayesian paradigm is constructing simultaneous credible regions for vectors and functions, guaranteed to have at least a pre-specified posterior probability content. However, such a problem has received little attention and most of the existing methods seem to rely on marginal distributions to construct point-wise credible intervals. We propose a methodology for this important problem which directly uses joint distributions of parameters to construct credible regions, is straightforward to implement, and not computationally intensive for small to moderate number of knots.

4.1 Credible regions for vectors

We first focus on estimating elliptical credible regions for vector-valued parameters, $\theta = (\theta_1, \ldots, \theta_p)'$. We use elliptical regions for tractability in inferences and computation. When p is not small it is particularly important for the estimated region

to be easy to store and visualize, while also allowing rapid inferences about whether particular parameter values of interest are contained in the region. One simple way to estimate an elliptical credible region based on MCMC samples from the posterior of θ is to rely on asymptotic normality assumptions. However, in practice such asymptotic approximations can be inaccurate, particularly in smaller sample sizes and when p is moderate to large. Elliptical regions calculated assuming normality of the posterior can be badly misaligned with the contours of the exact posterior, and can contain substantially less than $1-\alpha$ posterior probability. Instead our focus is on identifying a minimum volume ellipsoid that contains at least $1-\alpha$ probability, but will only correspond to the Gaussian elliptical region when the posterior is Gaussian.

An ellipsoid in p dimensions can be formulated as the set of points E satisfying

$$E = \{\theta : (\theta - \theta_0)' M (\theta - \theta_0) \leq 1\}, \tag{4.1}$$

where θ_0 is the center and M is the $p \times p$ shape matrix. A spectral decomposition yields $M = D\Lambda D'$, where D is a matrix having orthonormal eigenvectors as columns, and the diagonal matrix Λ has its elements as the inverse of the square of the axis lengths. A minimum volume covering ellipsoid (MVCE) is a special type of ellipsoid designed to enclose a given set of points in $\Re^p$ (in our case, the HPD set) with an ellipsoid of minimum volume and can be formulated as

$$\min_{M,\theta_0} \det(M^{-1}),\ (\theta - \theta_0)' M (\theta - \theta_0) \leq 1,\ \theta \in \text{HPD set}, M \text{ is positive definite.} \tag{4.2}$$

By definition, the MVCE credible regions for multivariate normal posteriors converge to the exact hyperelliptical $100(1-\alpha)\%$ HPD credible region in the limit as the number of MCMC samples increase. For posteriors with nearly hyperelliptical HPD credible regions, the method is expected to yield credible regions with a slightly higher

posterior probability than the nominal level $(1-\alpha)$, with the difference depending on the proportion of points included in the tails outside the HPD region. Even when the true HPD region deviates substantially from a hyperellipse, the MVCE credible region is expected to preserve the orientation of the HPD credible region (with respect to the coordinate axes) for a wide variety of cases. The same can not be said of hyper-rectangular regions. Finally, the proposed approach also meets our objectives in terms of storing the credible region in a relatively inexpensive manner and being able to quickly check if particular parameter values of interest are contained inside the credible region using the parameters of the computed MVCE.

We use the approximate algorithm of Khachinayan (1996) to compute the MVCE. The algorithm (presented in the appendix) computes the MVCE iteratively, is computationally fast for moderate number of knots and scales well to higher dimensions. For n MCMC samples, Khachiyan's algorithm computes a $(1+\eta)$-approximation to the MVCE in $\mathrm{O}(np^2([(1+\eta)^{2/(p+1)}-1]^{-1}+\log(p)+\log\log(n)))$ time, $\eta \in [0,1]$. Depending on the specified tolerance limit, the relative ratio n/p, and the shape of the true HPD region, the algorithm might exclude some points near the boundary of the true HPD region or might include additional points outside the HPD region. Simulations suggest that the method has near optimal performance when the HPD credible regions are elliptical. For non-Gaussian data, the contrast between our MVCE approach, and an approach based on regions assuming asymptotic normality is illustrated through a two dimensional example in Figure C.1. It can be clearly seen that the MVCE approach has probability content greater than the nominal value of 0.95, whereas the approach based on asymptotic normality has a much lower probability content.

4.2 Credible regions for one dimensional curves

In this section, we extend our consideration to real-valued functions $f : [a, b] \to \Re$. The methods described in the previous section can be applied directly to obtain a credible region for the function evaluated at a set of grid points or knots, $a = x_1^* < \ldots < x_k^* = b$. One can use the resulting elliptical region to represent posterior uncertainty in f. To investigate whether a function of interest, f_0, is contained in the credible region, one can assess whether $f_0(\mathbf{x}^*)$ is contained in the MVCE credible region for $f(\mathbf{x}^*)$. For a fine enough grid and sufficiently smooth functions, such grid-based inferences may provide an useful approximation. However, such grid based approaches fail to satisfy our goal of constructing infinite dimensional credible regions for the entire function over [a,b] and can be computationally expensive for a dense grid of points. Hence, it is appealing to obtain credible bands for the entire curve based on small to moderate number of knots, which provide a simple summary of uncertainty that is easy to plot and visualize.

Let $f \sim \Pi$, with Π being a prior over the function space $\mathcal{F}$, and let $\Pi(f|Y^n)$ denote the posterior. Denote MCMC samples from the posterior at $\mathbf{x}^* = (x_1^*, \ldots, x_k^*)$ as $\{f^j(\mathbf{x}^*), j = 1, \ldots, J\}$, where $f^j(\mathbf{x}^*)$=$(f^j(x_1^*), \ldots, f^j(x_k^*))$ is the vector of function values at the jth iteration. We first apply the method of Section 4.1 to estimate an elliptical credible region for $f(x_1^*), \ldots, f(x_k^*)$ (denoted as MVCE($\mathbf{x}^*$)). Subsequently we estimate the upper and lower credible limits at the mth knot as

$$\{\tilde{f}^U(x_m^*) = \max_j f^j(x_m^*), \quad \tilde{f}^L(x_m^*) = \min_j f^j(x_m^*), \quad f^j(\mathbf{x}^*) \in \text{MVCE}(\mathbf{x}^*)\}, m = 1, \ldots, k, \tag{4.3}$$

noting that the posterior probability allocated to the region inside MVCE($\mathbf{x}^*$) but outside the hyperectangle specified by the credible limits is vanishingly small for large J. We now interpolate between these credible limits at the knots to obtain the upper and lower credible bands, thus obtaining infinite dimensional credible regions. Finally,

our simultaneous credible region is defined as the set of all functions contained within MVCE($\mathbf{x}^*$) at the knots and within the credible bands in between the knots. Thus, we obtain credible regions of the form

$$\left\{ f : f(\mathbf{x}^*) \in \text{MVCE}(\mathbf{x}^*), \quad \tilde{f}^L(x) \leq f(x) \leq \tilde{f}^U(x), \quad x \notin \mathbf{x}^*, x \in (x_1^*, x_k^*) \right\}. \quad (4.4)$$

We consider linear interpolation and interpolation based on Lipschitz continuous functions, each having their own justifications.

To maintain the posterior probability allocated to MVCE($\mathbf{x}^*$), it is important to interpolate between the knots in such a manner that the conditional posterior probability assigned to aberrant curves that are contained within credible limits at knots but have at least one violation between the knots is small. If the support of Π is Lipschitz (denoted as $\mathcal{F}_{c_L}$) implying that $||f(x) - f(x')|| \leq c_L||x - x'||$ for all $x, x' \in [a, b]$ and $0 < c_L < \infty$, our interpolation approach based on Lipschitz continuous functions guarantees the conditional posterior probability for aberrant curves goes to zero. This is achieved by inflating the widths of the intervals between the knots so that only curves having absolute slope greater than c_L are ruled out, by constructing credible bands in the following manner

$$\tilde{f}^U_{c_L}(z) = \tilde{f}^U(u) + c_L|z - u|, \quad \tilde{f}^L_{c_L}(z) = \tilde{f}^L(u) - c_L|z - u|, \quad z \in [x_1^*, x_k^*] \quad (4.5)$$

where $u = \arg\min_{t \in \mathbf{x}^*}|z - t|$ and $\{\tilde{f}^t(u) : u \in \mathbf{x}^*, \quad t = U, L\}$ are the credible limits at the knots $\mathbf{x}^*$. Let us denote the credible region obtained by using $\tilde{f}^t_{c_L}$ (t=U,L) in (4.4)

as $\mathcal{F}_{CR}(\mathbf{x}^*)$ and let $\mathcal{F}_{\mathbf{x}^*} = \{f : \tilde{f}^L(x_m^*) \leq f(x_m^*) \leq \tilde{f}^U(x_m^*), m = 1, \ldots, k\}$. Then

$$\begin{aligned}
\Pi\Big(\mathcal{F}_{CR}(\mathbf{x}^*)\Big|Y^n\Big) &= \Pi\Big(\mathcal{F}_{CR}(\mathbf{x}^*) \cap \mathcal{F}_{\mathbf{x}^*}\Big|Y^n\Big) + \Pi\Big(\mathcal{F}_{CR}(\mathbf{x}^*) \cap \mathcal{F}_{\mathbf{x}^*}^c\Big|Y^n\Big) && (4.6)\\
&\approx \Pi\Big(MVCE(\mathbf{x}^*) \cap \mathcal{F}_{c_L} \cap \mathcal{F}_{\mathbf{x}^*}\Big|Y^n\Big) && (4.7)\\
&= \Pi\Big(MVCE(\mathbf{x}^*) \cap \mathcal{F}_{\mathbf{x}^*}\Big|Y^n\Big) \approx \Pi\Big(MVCE(\mathbf{x}^*)\Big|Y^n\Big), && (4.8)
\end{aligned}$$

where we use the fact that $\Pi\Big(f \in \text{MVCE}(\mathbf{x}^*) \cap \mathcal{F}_{\mathbf{x}^*}^c\Big|Y^n\Big) \approx 0$ for large J.

In practice, it may be too restrictive to assume in advance that $\mathcal{F}$ is Lipschitz with known c_L, but it might be reasonable to assume the following:
(A) It is possible to estimate some $\hat{c}_{k,L}$ such that $\mathcal{F} \cap \mathcal{F}_{\hat{c}_{k,L}}^c$ has vanishingly small posterior probability for large k, J, with $\mathcal{F}_{\hat{c}_{k,L}}$ = set of Lipschitz functions with constant $\hat{c}_{k,L}$.
We estimate $\hat{c}_{k,L}$ as $\hat{c}_{k,L} = \max_{j\in\{1,\ldots,J\}} s^j(\mathbf{x}^*)$, where $s^j(\mathbf{x}^*)$ is the maximum absolute slope evaluated over $\mathbf{x}^*$ for the jth MCMC iteration. We now construct credible bands $\tilde{f}_{\hat{c}_{k,L}}^t$ (t=U,L) similarly as in (4.5), but with c_L replaced by $\hat{c}_{k,L}$, and we denote the corresponding credible region as $\hat{\mathcal{F}}_{CR}(\mathbf{x}^*)$. Under assumption (A), such simultaneous credible region is designed to have posterior probability close to MVCE($\mathbf{x}^*$), as shown by the following arguments.

Let us express the support of Π as $\mathcal{F} = \mathcal{F}_{\hat{c}_{k,L}} \bigcup \mathcal{F}_{\hat{c}_{k,L}}^c$, with the decomposition varying as k, J change. It is straightforward to see that the set of aberrant curves can be expressed as $\mathcal{F}_{ab} = \mathcal{F}_{\hat{c}_{k,L}}^c \cap \mathcal{F}_{\mathbf{x}^*}$. Since $\hat{c}_{k,L}$ is non-decreasing in k, J, the size of the set of such aberrant curves is non-increasing in k, J. Under assumption (A), there exists constants k_0, J_0, such that $\Pi\Big(f \in \mathcal{F}_{\hat{c}_{k,L}}^c\Big|Y^n\Big) \leq \epsilon$ for $k > k_0$, $J > J_0$. We have

$$\begin{aligned}
\Pi\Big(f \in \hat{\mathcal{F}}_{CR}(\mathbf{x}^*)\Big|Y^n\Big) &= \Pi\Big(\text{MVCE}(\mathbf{x}^*) \cap \mathcal{F}_{\hat{c}_{k,L}} \cap \mathcal{F}_{\mathbf{x}^*}\Big|Y^n\Big) + \Pi\Big(\text{MVCE}(\mathbf{x}^*) \cap \mathcal{F}_{\hat{c}_{k,L}} \cap \mathcal{F}_{\mathbf{x}^*}^c\Big|Y^n\Big)\\
&= \Pi\Big(\text{MVCE}(\mathbf{x}^*) \cap \mathcal{F}_{\mathbf{x}^*}\Big|Y^n\Big) - \Pi\Big(\text{MVCE}(\mathbf{x}^*) \cap \mathcal{F}_{ab}\Big|Y^n\Big) + \epsilon_2\\
&= \Pi\Big(\text{MVCE}(\mathbf{x}^*)\Big|Y^n\Big) - \epsilon_1 + \epsilon_2, \text{ for } k > k_0, J > J_0, \text{ and } 0 < \epsilon_1 < \epsilon,
\end{aligned}$$

where we use the fact that $\Pi\Big(\text{MVCE}(\mathbf{x}^*) \cap \mathcal{F}_{\hat{c}_{k,L}} \cap \mathcal{F}^c_{\mathbf{x}^*} \Big| Y^n\Big) \approx 0$ for large J. When $\epsilon_1 \approx 0$ under (A) and $\epsilon_2 \approx 0$, we have $\Pi\Big(f \in \hat{\mathcal{F}}_{CR}(\mathbf{x}^*) \Big| Y^n\Big) \approx \Pi\Big(\text{MVCE}(\mathbf{x}^*) \Big| Y^n\Big)$.

As an alternative approach which is free from any underlying assumptions on the support of Π, we propose the linear interpolation approach having asymptotic justifications in the special case when Π is a Gaussian process (GP). The linear interpolation approach constructs credible bands as:

$$\tilde{f}^t(z) = \frac{x^*_m - z}{\Delta_m}\tilde{f}^t(x^*_{m-1}) + \frac{z - x^*_{m-1}}{\Delta_m}\tilde{f}^t(x^*_m), z \in D_m = [x^*_m, x^*_{m+1}), \quad \Delta_m = |x^*_m - x^*_{m-1}| \quad (4.9)$$

where $\tilde{f}^t(x^*_m)$ (t=L, U), m=1,...,k, are the credible limits at the knots computed as in (4.3). First, we state the following prior approximation result which is the basis for our linear interpolation approach.

Theorem 8 *Suppose* Π *is a zero mean GP with squared exponential covariance kernel and define* $w^j(z) = \sum_{m=2}^{k} E[f(z)|f^j(x^*_{m-1}), f^j(x^*_m)]I(z \in D_m)$, *j=1,...,J. Then for* $z \in D_m$, $w^j(z) \approx \frac{x^*_m - z}{\Delta_m} f^j(x^*_{m-1}) + \frac{z - x^*_{m-1}}{\Delta_m} f^j(x^*_m)$ *when* $(\Delta_m)^b \approx 0, b \geq 4$.

Theorem 8 intuitively suggests that when $z \in [x^*_{m-1}, x^*_m)$ and in the limiting case when the distance between the knots is small ($\Delta^b_m \approx 0, b \geq 4$), $f(z)$ can be approximated by a linear combination of $f^j(x^*_{m-1}), f^j(x^*_m)$ (for the jth iteration) under an appropriate GP prior. Hence given posterior samples in MVCE($\mathbf{x}^*$), it would be meaningful to construct credible bands designed to contain pairwise linear combinations of $f^j \in \text{MVCE}(\mathbf{x}^*)$ as:

$$\text{for } z \in D_m, \tilde{f}^U(z) = max_{f^j \in \text{MVCE}(\mathbf{x}^*)} \left\{ \frac{x^*_m - z}{\Delta_m} f^j(x^*_{m-1}) + \frac{z - x^*_{m-1}}{\Delta_m} f^j(x^*_m) \right\}$$
$$\tilde{f}^L(z) = min_{f^j \in \text{MVCE}(\mathbf{x}^*)} \left\{ \frac{x^*_m - z}{\Delta_m} f^j(x^*_{m-1}) + \frac{z - x^*_{m-1}}{\Delta_m} f^j(x^*_m) \right\}. \quad (4.10)$$

However in practice, the credible region specified by (4.10) might have posterior

probability content much smaller than $1-\alpha$ when the knots are not close enough. To adjust for this, we construct a more conservative credible region by linearly interpolating between the credible limits at knots, thus obtaining piecewise linear credible bands as defined in (4.9).

4.3 Functions with vector valued arguments

In the case of functions over the line, there is a natural ordering of knots which can be used to construct the credible bands. However for functions over a surface $f : S \to \Re$ (S is a convex hull of the sample design points in $\Re^d$, $d \geq 2$), such a natural ordering is lost. A meaningful way of connecting the points in a set $S \subset \Re^d$ ($d \geq 2$) is through *triangulations*, which refers to a subdivision of S such that the bounded faces are d-simplices, and the vertices are points in S. For example when $d = 2$, a triangulation would correspond to a planar subdivision whose bounded faces are triangles having vertices as points in $S \subset \Re^2$.

For a given set of points in S, such triangulations can be achieved in many possible ways. One optimal method is the Delaunay triangulation (DT), which ensures that the circumcircle of any triangle in the triangulation does not contain any additional point of the design set S. DT is optimal in several respects, including maximizing the minimum angle and minimizing the maximum circumcircle over all possible triangulations of S (Fortune, 1992). Thus DT has become an important tool for high quality mesh generation for a finite point set (Bern and Eppstein, 1992).

Based on DT, we construct piece-wise hyperplanar credible surfaces as follows: for the vth d-simplex having vertices $(s_{0,v}, \ldots, s_{d,v})$ obtained by the DT, we specify the credible surface using a hyperplane passing through the $(d+1)$ dimensional points $(s_{0,v}, \tilde{f}^t(s_{0,v}))$,$\ldots$,$(s_{d,v}, \tilde{f}^t(s_{d,v}))$ (t=U, L). For example in the two dimensional case, we would obtain the upper and lower credible surfaces as piece-wise hyperplanar surfaces

connected together at the edges of the neighboring triangles obtained from DT. These piece-wise hyperplanar credible surfaces can be considered as higher dimensional generalization of the piece-wise linear credible bands obtained using the linear interpolation approach in section 4.2. We can use the equations of the credible hyperplanes to check if a parametric function defined on the convex hull of $S \subset \Re^d$ $(d \geq 2)$ is contained within the simultaneous credible region - an useful tool for hypothesis testing involving higher dimensional curves.

Let $S' \in \Re^{d+1}$ denote the (d+1)-dimensional set constructed by lifting the points in S to a paraboloid in $\Re^{d+1}$. For example in two dimensions, a point $(z_1, z_2) \in S$ is lifted to $(z_1, z_2, z_1^2 + z_2^2) \in S'$. Then, DT of S can be computed as the projection of the downward facing faces of the convex hull of S'. Thus, most of the algorithms for Delaunay triangulations are based on computing convex hulls. We shall use the 'geometry' package in R to implement DT, which essentially uses the Quickhull algorithm to compute convex hulls. Obviously this would entail additional computation time compared to the approaches for one dimensional curves.

4.4 Simulation Studies

4.4.1 One Dimensional Functions

To assess how our interpolation methods work in practice, we generate data involving a one-dimensional mean function with an additive residual, using the following model:

$$y_i = \frac{1}{\max(x_i^2, 0.1)} + \cos\left(\pi(\frac{x_i - 5}{10})\right) + \sum_{k=1}^{20} w_k \exp(-|x_i - t_k^*|) + \epsilon_i, \quad \epsilon_i \sim N(0,1),$$
$$x_i \sim U(-5, 10), \quad t_k^* \sim U(\min(\mathbf{x}) + 0.5, \max(\mathbf{x}) - 0.5), \quad w_k \sim N(0, 10),$$

where $\mathbf{x}=(x_1,\ldots,x_n)$. We use a sample size of 100 for our simulations. We generate 100 replicates, and for each replicate we fit the data using the following Gaussian process mean regression model:

$$\begin{aligned} y_i &= f(x_i)+\epsilon_i, \quad \epsilon_i \sim N(0,\tau^{-1}), \quad \tau \sim Ga(a_\tau, b_\tau), \\ f(\cdot) \sim GP(0,K), K(x_i,x_j) &= \phi_1^{-1}\exp(-\phi_2(x_i-x_j)^2), \quad \phi_1 \sim Ga(a_{\phi_1}, b_{\phi_1}), \end{aligned} \tag{4.11}$$

and we use a Metropolis random walk to update ϕ_2. The posterior computation proceeds via straightforward Gaussian process mean regression algorithm.

We assess the performance of our method by computing the frequentist coverage of the mean function by the credible regions constructed using the method relying on Lipschitz continuity (Lip) and the linear interpolation method (Lin). The frequentist coverage by the credible region is computed as the percentage of replicates (out of a total of 100 replicates) when the credible bands completely contain the true mean function evaluated at a fine grid of 10000 points lying within $[x_{(1)}, x_{(n)}]$, the extreme values of $\mathbf{x}$. We also compute the frequentist coverage by the finite dimensional credible region at the knots. As a measure of the size of the infinite dimensional credible region, we determine the area enclosed inside the credible bands, as well as compute the volume of the finite dimensional credible region at the knots. We assess the performance over varying number of knots, and we choose knot locations as equispaced sample quantiles. When using number of knots greater than sample size, we choose equispaced knots.

As a comparison, we also construct credible bands by computing credible limits at the knots using Crainiceanu et al. (2007) method (CA) and subsequently interpolating relying on Lipschitz continuity as well as linearly. In addition, we consider the approach based on asymptotic normality (Pnorm) which constructs a credible region similar to the MVCE approach in (4.4), but with parameters $\theta_0 = \frac{1}{T-B}\sum_{j=B+1}^{T} f^j(\mathbf{x}^*)$ (mean of

posterior samples at knots after burn in) and shape matrix $M = \left(\frac{1}{T-B} \sum_{j=B+1}^{T} (f^j(\mathbf{x}^*) - \theta_0)(f^j(\mathbf{x}^*) - \theta_0)' \right)^{-1}$. Here T is the total number of MCMC iterations and B is the burn in.

As an alternative, we also considered applying Crainiceanu et. al. (2007) approach directly to the fine grid of 10000 points to compute credible limits, and subsequently checking if the true mean function at those grid points lie within the credible limits. Although such a sequence of credible limits do not yield a simultaneous credible region for the true mean function (which is our primary objective), it can be considered as a finite dimensional approximation. However, the implementation of this approach is expensive requiring krigging at 10000 points and the computational burden increases for increasing sample sizes. We found that even for moderate sample sizes, such an approach is not practically applicable. Hence we do not consider it further.

While computing the density level sets for the MVCE approach, the conjugate priors specified in (4.11) allow us to marginalize over the nuisance parameter τ and obtain the likelihood approximation at the knots $\mathbf{x}^* \subseteq \mathbf{x}$ as $\hat{L}(Y^*, f(\mathbf{x}^*))$

$$\propto \left(b_\tau + \frac{(Y^* - f(\mathbf{x}^*)'(Y^* - f(\mathbf{x}^*)}{2} \right)^{-(a_\tau + n/2)} \exp(-\frac{1}{2} f(\mathbf{x}^*)' \hat{K}^{-1} f(\mathbf{x}^*)), \qquad (4.12)$$

where $Y^* = \{y_i : x_i \in \mathbf{x}^*\}$, and $\hat{K}(\mathbf{x}^*)$ is the covariance kernel of the Gaussian process evaluated at $\mathbf{x}^*$ with elements $\hat{K}(x_i^*, x_j^*) = \hat{\phi}_1^{-1} \exp(-\hat{\phi}_2 (x_i^* - x_j^*)^2)$ and $\hat{\phi}_1, \hat{\phi}_2$ are the posterior means of the hyperparameters ϕ_1, ϕ_2. As an alternative, we considered computing the above likelihood using a Monte Carlo approximation involving realizations of ϕ_1, ϕ_2, from their respective priors. However such an approach is computationally expensive, and moreover, we obtain desirable frequentist coverage using the likelihood approximation (4.12).

Table C.1 summarizes the results from our simulation study. We see that for small

to moderate number of knots and under the interpolation approach based on Lipschitz continuity, the frequentist coverage at the knots as well as for the simultaneous credible region under the MVCE is greater than the nominal value. On the other hand, the corresponding frequentist coverage under the competing approaches are far lower than the desired nominal value. However, the performance of the competitors improve for high number of knots, yielding frequentist coverage close to the nominal value, as evident from Table C.1.

For linear interpolation, the frequentist coverage is poor for all the approaches when the number of knots is small or moderate, and seems to be sensitive to the location of the knots. From Table C.1 we note that for small to moderate number of knots, the frequentist coverage by the MVCE approach is far higher than the competitors. However, the frequentist coverage increases considerably for high number of knots, as reported in Table C.1. In general, it is our experience in simulations, that for high number of knots and a true mean function which is not terribly bumpy, the Lipschitz continuity approach and the linear interpolation approach seem to perform well, achieving a frequentist coverage close to the nominal value. On the other hand, for smooth true mean function, a frequentist coverage greater than or equal to the nominal value is attained for smaller number of knots.

As is obvious, we see from Table C.1 that the area under the credible bands for Lip decreases with the number of knots. We also see that the credible bands for the MVCE approach is wider than the competitors, across varying number of knots. This is probably one of reasons for the greater frequentist coverage for the MVCE approach. However, we note that the difference in area under the credible bands between the MVCE and competitors decreases as the number of knots increase.

4.5 Discussion and Future Directions

In this chapter, we attempt to address an important but understudied problem of constructing simultaneous credible regions for vectors and functions, guaranteed to have a prespecified posterior probability content. The proposed methodology is based on the joint posterior distributions of vector valued parameters and functions evaluated at a finite number of knots, and hence bypasses the approach of computing simultaneous credible regions based on marginal distributions, which is clearly inadequate from a philosophical as well as practical perspective. Among other uses, our methodology can be applied in hypothesis testing examples where we can use the simultaneous credible region to test a point null for the vector case, and test if two functions are significantly different for the function case. The proposed methodology is easy to implement and is not computationally expensive for small to moderate number of knots.

Currently we are working on applying our approach to the premature delivery application discussed in Chapter 2. Our analysis in this chapter is different from Chapter 2, in that we now model the probability of preterm birth with DDE as a covariate, using logistic regression involving a unknown log odds function modeled using a Gaussian process. Our aim is to construct simultaneous credible region for this unknown log odds function and to obtain an estimate of the DDE dose where the lower credible band crosses zero. This will provide us with an estimate of the lowest dose at which DDE significantly impacts preterm delivery.

Chapter 5

Future Directions

My future research seeks to extend the Gaussian process latent variable model proposed in Chapter 2 to include predictors in a manner which will allow variable selection. Variable selection using Gaussian process mean regression has been proposed in the literature (Linkletter et al., 2006; Zou et al., 2010; Stavitsky et al., 2011) and usually proceeds by examining the posterior summaries of the length scale parameter of the covariance kernel. Such models have important applications including detecting significant genes in multiple quantitative loci mapping with epistasis and gene-environment interactions. However, the drawback of such approaches using Gaussian process mean regression is that the residual density after adjusting for the unknown mean function is assigned a Gaussian distribution, which is restrictive. Such a restrictive structure does not allow the shape of the distribution to change with predictors and hence might impact variable selection adversely.

We propose the following model which combines the flexibility of the proposed Gaussian process latent variable model in Chapter 2 with variable selection approaches involving the standard Gaussian process mean regression model:

$$y_i = \eta(u_i, \mathbf{x}_i) + \epsilon_i, \quad u_i \sim U(0,1), \quad \mathbf{x}_i \sim \Re^d, d \geq 1, \tag{5.1}$$

where $\eta \sim GP(0, K)$ and K is the covariance kernel of the Gaussian process. By choosing a suitable covariance kernel and assigning mixture priors to suitable hyperparameters as in Savitsky et. al. (2011), we can proceed with variable selection. Further, the resulting conditional distribution allows the shape of the density to change with predictors, thus yielding greater flexibility.

While model (5.1) is expected to be flexible in detecting important predictors, it has the disadvantage of not being able to point out important interactions between covariates. For example in multiple quantitative loci studies, besides being able to study any potential main effects, it is often times of interest to study interactions (of arbitrary order) between markers and gene-environment interactions. Such interactions provide investigators greater understanding of the more complex underlying genetic pathways and gene-environment interactions. Hence, we are currently working on extending the above model to develop an approach which can select predictors having significant main effects, as well as detect significant interactions between predictors.

Appendix A

Chapter 2

Proof of Theorem 1:

We have $KL(f_{\mu,\sigma}(y), f_0(y)) = \int f_0 \log \frac{f_0}{f_{\mu,\sigma}} = \int f_0(y) \left[\frac{\lim_{\sigma\to 0} \int_0^1 \Gamma_\sigma(y-\mu_0(x))dx}{\int_0^1 \Gamma_\sigma(y-\mu(x))dx} \right] dy.$ Now for all $y \in \Re$ and fixed $\sigma \in \Re^+$,

$$\frac{f_{\mu_0,\sigma}}{f_{\mu,\sigma}} = \frac{\int_0^1 \Gamma_\sigma(y-\mu_0(x))dx}{\int_0^1 \Gamma_\sigma(y-\mu(x))dx} \leq sup_{x\in(0,1)} h_\sigma\left(y, \mu\left(x\right) - \mu_0\left(x\right)\right) \to 1, \text{ as } \frac{||\mu-\mu_0||_\infty}{\sigma^2} \to 0,$$

where $\mathrm{h}_\sigma(y, \mu-\mu_0) = e^{\frac{1}{2\sigma^2}(\mu-\mu_0)^2 - \frac{1}{\sigma^2}(y-\mu_0)(\mu-\mu_0)}$ for Γ_σ = Gaussian, while $\mathrm{h}_\sigma(y, \mu-\mu_0) = e^{\frac{1}{\sigma}(|\mu_0-\mu|)}$ for Γ_σ = Laplace. Hence $\lim_{\sigma\to 0} \log \frac{f_0(y)}{f_{\mu,\sigma}(y)} \to 0$, as $\frac{||\mu-\mu_0||_\infty}{\sigma^2} \to 0$. Under *(A1)* and Gaussian or Laplace residuals, $\log \frac{f_0}{f_{\mu,\sigma}}$ is bounded for all $\sigma \in \Re^+$ and μ close to μ_0. Hence we can use dominated convergence theorem to obtain,
$\lim_{\sigma\to 0, \frac{||\mu-\mu_0||_\infty}{\sigma^2}\to 0} \int_\Re f_0(y) \log \frac{f_0(y)}{f_{\mu,\sigma}(y)} dy \to 0$. Hence we can choose a suitably small η_1, η_2, η_2^* with $0 < \eta_1 < \eta_2 < \eta_2^*$ such that $\left\{ ||\mu - \mu_0||_\infty \leq \eta_1, \eta_2^* < \sigma \leq \eta_2 \right\} \Rightarrow KL(f_{\mu,\sigma}, f_0) \leq \epsilon$. Given positive support of priors Π^* and ν, we have $\Pi(KL_\epsilon(f_0)) > 0$.

Proof of Theorem 2:

We can use Taylor's series expansion to obtain,

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

where $\delta_1^y\left(n_0\right) - \delta_2^y\left(n_0\right)$ is bounded and decreases with n_0 under $(A1)$, for μ close to μ_0 and Γ_σ = Gaussian. Using the identity a^n-b^n=(a-b)$(\sum_{k=1}^n a^{n-k}b^{k-1})$, and denoting

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

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

where $K(n_0)=\sum_{k=1}^{n_0} \sup_y |\frac{(-1)^k}{k}(\sum_{l=1}^{k} g_{\mu,\sigma}^{k-l} g_0^{l-1})|$ is a finite constant depending on n_0 under $(A1)$, for μ close to μ_0 and Γ_σ = Gaussian. Further, using similar methods as in the proof of theorem 3, we can show that for $0 < \epsilon_1 < \epsilon_2 < \epsilon_2^*$,

$$\{\mu \in N_{\epsilon_1}(\mu_0), \sigma \in (\epsilon_2, \epsilon_2^*)\} \Rightarrow \int |f_0(y) - f_{\mu,\sigma}(y)|dy < \frac{\epsilon_1}{\epsilon_2}. \tag{A.1}$$

Using inequality $(A.1)$, we have for $\mu \in N_{\epsilon_1}(\mu_0)$ and $\sigma \in (\epsilon_2, \epsilon_2^*)$,

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

for suitably small $\Delta(n_0) = \int f_0(y)|\delta_1^y(n_0) - \delta_2^y(n_0)|dy$. The rest follows since $\Pi^* \otimes \nu \{\mu \in N_{\epsilon_1}(\mu_0), \sigma \in (\epsilon_2, \epsilon_2^*)\} > 0$

Proof of Theorem 3: For a fixed $\sigma \in \Re^+$, we have

$$\begin{aligned} \int |f_{\mu,\sigma} - f_{\tilde{\mu},\sigma}|dy \quad &\leq \quad \int\int_0^1 |\phi_\sigma(y - \mu(x)) - \phi_\sigma(y - \tilde{\mu}(x))|\text{dx dy} \\ &= \quad \int_0^1 \int |\phi_\sigma(y - \mu(x)) - \phi_\sigma(y - \tilde{\mu}(x))|\text{dy dx (Fubini's Theorem)} \\ &= \quad \int_{x|\mu_0>\mu} \int |\phi_\sigma(y - \mu(x)) - \phi_\sigma(y - \tilde{\mu}(x))|\text{dy dx} \\ &+ \quad \int_{x|\mu_0<\mu} \int |\phi_\sigma(y - \mu(x)) - \phi_\sigma(y - \tilde{\mu}(x))|\text{dy dx.} \end{aligned}$$

In the proof of lemma 1 of Ghosal, Ghosh and Ramamoorthy (1999), it was shown that for fixed $\theta_1 < \theta_2$, $||\phi_\sigma(y-\theta_1) - \phi_\sigma(y-\theta_2)|| < \frac{\theta_2-\theta_1}{\sigma}$ (where $||.|| = L_1$ norm), which would imply

$$\int |f_{\mu,\sigma} - f_{\tilde{\mu},\sigma}|dy \;\leq\; \int_{x|\tilde{\mu}>\mu} \frac{\tilde{\mu}(x)-\mu(x)}{\sigma}dx + \int_{x|\tilde{\mu}<\mu} \frac{\mu(x)-\tilde{\mu}(x)}{\sigma}dx = \int_0^1 \frac{|\mu-\tilde{\mu}|}{\sigma}dx \quad \text{(A.2)}$$

Under $(A1)$ and Γ_σ=Gaussian, $|f_{\mu,\sigma}(y) - f_{\tilde{\mu},\sigma}(y)|$ is bounded for μ close to $\tilde{\mu}$ and for all $\sigma \in \Re^+$. Thus, using dominated convergence theorem and (A.2), we have

$$\int \lim_{\sigma\to 0} |f_{\mu,\sigma} - \tilde{f}| = \lim_{\sigma\to 0} \int |f_{\mu,\sigma} - f_{\tilde{\mu},\sigma}| \leq \lim_{\sigma\to 0} \int_0^1 \frac{|\mu(x)-\tilde{\mu}(x)|}{\sigma}dx.$$

Hence, we can choose sufficiently small $\epsilon_1, \epsilon_2, \epsilon_2^*$ with $0 < \epsilon_1 < \epsilon_2 < \epsilon_2^*$ such that for $\mu \in N_{\epsilon_1}(\tilde{\mu})$ and $\sigma \in (\epsilon_2, \epsilon_2^*)$, we would have $\int |f_{\mu,\sigma} - \tilde{f}|dy < \frac{\epsilon_1}{\epsilon_2}$.

Proof of Theorem 4:

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

DEFINITION 1. For $\mathcal{G} \subset \mathcal{F}$ and $\delta > 0$, L-1 metric entropy $J(\delta, \mathcal{G})$ is defined as the minimum of $\log(k : \mathcal{G} \subset \cup_{i=1}^k \{f : \int |f - f_i|dy < \delta,\ f_1, f_2, \ldots, f_k \in \mathcal{F}\})$.

Theorem 5(Ghosal, Ghosh and Ramamoorthi) *Let Π be a prior on $\mathcal{F}$. Suppose $f_0 \in \mathcal{F}$ is in the Kullback-Leibler support of Π and let U=$\{f : \int |f - f_0|dy < \epsilon\}$. If there is a $\delta < \epsilon/4$, $c_1,\ c_2 > 0,\ \beta < \epsilon^2/8$ and $\mathcal{F}_n \subset \mathcal{F}$ such that for all large n:*

(1) $\Pi(\mathcal{F}_n^c) < c_1 \exp(-nc_2)$, and,

(2) The L-1 metric entropy, $J(\delta, \mathcal{F}_n) < n\beta$,

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

The constants δ, c_1, c_2, β *and* $\mathcal{F}_n$ *are allowed to depend on* ϵ.

Let U=$\left\{ f_{\mu,\sigma} : \int |f_{\mu,\sigma} - f_0| dy < \epsilon, \mu \in \Theta, \sigma \in (0, \infty) \right\}$. Let the parameter space for (μ, σ) be denoted as $\mathcal{H}$. Consider the subsets of the parameter space $\mathcal{H}_n = \mathcal{H}_{1n} \otimes \mathcal{H}_{2n}$, where $\mathcal{H}_{1n} = \{\mu : ||\mu||_\infty < M_n, ||\mu'||_\infty < M_n\}$ and $\mathcal{H}_{2n} = [L_n, \infty)$, with $L_n \to 0$ such that $\nu(\sigma \in (0, L_n)) < d_1 \exp(-d_2 n)$, $d_1, d_2 > 0$ and $M_n = O(n^{1/2})$. Using lemma 5 of Choi and Schervish (2004), the first derivative $\eta'(\cdot)$ exists and is also a Gaussian process under (*A2*). Using lemma 4 of Choi and Schervish (2004) who showed an upper bound on sup-norm metric entropy of $\mathcal{H}_{1n}$, we have the upper bound on L1 metric entropy as

$$J(\delta, \mathcal{H}_{1n}) < K_1 M_n / \delta. \tag{A.3}$$

This implies there are $\mathrm{K}^* = \exp(K_1 M_n/\delta)$ elements $\mu_1, \mu_2, \ldots, \mu_{K^*}$ such that

$$\mathcal{H}_{1n} \subset \cup_{j=1}^{K^*} \left\{ \mu : \int_0^1 |\mu - \mu_j| dx < \delta \right\}. \tag{A.4}$$

Let us consider the sieve $\mathcal{F}_n = \{f_{\mu,\sigma} \in \mathcal{F} : (\mu, \sigma) \in \mathcal{H}_n\}$. Clearly $\mathcal{F}_n \subseteq U$ and $\mathcal{F}_n \uparrow U$. Further, let us consider densities $\mathrm{f}_{i,n} = f_{\mu_i, L_n} \in \mathcal{F}_n$ defined as in section 2.1, where the $\mu_i, i = 1, \ldots, K^*$ correspond to the ones just defined to cover $\mathcal{H}_{1n}$. Using similar techniques as in lemma 1 of Ghosal, Ghosh and Ramamoorthi (1999), it can be shown that for fixed x and $\mu > \mu_i$, $\int |\phi_\sigma(y - \mu) - \phi_{L_n}(y - \mu_i)| dy$

$$\begin{aligned}
&= \frac{1}{\sqrt{2\pi}\sigma} \int_{y > \frac{\mu_i + \mu}{2}} \exp(-\frac{1}{2\sigma}(y-\mu)^2) - \frac{1}{\sqrt{2\pi}L_n} \int_{y > \frac{\mu_i + \mu}{2}} \exp(-\frac{1}{2L_n}(y-\mu_i)^2) \\
&+ \frac{1}{\sqrt{2\pi}L_n} \int_{y < \frac{\mu_i + \mu}{2}} \exp(-\frac{1}{2L_n}(y-\mu_i)^2) - \frac{1}{\sqrt{2\pi}\sigma} \int_{y < \frac{\mu_i + \mu}{2}} \exp(-\frac{1}{2\sigma}(y-\mu)^2) \\
&\leq 2\left(\frac{\mu - \mu_i}{\sqrt{2\pi}\sigma}\right) + 2\left(\frac{\mu - \mu_i}{\sqrt{2\pi}L_n}\right) \leq 4\left(\frac{\mu - \mu_i}{\sqrt{2\pi}L_n}\right), \text{ for small enough } L_n.
\end{aligned}$$

Similarly, for fixed x and $\mu < \mu_i$, $\int |\phi_\sigma(y-\mu) - \phi_{L_n}(y-\mu_i)|dy \leq 4\left(\frac{\mu_i-\mu}{\sqrt{2\pi}L_n}\right)$. This implies,

$$\int |f_{\mu,\sigma} - f_{i,n}|dy \leq 4\frac{1}{\sqrt{2\pi}L_n}\int_0^1 |\mu-\mu_i|dx \leq 4\frac{\delta}{\sqrt{2\pi}L_n}, \tag{A.5}$$

when $\int_0^1 |\mu-\mu_i|dx \leq \delta$. This clearly implies that (upto a constant)

$$J(\delta/L_n, \mathcal{F}_n) \;=\; J(\delta, \mathcal{H}_{1n}) < K_1 M_n/\delta \Rightarrow J(\delta, \mathcal{F}_n) \leq K_1 M_n L_n/\delta < n\beta, \tag{A.6}$$

where we can choose $\delta < \epsilon/4$ such that $\beta < \epsilon^2/8$. Thus the second condition in theorem 5 is satisfied. Also note that the prior probability of $\mathcal{F}_n$ can be calculated in terms of Π^* and ν. Under $(A2)$, we can use lemma 5 of Choi and Schervish (2004) to obtain $\Pi^*(\mathcal{H}_{1n}^c) \leq A\exp(-dM_n^2)$, where A, d$>$ 0. This implies

$$\Pi(\mathcal{F}_n^c) = (\Pi^* \otimes \nu)(\mathcal{H}_n^c) = (\Pi^* \otimes \nu)((\mathcal{H}_{1n}^c \otimes \mathcal{H}_{2n}) \cup (\mathcal{H}_{1n} \otimes \mathcal{H}_{2n}^c)) \leq c_1\exp(-c_2 n), c_1, c_2 > 0.$$

Thus the first condition in theorem 5 is satisfied. Hence $\Pi(U|Y_1, Y_2, \ldots, Y_n) \to 1$ a.s. P_{f_0}.

A.1 Tables

Table A.1: Marron-Wand Curves: L-1 error

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

Table A.2: Predictive MSE & L-1 error

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

A.2 Figures

Figure A.1: Prior realizations from the GPT for gestational
age at delivery (solid lines) along with frequentist kernel density estimate (dotted lines).
The rows correspond to ϕ_1=(0.01, 0.1); the columns correspond to ϕ_2=(0.1,1,25,100)

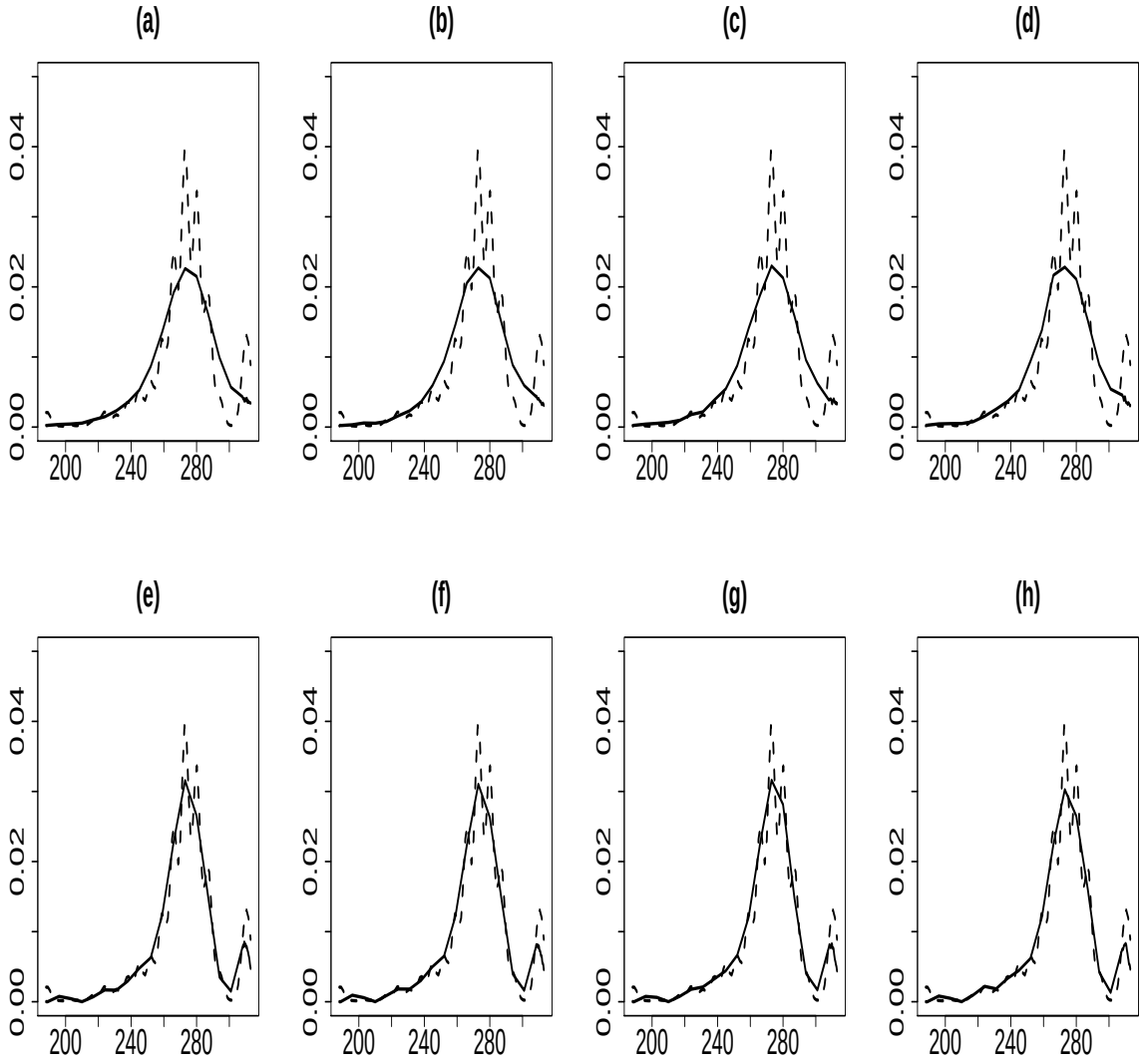


Figure A.2: Marron-Wand curves - density estimates for GPT, DPM and Polya tree mixtures

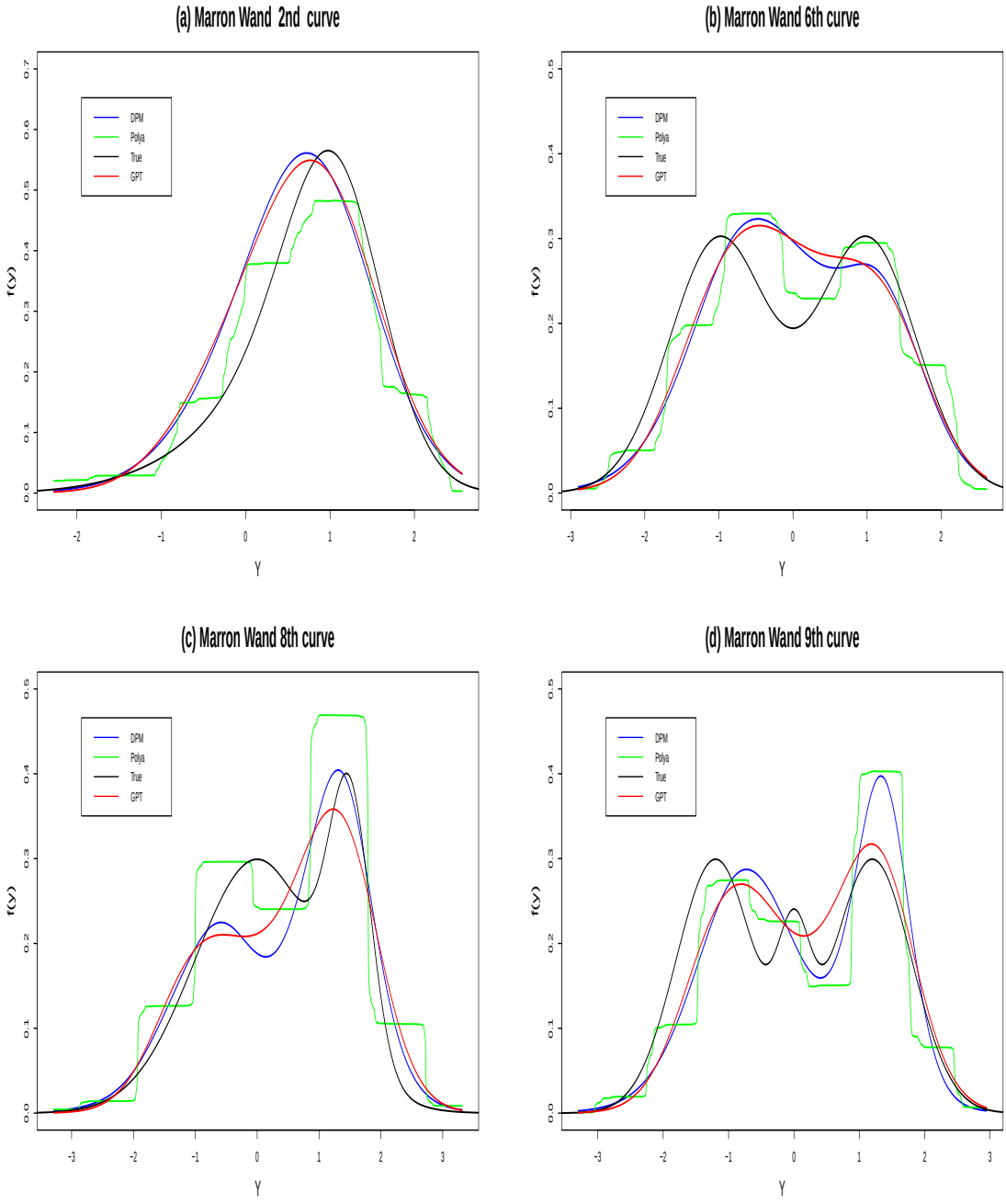


Figure A.3: GPT conditional density estimates and 90% credible intervals for 10th, 60th, 90th, 99th DDE quantiles. Vertical dashed line for cut-off at 37 weeks

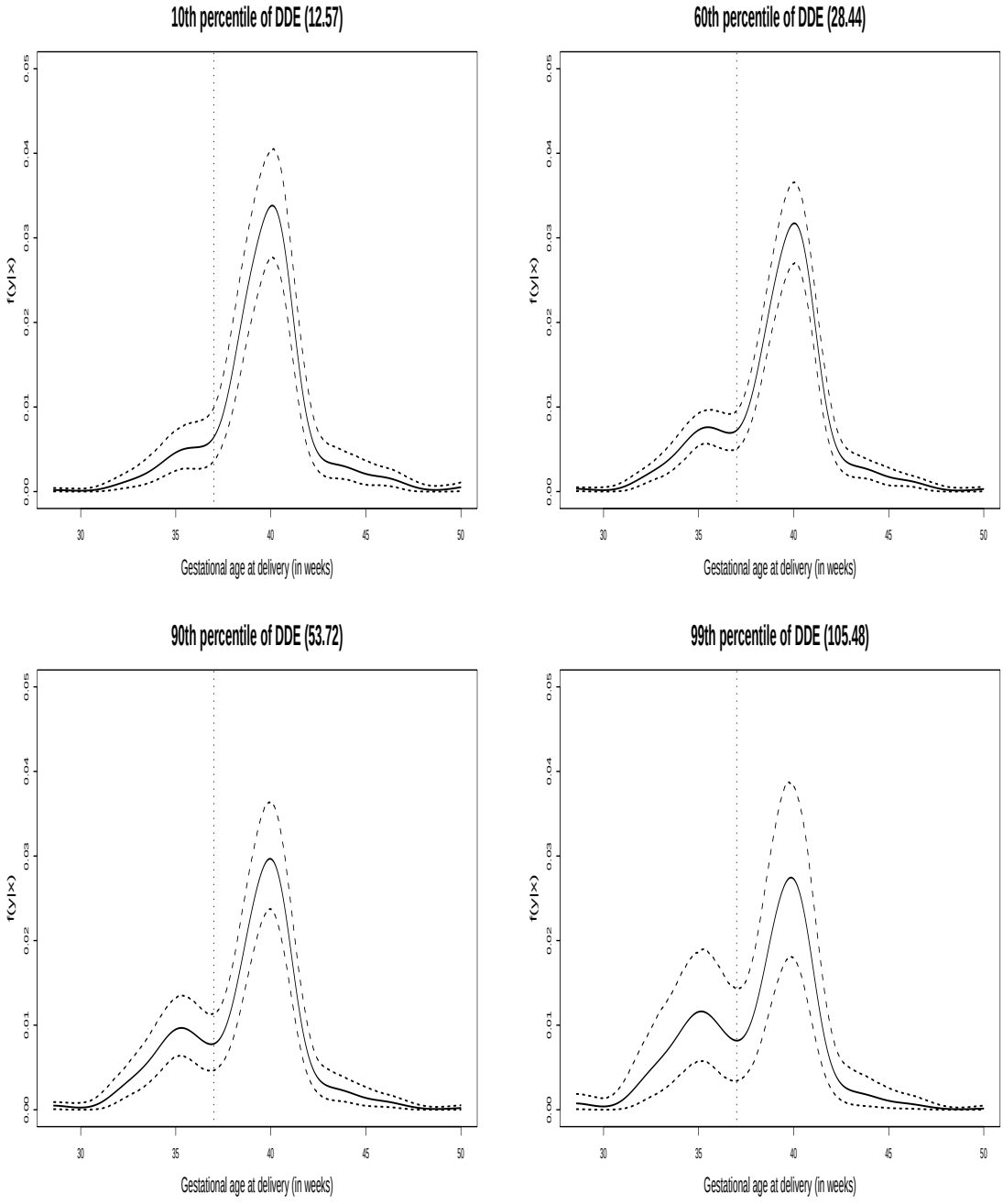


Figure A.4: DPM conditional density estimates and 90% credible intervals for 10th, 60th, 90th, 99th DDE quantiles. Vertical dashed line for cut-off at 37 weeks

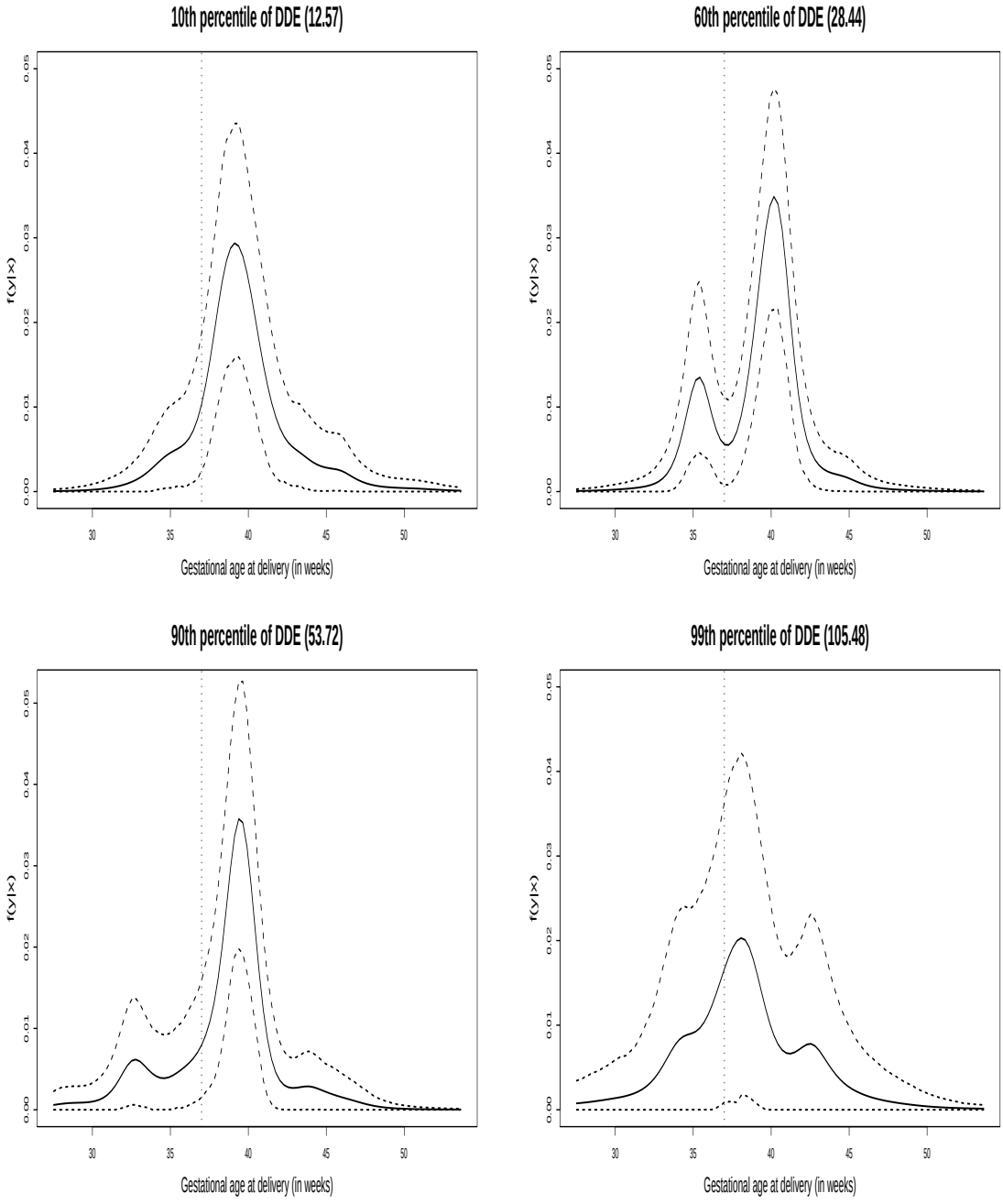
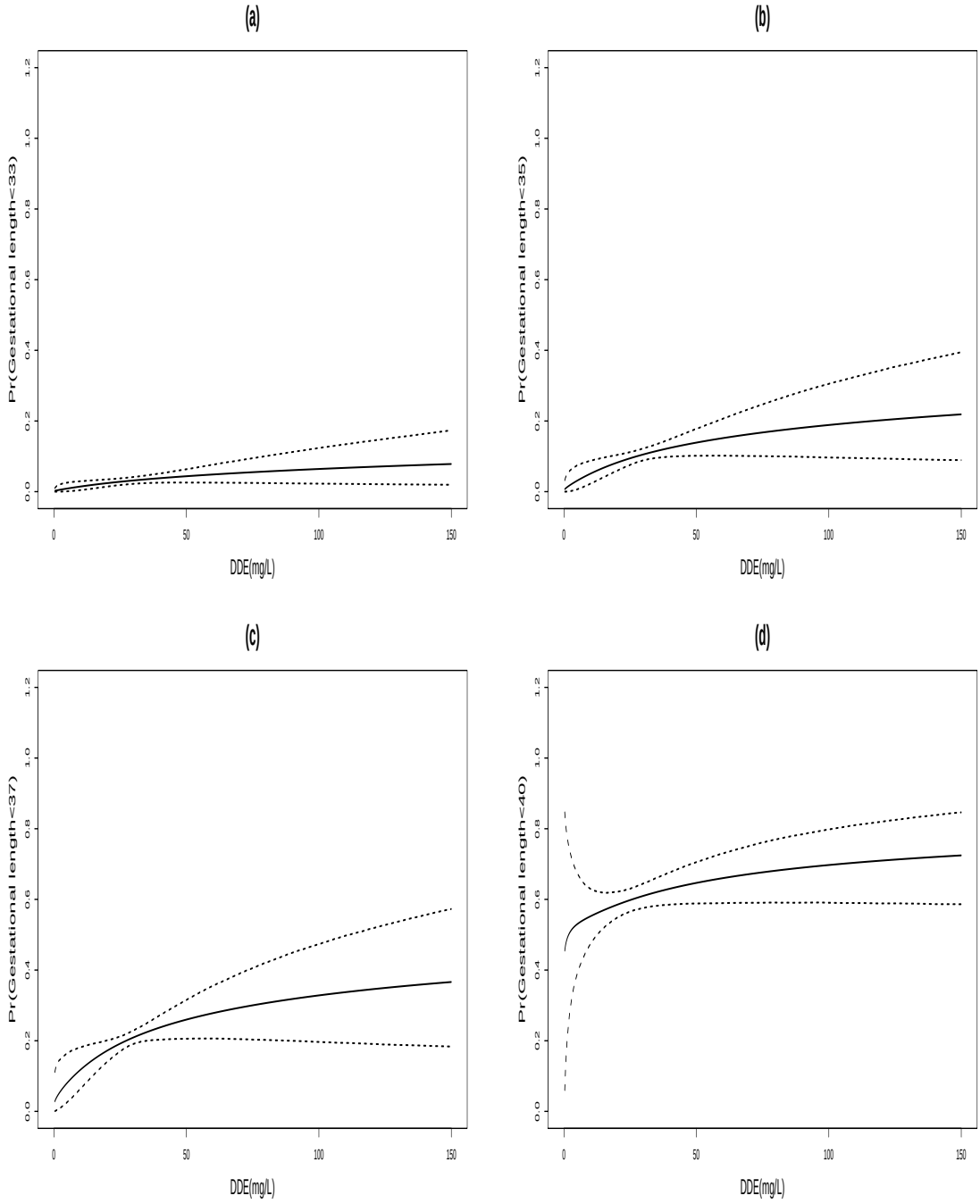


Figure A.5: Estimated probability that gestational age at delivery is less than T weeks versus DDE dose, for (a) T = 33, (b) T = 35, (c) T = 37, (d) T = 40. Solid lines are posterior means and dashed lines are pointwise 90% credible intervals



Appendix B

Chapter 3

Proof of Theorem 6: Using similar methods as in the proof of Theorem 2 in Guo and Speckman (2009), it can be shown that conditional on A and assumptions *(A3)* and *(A4)*, the upper and lower bounds of $L(Y^n|A,M_1)=\int_0^\infty(1+g)^{-p_1/2}\left[1-\frac{g}{1+g}\tilde{R}^2_{A,1}\right]^{-n/2}\pi(dg)$ are

$$\begin{aligned} L(Y^n|A,M_1) &\leq \left(\frac{p_1+2k_u}{n-p_1-2k_u}\right)^{p_1/2+k_u}\left(\frac{1-\tilde{R}^2_{A,1}}{\tilde{R}^2_{A,1}}\right)^{p_1/2+k_u}\left(\frac{n}{n-p_1-2k_u}\right)^{-n/2}\left(1-\tilde{R}^2_{A,1}\right)^{-n/2} \\ &\approx \left(\frac{p_1+2k_u}{n-p_1-2k_u}\right)^{p_1/2+k_u}\left(\frac{1-\tilde{R}^2_{A,1}}{\tilde{R}^2_{A,1}}\right)^{p_1/2+k_u}\left(1-\tilde{R}^2_{A,1}\right)^{-n/2}=U_{A,1}(n), \end{aligned}$$

and $L(Y^n|A,M_1)\geq n^{-p_1/2-k}\left(1-\tilde{R}^2_{A,1}\right)^{-n/2}=L_{A,1}(n)$. Similarly,

$$L_{A,2}(n) \leq L(Y^n|A,M_2)=\int_0^\infty(1+g)^{-p_2/2}\left[1-\frac{g}{1+g}\tilde{R}^2_{A,2}\right]^{-n/2}\pi(dg)\leq U_{A,2}(n).$$

Therefore, $\mathrm{BF}^n_{21,A}\leq\frac{U_{A,2}(n)}{L_{A,1}(n)}$

$$=\left(\frac{p_2+2k_u}{n-p_2-2k_u}\right)^{p_2/2+k_u}\left(\frac{1-\tilde{R}^2_{A,2}}{\tilde{R}^2_{A,2}}\right)^{p_2/2+k_u}\left(1-\tilde{R}^2_{A,2}\right)^{-n/2}/\left(n^{-p_1/2-k}(1-\tilde{R}^2_{A,1})^{-n/2}\right) \quad \text{(B.1)}$$

Case (I): For fixed p_j $(j=1,2)$, $\mathrm{BF}^n_{21,A}\leq\zeta(A,n)=n^{\frac{p_1-p_2}{2}+k-k_u}\left(\frac{1-\tilde{R}^2_{A,2}}{1-\tilde{R}^2_{A,1}}\right)^{-n/2}$, ignoring terms independent of n. Then conditional on A, we have directly from our Lemma 1 and the proof of Theorem 3 in Guo and Speckman (2009): for $\mathcal{M}_1\subset\mathcal{M}_2$, and under $\mathcal{M}_1$ and assumptions (A3), (A4), $\zeta(A,n)\xrightarrow{P}0$ as n$\to\infty$, and if $p_2-p_1>2+2(k-k_u)$,

$\zeta(A,n) \xrightarrow{a.s.} 0$ as $n \to \infty$. Further, if $\mathcal{M}_1 \not\subseteq \mathcal{M}_2$, then under $\mathcal{M}_1$ and assumption (A3), $\zeta(A,n) \xrightarrow{a.s.} 0$ as $n \to \infty$.

$$\begin{aligned} \text{Further,} \quad \mathrm{BF}^n_{21,A} &\leq \zeta(A,n) \Leftrightarrow L(Y^n|A,\mathcal{M}_2) \leq \zeta(A,n) L(Y^n|A,\mathcal{M}_1) \\ \Rightarrow L(Y^n|\mathcal{M}_2) &\leq \sum_{A_l \in \mathcal{C}_n} w_l \zeta(A_l,n) L(Y^n|A_l,\mathcal{M}_1) \leq \max_{A \in \mathcal{C}_n} \zeta(A,n) L(Y^n|\mathcal{M}_1) \end{aligned} \tag{B.2}$$

In the limiting sense as $n \to \infty$, the maximum in the upper bound in (B.2) is computed over $A \in \mathcal{C}_\infty$. From the preceding discussion, $\zeta(A,n) \to 0$ under $\mathcal{M}_1$ for all A as $n \to \infty$ implies $\lim_{n\to\infty} \max_{A\in\mathcal{C}_n} \zeta(A,n) \to 0$. Dividing both sides of (B.2) by $\mathrm{L}(Y^n|\mathcal{M}_1)$, this implies $\mathrm{BF}^n_{21} \to 0$ under $\mathcal{M}_1$. Further, the mode of convergence of BF^n_{21} is the same as $\mathrm{BF}^n_{21,A}$, and the rest follows accordingly.

Case (II): For increasing model dimensions $p_1 = O(n^{a_1})$ and $p_2 = O(n^{a_2})$ with $0 \leq a_1 < a_2 < 1$, for $g \sim \pi(g)$ we will only assume *(A3)* so that $k_u = 0$. We have using (B.1)

$$BF^n_{21,A} \leq n^{p_1/2-(1-a_2)p_2/2+k} \left(\frac{1-\tilde{R}^2_{A,2}}{\tilde{R}^2_{A,2}}\right)^{p_2/2} \left(\frac{1-\tilde{R}^2_{A,2}}{1-\tilde{R}^2_{A,1}}\right)^{-n/2}. \tag{B.3}$$

Let us consider the following cases under $0 \leq a_1 < a_2 < 1$.

Case C1: $\mathcal{M}_1 \subset \mathcal{M}_2$. We have $Q_j = \tau(Z_A'Z_A - Z_A'\tilde{H}_{A,j}Z_A) \sim \chi^2_{n-p_j}(0)$, j=1,2, and $Q_1 - Q_2 = \tau\left(Z_A'(\tilde{H}_{A,2} - \tilde{H}_{A,1})Z_A\right) \sim \chi^2_{p_2-p_1}(0)$. Using Lemma 1 of Guo et. al. (2009),

$$\frac{1-\tilde{R}^2_{A,1}}{1-\tilde{R}^2_{A,2}} = \frac{Z_A'Z_A - Z_A'\tilde{H}_{A,1}Z_A}{Z_A'Z_A - Z_A'\tilde{H}_{A,2}Z_A} = \frac{Q_1}{Q_2} = 1 + \frac{(Q_1-Q_2)/(p_2-p_1)}{Q_2/(n-p_2)} \frac{p_2-p_1}{n-p_2} \xrightarrow{a.s.} 1.$$

Moreover $\left(\frac{1-\tilde{R}^2_{A,2}}{\tilde{R}^2_{A,2}}\right) \xrightarrow{a.s.} \left(\frac{\tau^{-1}}{b_{A,1}}\right)$ under $\mathcal{M}_1$, which implies that $\left(\frac{1-\tilde{R}^2_{A,2}}{\tilde{R}^2_{A,2}}\right)^{p_2/2}$ blows up at a rate strictly slower than the rate at which $n^{p_1/2-(1-a_2)p_2/2+k} \to 0$. This implies that $\mathrm{BF}^n_{21,A} \xrightarrow{a.s.} 0$ under $\mathcal{M}_1$.

Case C2: $\mathcal{M}_1 \not\subseteq \mathcal{M}_2$. Using Lemma 1,

$$\frac{1-\tilde{R}^2_{A,2}}{\tilde{R}^2_{A,2}} \overset{a.s.}{\to} \frac{\tau^{-1}+b_{A,1}-b_{A,2}}{b_{A,2}} > 1, \quad \frac{1-\tilde{R}^2_{A,1}}{1-\tilde{R}^2_{A,2}} \overset{a.s.}{\to} \frac{\tau^{-1}}{\tau^{-1}+b_{A,1}-b_{A,2}} < 1, \text{ under } \mathcal{M}_1.$$

For fixed τ^{-1} and $b_{A,2} > 0$ (under $(A2)$), $\left(\frac{1-\tilde{R}^2_{A,2}}{\tilde{R}^2_{A,2}}\right)^{p_2/2}\left(\frac{1-\tilde{R}^2_{A,2}}{1-\tilde{R}^2_{A,1}}\right)^{-n/2} \overset{a.s.}{\to} 0$. In addition, we have $p_1-(1-a_2)p_2+k<0$ for $0 \le a_1 < a_2 < 1$, which implies $\mathrm{BF}^n_{21,A} \overset{a.s.}{\to} 0$ under $\mathcal{M}_1$.

Subsequently using similar arguments as in Case (I), $\mathrm{BF}^n_{21} \overset{a.s.}{\to} 0$ under $\mathcal{M}_1$ for both C1, C2.

Proof of Theorem 7: Given the assumptions *(A1)- (A4)*, Bayes factor consistency holds under the different cases elaborated in Theorem 6. For fixed p, the proof follows trivially using Bayes factor consistency. For increasing $p_n = O(n^a)$ $(a>0)$, our prior is $\left\{\pi^n(\gamma_{j_l}) \propto 2^{-p_j} I[\gamma_{j_l} \in \mathcal{M}_F] + N_j^{-1} I[\gamma_{j_l} \in \mathcal{M}_I]\right\}$, $N_j = \binom{p_n}{p_j}$. Let $\mathrm{BF}^n_{\gamma 1}$= Bayes factor between models γ and $\mathcal{M}_1$, let $D = \{p_j : \gamma \in \mathcal{M}_I\}$ and denote $\mathcal{H}_j = \{\gamma \in \mathcal{M}_I : \dim(\gamma) = p_j\}$. Note that under *(A1)-(A4)*, $\mathrm{BF}^n_{\gamma 1} \overset{P}{\to} 0$ for all $\gamma \in \mathcal{M}_F \bigcup \mathcal{M}_I$, using Theorem 6. Also,

$$\begin{aligned} P(\mathcal{M}_1|Y^n) &= [1 + \sum_{\gamma \in \mathcal{M}_F \cap \mathcal{M}_1^c} 2^{(p_1-p_j)/2} BF^n_{\gamma 1} + 2^{p_1/2} \sum_{\gamma \in \mathcal{M}_I} N_j^{-1} BF^n_{\gamma 1}]^{-1} \\ &= [1 + \sum_{\gamma \in \mathcal{M}_F \cap \mathcal{M}_1^c} 2^{(p_1-p_j)/2} BF^n_{\gamma 1} + 2^{p_1/2} \sum_{p_j \in D} \sum_{\gamma_{jl} \in \mathcal{H}_j} N_j^{-1} BF^n_{\gamma_{jl}1}]^{-1}. \end{aligned}$$

We note that $\sum_{\gamma \in \mathcal{M}_F \cap \mathcal{M}_1^c} 2^{(p_1-p_j)/2} BF^n_{\gamma 1} \to 0$ as $n \to \infty$ since all the individual terms in the finite summation $\to 0$ using Theorem 6. Further, the upper bound of $BF^n_{\gamma_{jl}1}$ for any $\gamma_{jl} \in \mathcal{H}_j$ is given by (using the preceding proof of Theorem 6), (a) for nested case, $\bar{U}^n_{j1} \approx \kappa^{p_j/2} n^{-(1-a_j)p_j/2+p_1/2+k}$ for some $0 < \kappa < \infty$ and large n, (b) for non-nested case, $\bar{U}^n_{j2} \le n^{-(1-a_j)p_j/2+p_1/2+k}$, for large n. Noting that the cardinality of $\mathcal{H}_j \le N_j = \binom{p_n}{p_j}$,

and denoting $\bar{U}_j^n = \max(\bar{U}_{j1}^n, \bar{U}_{j2}^n)$, we have for large n,

$$\sum_{\gamma_{jl} \in \mathcal{H}_j} N_j^{-1} BF_{\gamma_{jl}1}^n \leq \bar{U}_j^n \quad \Rightarrow P(\mathcal{M}_1|Y^n) \geq [1 + 2^{p_1/2} \sum_{p_j \in D} \bar{U}_j^n]^{-1}, \text{ under } \mathcal{M}_1.$$

Noting that p_1 is fixed and the cardinality of $D < \kappa_0 n$ for some constant $\kappa_0 > 0$, it is clear that $2^{p_1/2} \sum_{p_j \in D^n} \bar{U}_j^n \to 0$ as $n \to \infty$, using (a), (b). Hence the result is proved.

Computational steps for MCMC

The posterior computation steps are:

Step 1.1: Update the ν's after marginalizing out the augmented uniform variable using $\pi(\nu_h|-) = Be(1 + n_h, \sum_{j>h} n_j + m)$.

Step 1.2: Update the augmented uniform variables from its full conditional as described in Walker (2007).

Step 2: Update the allocation of atoms to different subjects using $f(y_i|u_i, S_i = h) \propto N(y_i|\alpha_h, x_{\gamma,i}, \beta_\gamma, \tau^{-1}) I(h \in B_w(u_i))$, h=1,. . . ,M

Step 3: Update the precision parameter of the DP using $\pi(m|-) = Ga(a_m + M, b_m - \sum_{l=1}^{M} \log(1 - \nu_l))$, where M is the number of clusters in the particular iteration.

Step 4: Letting $p_\gamma = \sum_{j=1}^{p} \gamma_j$, update precision τ using
$\pi(\tau|-) = Ga\left(a_\tau + \frac{n+p_\gamma}{2}, b_\tau + \frac{1}{2}\left\{(Y^n - X_\gamma\beta_\gamma)'\Sigma_A^{-1}(Y^n - X_\gamma\beta_\gamma) + \frac{1}{g}\beta_\gamma'(X_\gamma'\Sigma_A^{-1}X_\gamma)\beta_\gamma\right\}\right)$.

Step 5: Using the hyper-g prior and the fact that $\frac{g}{1+g} \sim Be(1, 1)$ for $a = 4$, we can subsequently adopt the griddy Gibbs approach (Ritter and Tanner, 1992) to update g.

Step 6: Update the prior inclusion probability $\pi = \Pr(\gamma_j = 1)$ using
$f(\pi|-) = Be(a_1 + p_\gamma, b_1 + p - p_\gamma)$, j=1,. . . ,p.

Step 7: Update γ_j's one at a time by computing their posterior inclusion probabilities after marginalizing out β_γ and conditional on inclusion indicators for the remaining predictors as well as g, τ and A. Denoting $\gamma(j)$ as the vector of variable inclusion

indicators with $\gamma_j = 1$, and $\mathrm{p}_{\gamma(j)}$ as the vector sum of $\gamma(j)$, we can sample γ_j from the Bernoulli conditional posterior distribution with probabilities $\Pr(\gamma_j = 1|-) \propto$

$$\pi(1+g)^{-p_{\gamma(j)}/2} \exp\left\{\frac{\tau}{2}\frac{g}{1+g}\left(Y^{n\prime}\Sigma_A^{-1}X_{\gamma(j)}(X'_{\gamma(j)}\Sigma_A^{-1}X_{\gamma(j)})^{-1}X'_{\gamma(j)}\Sigma_A^{-1}Y^n\right)\right\}.$$

Step 8: Set $\{\beta_j : \gamma_j = 0\} = 0$ and update $\beta_\gamma = \{\beta_j : \gamma_j = 1\}$ using $\pi(\beta_\gamma|-) = N(\beta_\gamma; E, V)$, where $V = \left(\frac{\tau}{g}(X'_\gamma\Sigma_A^{-1}X_\gamma) + \tau(X'_\gamma X_\gamma)\right)^{-1}$ and $E = V\left(\tau X'_\gamma(Y^n - \alpha)\right)$.

B.1 Tables

Table B.1: Estimates and MIPs for fixed effects for Case I when n=100

	MIP_{SLM}	β_{SLM}	MIP_{NLM}	β_{SLM}	β_{L1}	β_{EL}	β_{LMR}	β_{QR}
3	1.00	2.88	1.00	2.83	3.08	3.08	3.15	2.92
2	0.99	1.89	0.98	1.95	2.06	2.06	2.11	1.84
-1	0.93	-0.91	0.75	-0.78	-0.98	-0.98	-0.87	-0.78
0	0.45	-0.01	0.53	0.006	0.01	0.009	-0.003	-0.02
1.5	0.98	1.43	0.90	1.35	1.54	1.54	1.57	1.29
1	0.90	0.79	0.68	0.54	0.74	0.74	0.66	0.42
0	0.43	-0.005	0.53	-0.05	-0.04	-0.04	-0.09	-0.06
-4	1.00	-3.89	1.00	-3.75	-4.05	-4.04	-4.14	-3.95
-1.5	0.99	-1.54	0.92	-1.43	-1.57	-1.57	-1.54	-1.30
0	0.42	0.008	0.54	-0.12	-0.12	-0.12	-0.06	-0.14

Table B.2: Summaries for Case I when n=100

Measure	SLM	NLM	L1	EL	LMR	QR
MSE around β_T	0.07	0.21	0.24	0.24	0.40	0.50
MSE for out of sample prediction	7.70	16.44	8.33	8.32	8.83	9.11

Table B.3: Fixed effects (times 100) for type-II diabetes example

Predictor	$\hat{\beta}_{SLM}$	$\hat{\beta}_{NLM}$	$\hat{\beta}_{L1}$	$\hat{\beta}_{EL}$	$\hat{\beta}_{LMR}$	$\hat{\beta}_{QR}$
TC	0.55(0.11,0.73)	0.74(0.25,1.20)	0.75	0.75	0.29	0.01
SG	2.11(1.75,2.48)	2.82(2.5,3.15)	2.83	2.82	2.99	3.23
HDL	-0.50(-1.4,0.015)	-0.36(-1.61,0)	-1.02	-1.02	-0.42	0
Age	0.34(-0.06,1.3)	0.98(0,2.35)	1.19	1.19	0.57	0.04
Gender	-3.72(-30.12,4.39)	-1.53(-25.46,3.22)	-19.66	-19.81	-7.87	-0.86
BMI(overwt)	1.55(-9.43,24.03)	2.04(-3.33,29.53)	4.33	4.27	15.12	1.84
BMI(obese)	-0.74(-20.33,13.44)	-0.91(-21.93,6.14)	-14.88	-15.03	8.16	0.62
SBP	0.53(0,1.35)	0.03(-0.13,0.65)	0.25	0.25	0.56	0.009
DBP	-0.03(-0.99,0.69)	0(-0.45,0.45)	0.018	0.017	-0.55	0.002
WHR	224.27(67.72,381.88)	3.16(-44.74,91.4)	90.47	91.53	90.79	129.23
PPT	21.42(1.89,57.49)	33.04(0,80.39)	47.31	47.32	37.55	18.99

Table B.4: Marginal Inclusion Probabilities for SLM, NLM and QR

MIP	TC	SG	HDL	Age	Sex	BMI1	BMI2	SBP	DBP	WHR	PPT
SLM	0.97	1	0.64	0.43	0.17	0.15	0.22	0.72	0.23	0.93	0.64
NLM	0.98	1	0.39	0.67	0.12	0.13	0.11	0.14	0.10	0.13	0.68
QR	0.02	1	0.002	0.03	0.08	0.10	0.08	0.01	0.004	0.71	0.42

Table B.5: Prediction (Cov: 95% coverage, CIW: 95% C.I. width)

Replicate	S 1	S 2	S 3	S 4	S 5	S 6	S 7	S 8
MSE_{SLM}	1.25	1.24	1.55	1.21	1.45	1.47	3.44	1.23
MSE_{NLM}	1.23	1.33	1.74	1.29	1.14	1.46	3.43	1.52
MSE_{L1}	1.28	1.45	2.49	2.34	1.13	1.45	3.47	1.75
MSE_{EL}	1.29	1.47	2.51	2.36	1.14	1.45	3.48	1.75
MSE_{LMR}	2.23	1.21	2.15	1.02	1.09	1.36	4.06	1.69
MSE_{QR}	1.82	1.91	2.64	1.15	1.64	2.68	3.98	2.44
Cov_{SLM}	100.00	97.14	100.00	97.14	100.00	100.00	91.42	100.00
Cov_{NLM}	97.12	97.14	94.28	97.14	100.00	97.14	91.42	100.00
CIW_{NLM}	5.92	5.41	5.84	5.94	5.93	5.91	5.59	5.90
CIW_{SLM}	6.93	6.16	6.80	6.81	6.84	6.86	6.13	6.77

Table B.6: Auto-correlations across lags for fixed effects

Predictor	Lag 1		Lag 5		Lag 10		Lag 25		Lag 50	
	SLM	NLM	SLM	NLM	SLM	NLM	SLM	NLM	SLM	NLM
TC	0.22	0.18	0.113	0.194	0.073	0.159	0.032	0.111	0.013	0.059
SG	0.59	0.06	0.386	0.038	0.285	0.022	0.14	0.009	0.06	0.016
HDL	0.19	0.02	0.081	0.012	0.041	0.013	0.01	0.021	0.0005	-0.006
Age	0.21	0.04	0.072	0.009	0.053	-0.0001	0.025	0.006	0.007	-0.014
Gender	0.06	-0.007	0.030	0.0003	0.013	-0.006	0.009	-0.014	0.005	0.019
BMI(overwt)	0.02	-0.002	0.01	-0.006	0.006	0.013	-0.006	0.009	0.0014	0.018
BMI(obese)	0.02	0.002	0.017	0.004	0.004	0.018	0.007	-0.003	0.000	0.000
SBP	0.29	0.0711	0.137	0.019	0.096	0.007	0.047	0.03	0.014	0.022
DBP	0.07	0.0239	0.021	0.019	0.019	0.031	0.009	-0.003	0.004	-0.012
WHR	0.44	0.0642	0.353	0.043	0.321	0.061	0.251	0.06	0.186	-0.003
PPT	0.22	0.0600	0.118	0.047	0.068	0.045	0.015	0.004	-0.002	0.019

B.2 Figures

Figure B.1: MIP for Case I: Solid lines - SLM, dashed lines - NLM

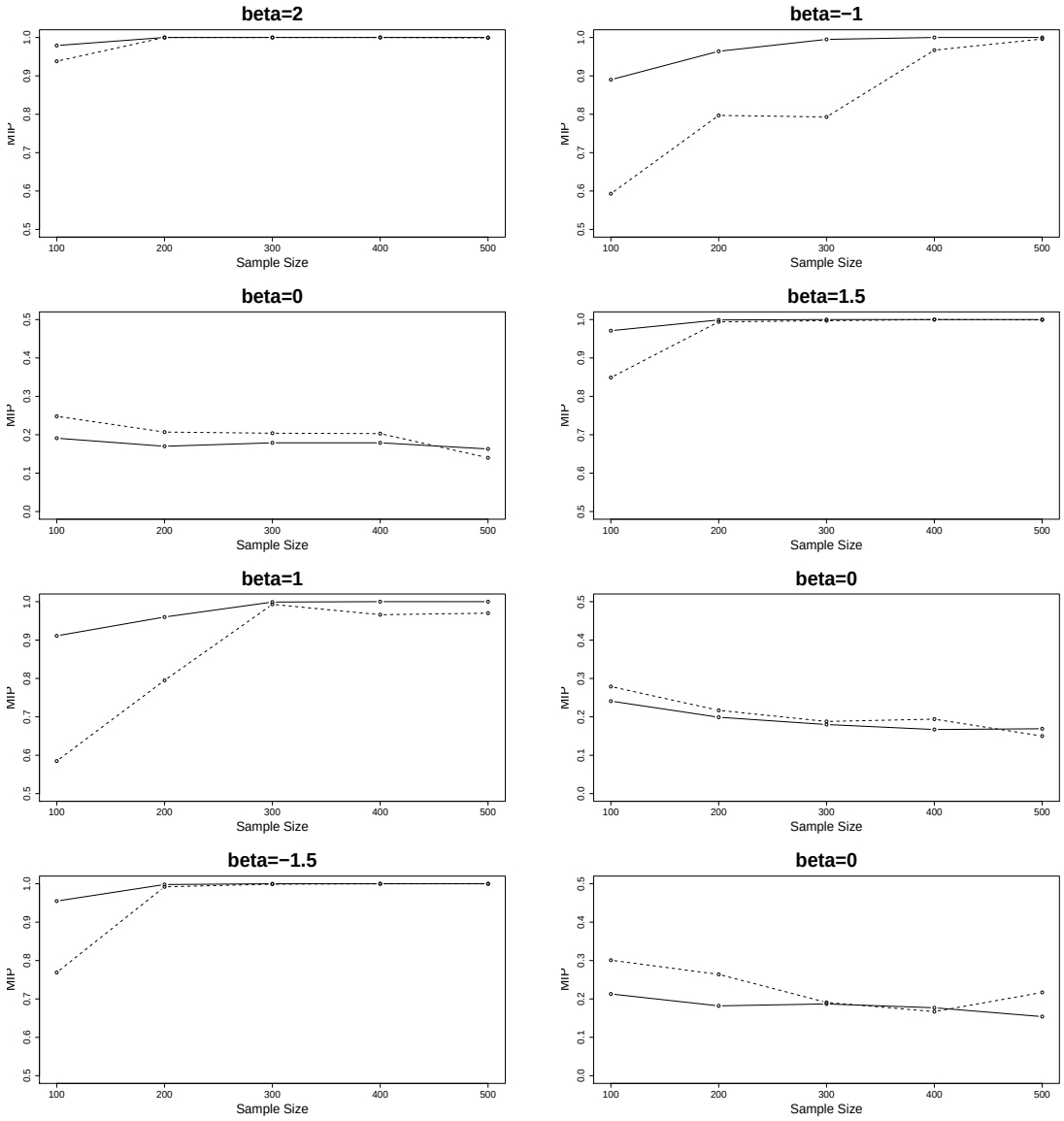


Figure B.2: MIP for Case II: Solid lines - SLM, dashed lines - NLM

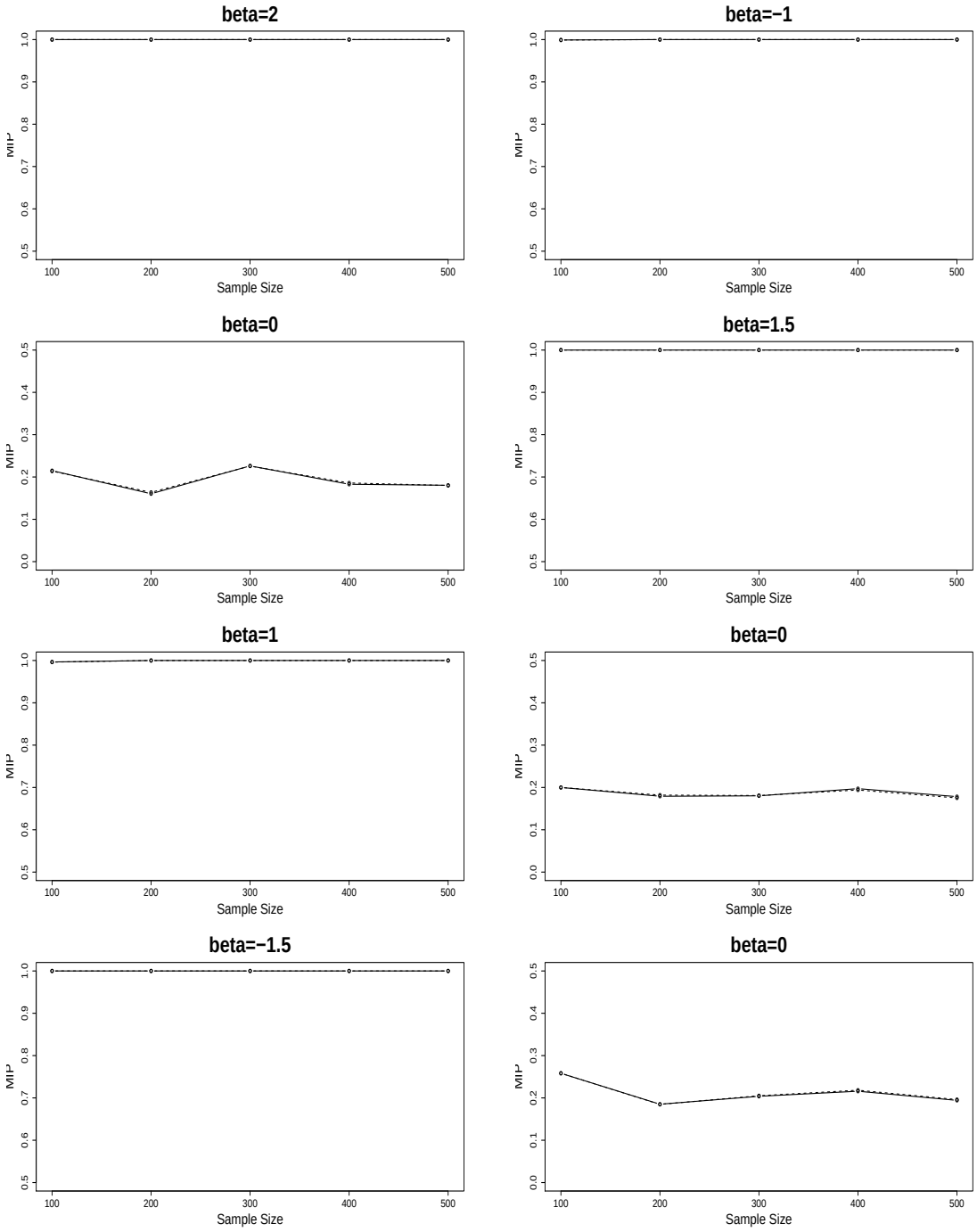
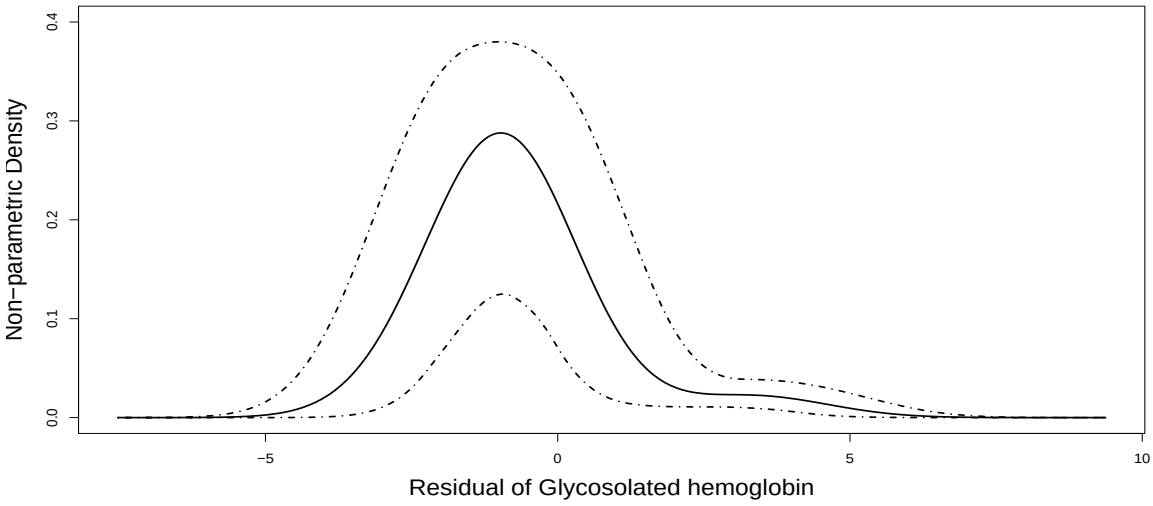


Figure B.3: Residual plots for Diabetes study for Semi-parametric Linear Model



Appendix C

Chapter 4

Proof of Theorem 8: Suppose $\Pi \sim GP(0, K)$, where $K(\cdot, \cdot)$ is the covariance kernel with parameters ϕ_1, ϕ_2. Then using the standard result for a multivariate normal distribution, we have $E[f(z)|f^j(x^*_{m-1}), f^j(x^*_m)] = \bar{\Sigma}_{m,m+1}\Big(f^j(x^*_m), \quad f^j(x^*_{m+1})\Big)'$, where

$$\bar{\Sigma}_{m,m+1} = \Big[\phi_1 e^{-\phi_2(z-x_m)^2}, \quad \phi_1 e^{-\phi_2(z-x_{m+1})^2}\Big] \begin{pmatrix} \phi_1 & \phi_1 e^{-\phi_2\Delta_m^2} \\ \phi_1 e^{-\phi_2\Delta_m^2} & \phi_1 \end{pmatrix}^{-1}$$

$$= \frac{1}{1-e^{-2\phi_2\Delta_m^2}} \Big[e^{-\phi_2(z-x_m)^2}, e^{-\phi_2(z-x_{m+1})^2}\Big] \begin{pmatrix} 1 & -e^{-\phi_2\Delta_m^2} \\ -e^{-\phi_2\Delta_m^2} & 1 \end{pmatrix}$$

$$= \frac{1}{1-e^{-2\phi_2\Delta_m^2}} \left(e^{-\phi_2(z-x_m)^2} - e^{-\phi_2(z-x_{m+1})^2-\phi_2\Delta_m^2}, \quad e^{-\phi_2(z-x_{m+1})^2} - e^{-\phi_2(z-x_m^2)-\phi_2\Delta_m^2}\right)$$

.

We can use Taylor's series expansion and the assumption $\Delta_m^b \approx 0$, $b \geq 4$ to obtain $\bar{\Sigma}_{m,m+1}$

$$\begin{aligned} &\approx \frac{1}{2\phi_2\Delta_m^2}\Big(\phi_2(z-x_{m+1})^2 + \phi_2\Delta_m^2 - \phi_2(z-x_m)^2, \quad \phi_2(z-x_m)^2 + \phi_2\Delta_m^2 - \phi_2(z-x_{m+1})^2\Big) \\ &= \frac{1}{2\phi_2\Delta_m^2}\Big(2\phi_2\Delta_m(x_{m+1}-z), \quad -2\phi_2\Delta_m(x_m-z)\Big). \end{aligned}$$

Then for $z \in [x^*_m, x^*_{m+1})$,

$$\begin{aligned} w^j(z) &\approx \frac{1}{\Delta_m^2}\Big(\Delta_m(x_{m+1}-z), \quad -\Delta_m(x_m-z)\Big)\Big(f^j(x^*_m), \quad f^j(x^*_{m+1})\Big)' \\ &= \frac{x^*_m - z}{\Delta_m} f^j(x^*_{m-1}) + \frac{z - x^*_{m-1}}{\Delta_m} f^j(x^*_m). \end{aligned}$$

Khachiyan's algorithm for computing minimum volume covering ellipsoids

1. **Input:** A set of points $P = \{a_1, \ldots, a_n\} \subset \Re^p$, and $\epsilon > 0$.

Further let $q_i \leftarrow ((a_i)^T, 1)^T$, i=1,. . . ,n, and $\Lambda(u) = \sum_{h=1}^{n} a_h q_h (q_h)^T$.

2. **Initialization:** $k \leftarrow 0$, $N \leftarrow p + 1$, $u^0 \leftarrow (1/n)1_n$.

Begin loop:

3. $j \leftarrow \arg\max_{i=1,\ldots,n} (q_i)^T \Lambda(u^k)^{-1}(q_i), \kappa \leftarrow (q_j)^T \Lambda(u^k)^{-1}(q_j)$.

4. $\beta \leftarrow \frac{\kappa - p - 1}{(p+1)(\kappa - 1)}$.

5. $u^{k+1} \leftarrow (1 - \beta)u^k + \beta e^j$, $k \leftarrow k + 1$, where e^j is a vector of zeros with jth element =1.

6. If $\text{norm}(u^{k+1} - u^k) > \epsilon$, go to step 2.

7. $u^* = u^{k+1}$.

End loop

8. **Output:** Let $U = \text{diag}(u^*)$. Then $c = Pu^*$ and $A = (1/p)(PUP^T - cc^T)^{-1}$.

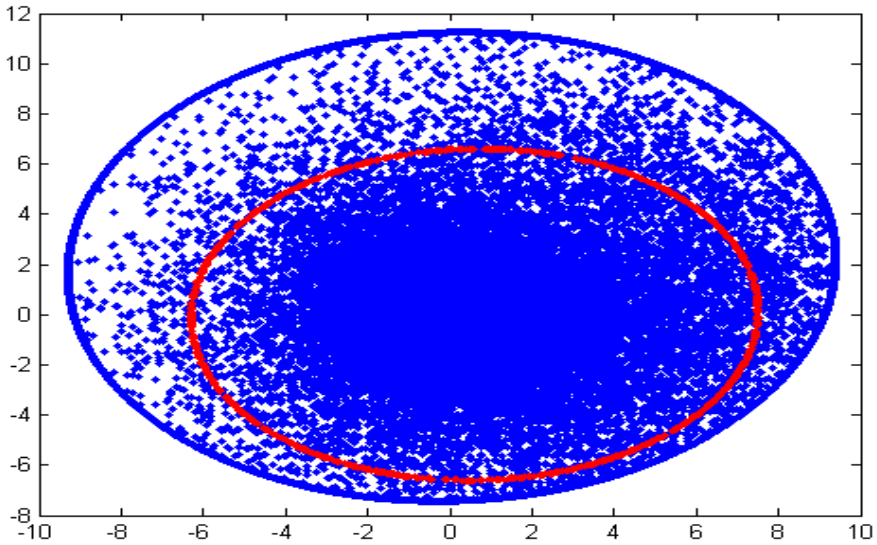
C.1 Tables

Table C.1: Frequentist Coverage (Fcov) of Credible Regions

Method	Fcov at knots	Fcov under Lip	Fcov under Lin
MVCE (knots = 30)	97	97	34
CA (knots = 30)	87	87	12
Pnorm(knots = 30)	84	83	24
MVCE (knots = 50)	97	96	74
CA (knots = 50)	84	84	41
Pnorm(knots = 50)	83	82	52
MVCE (knots = 70)	98	98	69
CA (knots = 70)	87	87	44
Pnorm(knots = 70)	82	82	47
MVCE (knots = 200)	97	97	97
CA (knots = 200)	94	94	94
Pnorm(knots = 200)	93	93	93
Method	Area (Lip)	Area (Lin)	Log(vol) at knots
MVCE (knots = 30)	3302.4	169.3	59.87
CA (knots=30)	3262.5	129.6	62.92
Pnorm(knots=30)	3281.5	148.5	51.76
MVCE (knots = 50)	1131.8	170.5	96.50
CA (knots=50)	1097.8	136.7	107.97
Pnorm(knots=50)	1112.5	151.3	83.45
MVCE (knots = 70)	1121.1	171.7	123.81
CA (knots=70)	1092.4	143.1	153.04
Pnorm(knots=70)	1104.8	155.4	105.20

C.2 Figures

Figure C.1: Comparison of two dimensional credible regions. Blue Dots: 95% HPD set generated from mixture of bivariate Gaussian and t distribution; Blue line: MVCE credible region (posterior prob = 0.9507); Red Line: Credible region using asymptotic normality (posterior prob = 0.895)



Bibliography

Bartlett, M. (1957), "A comment on D. V. Lindley's statistical paradox", Biometrika, 44, 533-534.

Berger, J. O. and Pericchi, L. R. (1996), "The intrinsic Bayes factor for model selection and prediction", J. Amer. Statist. Assoc., 91, 109 - 122.

Berger, J. O. and Pericchi, L. (2001), "Objective Bayesian methods for model selection: Introduction and comparison", Model Selection, vol. 38 of IMS Lecture Notes - Monograph Series, (ed. P. Lahiri), 135 - 193. Institute of Mathematical Statistics.

Brown, E.R., and Ibrahim, J.G. (2003), "A Bayesian semiparametric joint hierarchical model for longitudinal and survival data", Biometrics, 59, 221 - 228.

Brunham, L.R., Kruit, J.K., Pape, T.D., Timmins, J.M., Reuwer, A.Q., Vasanji, Z., Marsh, B.J., Rodrigues, B., Johnson, J.D., Parks, J.S., Verchere, C.B., and Hayden, M.R. (2007), "β-cell ABCA1 influences insulin secretion, glucose homeostasis and response to thiazolidinedione treatment", Nature Medicine, 13, 340 - 347.

Bush, C.A., and MacEachern, S.N. (1996), "A semiparametric Bayesian model for randomised block designs", Biometrika, 83, 275 - 285.

Carvalho, C.M., Chang, J., Lucas, J.E., Nevins, J.R., Wang, Q., and West, M. (2008), "High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics", Journal of the American Statistical Association, 103, 1438 - 1456.

Casella, G. and Moreno E. (2006), "Objective Bayesian variable selection", J. Amer. Statist. Assoc., 101, 157 - 167.

Casella, G., Gir$\acute{o}$n, F. J., Mart$\acute{i}$nez, M. L. and Moreno, E. (2009), "Consistency of Bayesian procedures for variable selection", Ann. Statist. 37 1207 - 1228.

Castillo, I. (2012). A semi-parametric Bernstein-von Mises theorem for Gaussian process priors. Probability Theory and Related Fields, 152, 53-99.

Chen, M., Silva, J., Paisley, J., Wang, C., Dunson, D.B., and Carin, L. (2010), "Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: algorithm and performance bounds", IEEE. Transactions on Signal Processing, 58, 6140 - 6155.

Chen, M.H. and Shao, Q.M. (1999). Monte Carlo estimation of Bayesian credible and HPD intervals. Journal of Computational and Graphical Statistics, 8, 69-92.

Chipman, H., George, E., and McCulloch, R. (2010), "BART: Bayesian Additive Regression Trees", The Annals of Applied Statistics , 4, 266 - 298.

Choi, T. (2007) "Alternative posterior consistency results in nonparametric binary regression using Gaussian process priors.", Journal of Statistical Planning and Inference, volume 137, Issue 9, 2975-2983

Choi, T., and Schervish, M. (2004), "Posterior Consistency in Nonparametric Regression Problems under Gaussian Process Priors", Technical Report.

Choi, T. and Schervish, M. J. (2007) "On posterior consistency in nonparametric regression problems", Journal of Multivariate Analysis, volume 98, Issue 10, 1969-1987

Chung, Y., and Dunson, D.B. (2009), "Nonparametric Bayes Conditional Distribution Modeling With Variable Selection", J. Amer. Statist. Assoc., 104(488), 1646-1660.

Cooper DA. Learning Lipschitz Functions. International Journal of Computer Mathematics 1995;59:1526

Croux, C, Haesbroeck, G, and Rousseeuw, P.J. (2002), Location adjustment for minimum volume ellipsoid estimator. Statistics and Computing, 12, 3, 191200.

De Iorio, M., Muller, P., Rosner, G.L., and MacEachern, S. (2004), "An ANOVA Model for Dependent Random Measures", Journal of the American Statistical Association, 99, 205-215.

Denison, D.G.T., Mallick, B.K., and Smith, A.F.M. (1998). Automatic Bayesian Curve Fitting. Journal of Royal Statistical Society, Series B, 60, 330-350.

Dunson, D.B. (2006), "Bayesian dynamic modeling of latent trait distributions", Biostatistics, 7, 551 - 568.

Dunson, D.B., and Herring, A.H. (2005), "Bayesian model selection and averaging in additive and proportional hazards models", Lifetime Data Analysis, 11(2), 213-232.

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

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

Epstein, M., and Sowers, J.R. (1992), "Diabetes mellitus and hypertension", Hypertension, 19, 403-418.

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

Ferguson, T. S. (1973), "A Bayesian analysis of some nonparametric problems", Annals of Statistics, 1, 209 - 230

Ferguson, T. S. (1974), "Prior distributions on spaces of probability measures", Annals of Statistics, 2, 615 - 629.

Fernándex, C., E. Ley, and M. F. J. Steel (2001), ''Benchmark priors for Bayesian model averaging", J. Econometrics, 100(2), 381 - 427.

Fokoue, E., and Titterington, D.M. (2003), "Mixtures of factor analysers: Bayesian estimation and inference by stochastic simulation", Machine Learning, 50, 73 - 94.

Fokoue, E. (2005), "Mixtures of factor analyzers: an extension with covariates", Journal of Multivariate Analysis, 95, 370 - 384.

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1999), "Posterior consistency of Dirichlet mixtures in density estimation", Annals of Statistics, 27, 143 - 158.

Ghosal, S., and Roy, S. (2006), "Posterior consistency of Gaussian process prior for nonparametric binary regression", Annals of Statistics, 34, 2413 - 2429.

George, E. I. and McCulloch, R. E. (1993), "Variable Selection Via Gibbs Sampling", J. Amer. Statist. Assoc., 88(423), 881-89.

George, E. I. and McCulloch, R. E. (1997), "Approaches for Bayesian Variable Selection", Statist. Sinica, 7(2), 339-74.

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1999), "Posterior consistency of Dirichlet mixtures in density estimation", Annals of Statistics, 27, 143 - 158.

Ghosal, S., Lember, J., and van der Vaart, A. (2008), "Nonparametric Bayesian model selection and averaging", Electronic J. Stat., 2, 63-89.

Ghosal, S., and Roy, S. (2006), "Posterior consistency of Gaussian process prior for nonparametric binary regression", Annals of Statistics, 34, 2413 - 2429.

Green, P.J. (1995), "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination", Biometrika, 82 (4), 711-732.

Gramacy R.B., and Lee, H. K. H, (2008), "Bayesian Treed Gaussian Process Models With an Application to Computer Modeling", Journal of the American Statistical Association, 103, 1119 - 1130.

Griffin, J. E. and Steel, M. F. J. (2006), "Order-based dependent Dirichlet processes", Journal of the American Statistical Association, 101, 179 - 194.

Guo, R. and Speckman, P. (2009), "Bayes factor consistency in linear models", In the 2009 International Workshop on Objective Bayes Methodology, Philadelphia, 2009.

Hanson, T. (2006), "Inference for Mixtures of Finite Polya Trees", Journal of the American Statistical Association, 101, 1548 - 1565.

Jara, A., and Hanson, T. (2010), "A class of mixtures of dependent tail-free processes", Biometrika, accepted.

Jeffreys, H. (1961), "Theory of Probability", Oxford Univ. Press.

Jiang, W. (2007), " Bayesian Variable Selection for high dimensional generalized linear models: convergence rates of the fitted densities", Ann. of Statist., 35(4), 1487 - 1511.

Johnson, V., and Rossell, D. (2010), "On the use of non-local prior densities in Bayesian hypothesis tests", J. Royal. Statist. Soc., Series B., 72(2), 143 - 170.

Kass, R. E.,and Raftery, A.E.(1995), "Bayes Factors", J. Amer. Statist. Assoc., 90, 773 - 795.

Khachiyan, L. G. (1996), "Rounding of polytopes in the real number model of computation", Mathematics of Operations Research 21, 307-320.

Kim, S., Tadesse, M.G., and Vannucci, M. (2006), "Variable selection in clustering via Dirichlet process mixture models", Biometrika, 93(4), 877-893.

Klee, V. (1966), "Convex polytopes and linear programming", In Proceedings of the IBM Scientific Computing Symposium: Combinatorial Problems. IBM, Armonk, N.Y., 123158.

Knapik, B.T., van der Vaart, A.W. and van Zanten, J.H. (2011), "Bayesian inverse problems", arXiv:1103.2692v1

Knorr, E.M., Ng, R.T. and Zamar, R.H., "Robust Space Transformations for Distance-based Operations", In Proceedings of the Seventh ACM SIGKDD international conference on Knowledge discovery and data mining, 2001.

Kleinman, K.P., and Ibrahim, J.G. (1998), "A semiparametric Bayesian approach to the random effects model", Biometrics, 54, 921 - 938.

Kuo, L. and Mallick, B. (1997), "Semiparametric inference for the accelerated failure time model", Can. J. Stat., 25, 457-472.

Kyung, M., Gill, J., and Casella, G. (2009), "Characterizing the variance improvement in linear Dirichlet random effects models", Statistics and Probability Letters, 79, 2343-2350.

Lavine, M. (1992), "Some Aspects of Polya Tree Distributions for Statistical Modelling", Annals of Statistics, 20, 1222 - 1235.

Lavine, M. (1994), "More Aspects of Polya Tree Distributions for Statistical Modelling", Annals of Statistics, 22, 1161 - 1176.

Lawrence, N. (2005), "Probabilistic Non-linear Principal Component Analysis with Gaussian Process Latent Variable Models", Journal of Machine Learning Research, 6, 1783 - 1816.

Lee, S.Y., Lu, B., and Song, X.Y. (2008), "Semiparametric Bayesian analysis of structural equation models with fixed covariates", Statistics in Medicine, 27, 2341 - 2360.

Lenk, P. J. (1988), "The Logistic Normal Distribution for Bayesian, Nonparametric, Predictive Densities", Journal of the American Statistical Association, 83, 509 - 516.

Lenk, P. J. (1991), "Towards a Practicable Bayesian Nonparametric Density Estimator", Biometrika, 78, 531 - 543.

Liang, F., Paulo, R., Molina, G., Clyde, M.A., and Berger, J.O. (2008), "Mixtures of g-priors for Bayesian Variable Selection.", J. Amer. Statist. Assoc., 103(481), 410-423.

Lindstrom, M.J. (1999), "Penalized estimation of free knot splines", Journal of Computational and Graphical Statistics, 8, 333-352.

Linkletter, C., Bingham, D., Hengartner, N., Higdon, D. and Ye K. Q. (2006). Variable Selection for Gaussian Process Model in Computer Experiments. Technometrics, 48, 478-490.

Longnecker, M. P., Klebanoff, M. A., Zhou, H. B., and Brock, J. W. (2001), "Association between maternal serum concentration of the DDT metabolite DDE and preterm and small-for-gestational-age babies at birth", Lancet, 358, 110 - 4.

MacEachern, S. N. (1999), "Dependent nonparametric processes", In ASA Proceedings of the Section on Bayesian Statistical Science, Alexandria, VA: American Statistical Association, 50-55.

Marin, J.M. and Robert, C. P. (2007), "Bayesian Core: A Practical Approach to Computational Bayesian Statistics", Springer-Verlag Inc.

Marron, J. S., and Wand, M. P. (1992), "Exact mean integrated squared error", Annals of Statistics, 20, 712 - 736.

Mauldin, R.D., Sudderth, W.D., and Williams, S.C. (1992), "Polya Trees and Random Distributions", Annals of Statistics, 20, 1203 - 1221.

Meyer, M. C. and Laud, P. W. (2002), "Predictive variable selection in generalized linear models", J. Amer. Statist. Assoc., 97, 859 - 871.

Mokdad, A.H., Bowman, B.A., Ford, E.S., Vinicor, F., Marks, J.S., Koplan, J.P. (2001), "The continuing epidemics of obesity and diabetes in the United States", J. Amer. Med. Assoc., 286, 1195-1200.

Moreno, E., Bertolino, F. and Racugno, W. (1998), "An intrinsic limiting procedure for model selection and hypothesis testing", J. Amer. Statist. Assoc., 93, 1451 - 1460.

Moreno, E., Girón, F.J., and Casella, G. (2010), "Consistency of objective Bayes factors as the model dimension grows", Ann. Statist., 38(4), 1937 - 1952.

Mostofi, A.G., and Behboodian, J. (2007), "On model selection in Bayesian regression", Metrika, 66(3), 259-268.

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

Neal, R.M. (2000), "Markov Chain sampling methods for Dirichlet process mixture models", Journal of Computational and Graphical Statistics, 9, 249 - 265.

O'Hagan, A., and Kingman, J. F. C. (1978), "Curve Fitting and Optimal Design for Prediction", Journal of the Royal Statistical Society B, 40, 1 - 42.

O'Hara, R.B. and Sillanpää, M.J. (2009), "Review of Bayesian variable selection methods: What, how and which", Bayesian Analysis, 4, 85 - 118.

Park, T., and Casella, G. (2008), "The Bayesian Lasso", J. Amer. Statist. Assoc., 103(482), 681-686.

Pati, D., Dunson, D.B. and Tokdar, S. (2011), "Posterior consistency in conditional distribution estimation", Annals of Statistics, revision submitted.

Polonik, W. (1995), "Measuring mass concentrations and estimating density contour clusters - an excess mass approach", The Annals of Statistics, 23, 855881.

Polonik, W. (1997), "Minimum volume sets and generalized quantile processes", Stochastic Processes and Their Applications, 69:124.

Raftery, A. E. and Richardson, S. (1993), "Model selection for generalized linear models via GLIB, with application to epidemiology", Bayesian Biostatistics, Berry, D. A. and Stangl, D. K., editors. Marcel Dekker, New York.

Ritter, C., and Tanner, M.A. (1992), "Facilitating the Gibbs sampler: the Gibbs stopper and the griddy-Gibbs sampler", J. Amer. Statist. Assoc., 87(419), 861-868.

Rivoirard, V. and Rousseau, J. (2009), "Bernstein-von Mises Theorem for linear functionals of the density", arXiv:0908.4167v1

Savitsky, T., Vannucci, M. and Sha N. (2011). Variable Selection for Nonparametric Gaussian Process Priors: Models and Computational Strategies. Statistical Science, 26, 130-149.

Schmidt, M.I., Duncan, B.B., Canani L.H., Karohl, C., and Chambless L. (1992), "Association of waist-hip ratio with diabetes mellitus. Strength and possible modifiers", Diabetes Care. 15(7), 912-4.

Schwartz, L. (1965), "On Bayes procedures", Z. Wahrsch. Verw. Gebiete, 4, 10 - 26.

Sethuraman, J. (1994), "A constructive definition of Dirichlet priors", Statistica Sinica, 4, 639-50.

Shao, J. (1997), "An asymptotic theory for linear model selection", Statist. Sinica, 7, 221 - 264.

Sheather, S. J., and Jones M. C. (1991), "A reliable data-based bandwidth selection method for kernel density estimation", Journal of the Royal Statistical Society B, 53, 683 - 690.

Silva, R., and Gramacy, R. (2010), "Gaussian process structural equation models with latent variables", Proceedings of the 26th Conference on Uncertainty on Artificial Intelligence, UAI.

Smith, M., and Kohn, R. (1996), "Nonparametric regression using Bayesian variable selection", J. Econometrics, 75(2), 317-343.

Stander, B.T, Hart, J.C. (1995), "A Lipschitz method for accelerated volume rendering", Proceedings of 1994 Symposium on Volume Visualization. IEEE, New York, NY, USA, 10714.

Strawderman, W. E. (1971), "Proper Bayes minimax estimators of the multivariate normal mean", Ann. Math. Statist., 42, 385-388.

Strongin, R.G. (1973), "On the convergence of an algorithm for finding a global extremum", Engineering Cybernetics 11,549-555.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso", J. Royal. Statist. Soc., Series B., 58(1), 267-288.

Tokdar, S. T., and Ghosh, J. K. (2007), "Posterior consistency of logistic Gaussian process priors in density estimation", Journal of Statistical Planning and Inference, 137, 34 - 42.

Tokdar, S. T., Zhu, Y. M., and Ghosh, J. K. (2010), "Bayesian Density Regression with Logistic Gaussian Process and Subspace Projection Bayesian Analysis", Bayesian Analysis, 5, 319 - 344.

Tokdar, S. T. (2006), "Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression", Sankhya: The Indian Journal of Statistics, 67 90-110.

Van der Vaart, A. W., and Wellner, J. A. (1996), "Weak Convergence and Empirical Processes", Springer-Verlag, New York.

Walker, S. (2007), "Sampling the dirichlet mixture model with slices", Comm. Statist. Sim. Comput., 36, 45 - 54.

Wood, G. R., and Zhang, B. P. (1996), "Estimation of the Lipschitz Constant of a Function", Journal of Global Optimization, Vol. 8, 1, 91-103.

Yang, M., and Dunson, D.B. (2010), "Bayesian semiparametric structural equation models with latent variables", Psychometrika, 75, 675-693.

Yau C., Papaspiliopoulos, O., Roberts, G. and Holmes, C. (2011), ''Bayesian non-parametric hidden Markov models with applications in genomics", J. Royal Stat. Soc., Series B, 73(Part 1), 33 - 57.

Yi, N., and S. Xu. (2008), "Bayesian LASSO for quantitative trait loci mapping", Genetics, 179, 1045-1055.

Zellner, A. and Siow, A. (1980), "Posterior odds ratios for selected regression hypotheses", Bayesian Statistics: Proceedings of the First International Meeting, Valencia, J.M. Bernardo, M.H. Degroot, D.V. Lindley and A.F.M. Smith. Valencia, Spain: University of Valencia Press, 585-603.

Zellner, A. (1986), "On assessing prior distributions and Bayesian regression analysis with g-prior distributions", Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti, (eds. P. K. Goel and A. Zellner), 233-243. North-Holland/Elsevier.

Zou, H., and Hastie, T. (2005), "Regularization and variable selection via the elastic net", J. Royal. Statist. Soc, Series B., 67(2), 301 - 320.

Zou, F., Huang, H., Lee, S. and Hoeschele, I. (2010). Nonparametric Bayesian Variable Selection with Applications to Multiple Quantitative Trait Loci Mapping with Epistasis and Gene-Environment Interaction. Genetics, 186, 385-394. Barber, C.B., Dobkin, D.P., and Huhdanpaa, H. (1996). "The quickhull algorithm for convex hulls", ACM Transactions on Mathematical Software (TOMS), 22, 4, 469-483.