ABSTRACT

WEI, RAN. Bayesian Variable Selection Using Continuous Shrinkage Priors for Nonparametric Models and Non-Gaussian Data. (Under the direction of Subhashis Ghoshal and Brian Reich.)

In this thesis, we study the properties and applications of Bayesian variable selection in regression models. We focus on Bayesian shrinkage priors that are global-local mixture priors to select an appropriate subset of covariates. First, a Bayesian non-parametric regression model is proposed with multivariate continuous shrinkage prior. This approach is to decompose the commonly-used linear regression model into the summation of nonlinear main effects and two-way interaction terms and apply the proposed computationally-advantageous continuous shrinkage prior to identify important effects. We construct a multivariate Dirichlet-Laplace prior that aggressively shrinks many of the terms towards zero, thus mitigating the noise of including unimportant exposures and allowing us to isolate the effects of important exposures. Theoretical studies demonstrate the asymptotic prediction and variable selection consistency properties, while numerical simulations present model performance of prediction and variable selection under practical scenarios. The method is applied on neurobehavioral data from Agricultural Health Study that investigates the associations between pesticide use and neurobehavioral outcomes in farmers. The proposed method shows improved accuracy in predicting the joint effects on neurobehavioral responses, while restricting the number of covariates included in the model through variable selection technique.

Next, we investigate the contraction properties of shrinkage priors in logistic regression model when the number of covariates is high. For the prior distribution that is heavy-tailed and concentrated around around zero with large probability, the logistic

coefficient estimates are asymptotically concentrated around the true sparse vector in $\mathcal{L}_2$-error contraction rate. It is shown that the proposed contraction rate is comparable with point mass prior that is studied in Atchadé (2017). The simulation study under logistic regression model verifies the theoretical results by showing that shrinkage priors such as horseshoe prior and Dirichlet-Laplace prior perform quite similar as the point mass prior in estimation, variable selection and prediction, but yield much better results than Bayesian lasso and non-informative normal prior.

Last, we propose spatial extension of Bayesian shrinkage prior in the application of forecasting calibration for fine particulate matter from wildland fire smoke. To improve the forecast, we consider expanding the space-time forecasts by computing spatial summaries of the forecast in surrounding areas and using the constructed covariates in a multivariate regression model. We apply an additive nonparametric regression model to delineate the associations between the fine particulate measurements at monitoring stations and multivariate spatial summaries. Our proposed model incorporates both spatial variation on the additive nonparametric effects and Bayesian variable selection across locations. The simulation study evaluates the additive regression model under different model assumptions (nonlinearity, spatial dependence in the mean functions and residuals) and demonstrates its advantages in both prediction and variable selection. The regression model is implemented to downscale the fine particulate matter forecasts of wildland fire smoke in Washington state in 2015 and find an improvement of 18% in mean squared error compared to the standard linear regression calibration.

Bayesian Variable Selection Using Continuous Shrinkage Priors for Nonparametric Models and Non-Gaussian Data

by
Ran Wei

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2017

APPROVED BY:

Jane Hoppin	Soumendra Lahiri
Subhashis Ghoshal Co-chair of Advisory Committee	Brian Reich Co-chair of Advisory Committee

DEDICATION

To my family,

To all of my friends,

And most importantly, to my husband,

Without whom none of this would be possible

BIOGRAPHY

The author was born in 1990 in Chongqing, China. In 2008, she was admitted to Zhejiang University that is one of the top-tier universities in China. With a first major in English Literature, she found great interests in applied Statistics so that a second major of Statistics is added in her Junior year. After four years of undergraduate study in Zhejiang University, she decided to pursue a Ph.D. degree in Statistics at North Carolina State University. During the five years of Ph.D. program, she is conducting statistical research in Bayesian methodology and environmental statistics.

ACKNOWLEDGEMENTS

I would like to thank all the people who contributed in some ways during the completion of my thesis and my Ph.D. degree.

First I want to express my appreciation to my co-advisors Dr. Subhashis Ghoshal and Dr. Brian Reich. It has been a pleasure to work with both of them. Dr. Ghoshal is always an excellent example as a responsible and humble scholar. I really appreciate all his dedication of time, ideas and mentoring to make my pursuit of Ph.D. degree always productive and insightful. I am also grateful for the opportunity working with Dr. Reich on the applications and development of statistics methods. Dr. Reich is always fun to work with. His humor, passion and admiration towards academic research have inspired me in different perspectives throughout my entire Ph.D. experience. I would like to express my deepest and most genuine gratitude to their mentorship.

I am also thankful for having Dr. Jane Hoppin and Dr. Soumendra Lahiri as my committee members. I met Dr. Jane Hoppin in an interdisciplinary research project. She showed me a wonderful example of a successful and passionate woman scientist. She has contributed tremendous work in my research project and provided her thoughtful advice outside of the statistical discipline. I would also like to extend my appreciation to Dr. Lahiri for his support and helpful comments on my research projects. He has always been very generous spending his time during the entire process.

When working with statisticians and scientists at Environmental Protection Agency (EPA), I truly appreciate the opportunity of interdisciplinary research for environmental problems. I am extremely thankful for Dr. Ana Rappold and Dr. Susan O'neill and their team members. They provided a friendly and cooperative atmosphere every time we discuss about the research problem. They also provided useful feedback and insightful

comments on my work. The experience of working with them has inspired me to apply my technique skills in statistics to solve real world problems in public health.

I would like to acknowledge the Department of Statistics at North Carolina State University. I am grateful for every faculty member and staff in the department. Their services and dedication have made the department a big family for all the students. My graduate experience benefitted greatly from every course I took, the opportunities of academic research and industrial training, the resources of cutting-edge research in Statistics and the fundings we received to help us continuing our graduate study without financial burden.

Most importantly, I want to thank my all my friends and family members for their unconditional love and supports, especially my father and mother. My parents are wonderful role model to me in my life. I am always thankful for growing up in a family filled with laughters and love. They have provided me great freedom since my childhood to encourage me exploring my true interests and discovering my potential. My greatest appreciation goes to the most important person in my life, my husband, Milton Mai. He is such a wonderful man that I could not have asked for a better one in my life. I am grateful for every amazing and blessed day we are together with each other. I just want to acknowledge him for everything he has done!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Bayesian Shrinkage Prior

For statistical models such as linear regression, logistic regression or normal mean models, high-dimensional data analysis is challenging due to its computational burden and the inherent limitations due to the high complexity of the model. Despite the difficulty of high-dimensional data, approximate solutions can be achieved with the assumption of sparsity that restricts the number of important covariates to be small. A variety of regularization methods such as LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001) and adaptive LASSO (Zou, 2006) have been proposed to identify sparse solutions for high-dimensional analysis.

Compared to regularization methods, Bayesian variable selection procedures naturally address model selection uncertainty. Since the value zero of a regression coefficient is equivalent to having the corresponding predictor removed from the model, priors are designed to give special emphasis to the value zero. For example, a point-mass prior combines a probability mass at zero and a non-zero continuous distribution. Computation is

generally carried out by Reversible Jump Markov Chain Monte Carlo methods (Green, 1995). A spike-and-slab prior (Ishwaran and Rao, 2005) is a mixture of two normal distributions with one highly concentrated at around zero. The Stochastic Search Variable Selection (SSVS) method (George and McCulloch, 1993) can be used to compute the posterior distribution corresponding to a spike-and-slab prior. Due to high computational costs, these methods are not scalable to very high dimensional situations commonly arising in recent applications. For high-dimensional model with sparsity assumption, the Bayesian LASSO (Park and Casella, 2008) uses a double exponential prior distribution on the coefficients. The posterior mode of the Bayesian LASSO is the LASSO estimator in linear regression model, and the posterior can be computed by a simple Gibbs sampling procedure. However, the Bayesian LASSO does not make the posterior distribution concentrate near the true value in large samples (Castillo et al., 2015), although its mode (the LASSO estimator) has good estimation and variable selection properties under appropriate conditions.

Among all the Bayesian variable selection methods that aim to quantify uncertainty of included subsets of covariates, Bayesian shrinkage priors that have heavy tails and dominant concentration around zero have attracted much attention in Bayesian analysis, especially in high-dimensional model. In the recent years, a variety of continuous prior densities with good shrinkage properties have been introduced in the literature, such as the horseshoe prior (Carvalho et al. 2010), normal-gamma prior (Griffin and Brown, 2010), double-Pareto prior (Armagan et al. 2013a), Dirichlet-Laplace (DL) prior (Bhattacharya et al. 2015) and the horseshoe-plus prior (Bhadra et al., 2015). Unlike a spike-and-slab prior, these priors have a single component like the double exponential prior, but have much higher concentration near zero and have typically thicker tails, so that they mimic a point-mass prior. Treated as computationally-efficient alternatives to

point mass prior, the shrinkage priors can be represented as a global-local scale mixtures of Gaussian distribution:

$$\beta_j \sim \mathrm{N}(0, \psi_j \tau), \quad \psi_j \sim f, \quad \tau \sim g, \tag{1.1}$$

where τ controls the overall shrinkage towards the origin while ψ_j allows individual deviations for different levels of shrinkage, for $j = 1, \ldots, p$. The dominated peak at zero aggressively shrinks covariates towards zero and the heavy tails avoid over-shrinkage for true signals. These priors are referred to as continuous shrinkage priors, one-component priors or global-local priors. Recent contributions in the literature show their promising posterior concentration properties near the true value and ability to identify true non-zero coefficients. Theoretically, shrinkage priors are shown to obtain almost the same contraction rate as the point-mass prior for recovering the model parameters and the true subset of covariates in the model, for both low-dimensional (Armagan et al. 2013b) and high-dimensional (Song and Liang, 2017) data. It may be noted that under such priors, the posterior probability of hitting the exact value zero is always zero, so some appropriate thresholding procedure needs to accompany the Bayesian procedure.

1.2 Notations

Throughout the thesis, we let $\boldsymbol{\beta}$ denote the coefficients in our proposed models. Then $\boldsymbol{\beta}_*$ and σ_*^2 are the model parameters under the truth. We also let $\xi_* \subset \{1, \ldots, p\}$ be the indices of covariates with nonzero effects, for example, $\beta_{*j} \neq 0$ for all $j \in \xi_*$. In the theorems and prood, P_* and E_* stand respectively for the probability and expectation on the data under the true distribution. When in nonparametric regression model, P_* and

E_* are represented as $\mathrm{P}_{f_*,\sigma_*}$ and $\mathrm{E}_{f_*,\sigma_*}$, where f_* is the true regression mean function. For the prior density $\pi_\alpha(\cdot)$ with hyper-parameter α, we let $\mathrm{P}_\alpha(\cdot)$ be the probability on the model parameters.

For two positive sequences a and b, the relation $a \prec b$ means $\lim a/b = 0$, $a \gtrsim b$ means a/b is bounded below, and $a \asymp b$ stands for $m_1 < \liminf a/b \leq \limsup a/b < m_2$ for some positive constants m_1 and m_2.

1.3 Outline of Thesis

In Chapter 2, an additive nonparametric regression model is proposed to investigate the associations between health endpoint and mixture of pesticide exposures. Motivated by neurobehavioral data from Agricultural Health Studies (AHS), we propose a robust regression model to delineate the joint effects of multiple exposures of pesticide. The regression mean function is decomposed as main effects and interaction effects in nonparametric functions, where model uncertainty of each individual function is addressed though B-spline basis expansion and shrinkage priors on the basis coefficients. Specifically, we expand the notion of the computationally efficient Dirichlet-Laplace (DL) prior introduced in Bhattacharya et al. (2015) to multidimensional variation for simultaneous shrinkage on the basis coefficient vector. We further evaluate the prediction consistency and variable selection consistency for the proposed nonparametric regression model based on the contraction results in linear regression model with shrinkage priors (Song and Liang, 2017). Furthermore, we incorporate the neurobehavioral data to explore the health effects of multiple pesticide measurements and how the effects on each health endpoint differ from the effects on overall neurobehavioral system.

When the response variable is binary, logistic regression model is often implemented.

We consider the incorporation of Bayesian shrinkage prior in logistic regression and compare the performance with point-mass prior for recovering the true model parameters. Chapter 3 studies logistic regression model in high-dimensional data. If the shrinkage prior is heavy-tailed and allocates large probability mass in a very small neighborhood of zero, we show that the Bayesian shrinkage prior in logistic regression model achieves optimal contraction rate that can be almost the same as the point-mass prior. The theoretical results illustrate the advantages of shrinkage priors for their promising posterior consistency and computational efficiency. Examples of shrinkage priors are demonstrated and posterior consistency results can be achieved with certain prior restrictions. The simulation study under logistic regression model verifies the theoretical results by showing that shrinkage priors such as horseshoe prior and Dirichlet-Laplace prior perform quite similar as the point mass prior in estimation, variable selection and prediction.

Motivated by forecasting fine particulate matters (PM2.5) from wildland fire smoke in the state of Washington, we extend continuous shrinkage priors to the spatial setting in Chapter 4. We incorporate spatial summaries of the numerical forecasts as constructed covariates in a spatial regression model. The proposed regression model defines spatially-varying regression mean function and spatio-temporal dependence on residuals. We characterize the effects of multiple covariates using Bayesian spatial variable selection. The simulation study evaluates the additive regression model under different model assumptions (nonlinearity, spatial dependence in the mean functions and residuals) and demonstrates its advantages in both prediction and variable selection. The regression model is implemented to downscale the fine particulate matter forecasts of wildland fire smoke in Washington State in 2015 and we find an improvement of 18% in mean squared error compared to the standard linear regression calibration.

Chapter 2

Bayesian Shrinkage for Additive Nonparametric Regression

2.1 Introduction

Traditional epidemiological studies analyze the correlation between chemical exposures and a single health endpoint. As data become more complex, advanced statistical methods are needed to estimate the relationships between mixture of multiple chemicals and a suite of health endpoints to increase statistical powers and paint a more realistic picture of health risk. As a motivating application, we analyze the neurobehavioral (NB) data collected as part of the Agricultural Health Study (AHS; `http://aghealth.nih.gov/`). The data consist of 701 farmers from Iowa and North Carolina, and include measurements of 20 organophosphate pesticides and 12 neurobehavioral health endpoints. In previous works, Starks et al. (2012a, b) used linear regression model to examine the associations between multiple pesticides and each health endpoint of NB tests separately. They conduct conventional hypothesis testing under normality assumption to study the

significance of linear coefficient for each pesticide measurements.

Since the pesticide exposures may have complex nonlinear associations with health endpoints, nonparametric models are preferable considering their robustness to model assumptions. Previous literature has introduced different nonparametric regression models to delineate the associations between covariates and response variables and overcome the limitations of linear regression. Friedman (1991) developed multivariate regression splines (MARS) that defines multivariate nonparametric regression model using splines. Lin and Zhang (2006) proposed the component selection and smoothing operator (COSSO) technique that uses penalized regression to select variables. In the Bayesian framework, Linkletter et al. (2006) and Savitsky et al. (2011) modeled the response surface as a Gaussian process and used Bayesian variable selection techniques in selecting subset of covariates. Bobb et al. (2014) implemented Bayesian kernel machine regression model that assumes nonparametric associations between mixtures and health response.

While fully nonparametric models are robust to model assumptions, they suffer from a lack of interpretability and are hard to fit in high dimensions. We retain flexibility but reduce complexity by assuming an additive model that decomposes the joint function as the summation of individual nonparametric functions of each predictor. This idea has been widely implemented in methodology development such as smoothing spline ANOVA (SS-ANOVA, Gu 2002) that allows higher-order interactions among predictors. Reich et al. (2009) implemented the SS-ANOVA model with Gaussian process priors, and searched for the best model using Stochastic Search Variable Selection (SSVS, George and McCulloch, 1993) in MCMC sampling. Curtis et al. (2014) used a multivariate Laplace prior on the basis coefficients in their additive nonparametric model. The former is computationally difficult for large problems while the latter relies on large sample approximations. In this chapter, we apply the same additive regression model with basis expansion technique

as in Curtis et al. (2014) but use a different Bayesian variable selection method.

In this chapter, an additive nonparametric regression model is assumed for both main effect of each pesticide and interaction effect between pesticide mixtures. We deliver two major findings to the existing empirical literature on the health effects of multiple chemical exposures. First, we address the uncertainty about the subset of covariates to be included through B-spline basis expansion with multivariate shrinkage prior so that the level of shrinkage on basis coefficients is reported to identify the significant main effects and interactions. Second, we expand the notion of the computationally-efficient Dirichlet-Laplace (DL) prior introduced in Bhattacharya et al. (2015) to multidimensional vectors to achieve simultaneous shrinkage on the basis coefficient vector for each main effect or interaction effect function. Theoretically, we expand the current prediction consistency and variable selection consistency results for shrinkage prior under linear regression model to the proposed additive nonparametric model, where induced bias from B-spline approximation and shrinkage on multidimensional vectors pose the major challenges. Further, we incorporate neurobehavioral data from AHS to explore the health effects of multiple pesticides measurements and how the effects on each health endpoint differ from the effects on overall neurobehavioral system.

The remainder of the chapter proceeds as follows. The additive model with DL prior is described in Chapter 2.2, as well as the restrictions imposed on model parameters for identifiability. Chapter 2.3 introduces the posterior sampling scheme for model parameters. The asymptotic properties and variable selection performance are discussed in Chapter 2.4 and the proof details of theorems are included. Chapter 2.5 presents a simulation study that compares our model with other nonparametric regression models in terms of selecting important variables. Chapter 2.6 implements our model into NB data from AHS study.

2.2 Model Description and Prior Specification

2.2.1 Main-effect-only model

We first demonstrate the main-effect-only model that assumes the regression mean function is the summation of main effect functions for individual covariate. Let the data be (Y, X), where Y is the response variable denoting the health endpoint, $X_{n\times p} = (X_1, \ldots, X_p)^T$ is a p-vector of chemical exposure measurements. For the nonparametric regression model of Y on the covariates X, we assume that

$$Y = \mu + f(X_1, \ldots, X_p) + \varepsilon, \tag{2.1}$$

where μ is the intercept and the error term $\varepsilon \sim \mathrm{N}(0, \sigma^2)$. Assuming the covariates in the data affect the response variable in mutually independent nonparametric functions, the joint effect is decomposed as the sum of individual main effect functions: $f(X_1, \ldots, X_p) = \sum_{j=1}^{p} f_j(X_j)$, where $f_j(X_j)$ is the univariate nonparametric function of X_j.

If each main effect function of individual covariate is sufficiently smooth, it can be approximated using B-spline basis expansions with a predetermined number of basis functions, m. For the covariate matrix X with elements all scaled to $(0, 1)$ interval, each main effect function is approximated by $f_j(X_j) \approx \sum_{r=1}^{m} B_r(X_j)\beta_{jr}$, where $B_r(X_j)$, $r = 1, \ldots, m$, are B-spline basis terms for X_j. With this approximation, the nonlinear effects of X_j is transformed into a linear combination of its basis terms with the m-vector basis coefficients $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jm})^T$. The regression model in (2.1) can be written as

$$Y = \mu + \sum_{j=1}^{p}\sum_{r=1}^{m} B_r(X_j)\beta_{jr} + \varepsilon. \tag{2.2}$$

With the regression model defined in (2.2), model uncertainty and main effect of each covariate X_j are addressed through the basis coefficients $\boldsymbol{\beta}_j$. The m-vector coefficients $\boldsymbol{\beta}_j$'s are assgined multivaraite nromal priors with mean zero and different variance factor-sacross $j = 1, \ldots, p$: $\boldsymbol{\beta}_j \overset{ind}{\sim} \mathrm{N}_m(0, \sigma^2\lambda_j \mathrm{I}_m)$. The local variance factors λ_j's determine the shrinkage on the m-vector basis coefficients so that the problem of selecting important main effects reduces to shrinkage of the λ_j's. Using the Dirichelt-Laplace (DL) prior in Bhattacharya et al. (2015), λ_j follows an exponential distribution with scale parameter $\tau\phi_j$, where τ is the global factor that determines the tail of the marginal distribution of λ_j's and $\phi_j > 0$, with $\sum_{j=1}^{p} \phi_j = 1$, is the proportion of variance allocated to covariate X_j. Further, a Dirichlet distribution on $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^T$ and a gamma distribution on τ are assumed. The hyper-parameter α in Dirichlet distribution controls the level of shrinkage so that smaller α gives more concentration around zero and thus a sparser model. Thus, the uncertainty on each nonparametric main effect function is addressed by the implementation of multivariate DL prior on the B-spline basis coefficients. The Bayesian model on the model parameters is formulated as follows,

$$\boldsymbol{\beta}_j | \lambda_j, \sigma^2 \overset{ind}{\sim} \mathrm{N}_m(0, \sigma^2\lambda_j \mathrm{I}_m), \ j = 1, \ldots, p, \tag{2.3}$$

$$\lambda_j | \phi_j, \tau \overset{ind}{\sim} \mathrm{Exp}(\phi_j\tau), \ j = 1, \ldots, p, \tag{2.4}$$

$$\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^T \sim \mathrm{Dirichlet}(\alpha, \ldots, \alpha), \quad \tau \sim \mathrm{Gamma}(p\alpha, 2). \tag{2.5}$$

The proposed Bayesian hierarchical method for the additive nonparametric model is a multivariate extension of the DL prior in linear regression. As a shrinkage prior on the B-spline basis coefficents, the proposed multivariate DL method leads to a slightly different prior from the original univariate DL prior of Bhattacharya et al. (2015). When $m = 1$ and $\sigma^2 = 1$, our multivariate DL prior model is $\beta_j | \lambda_j \sim \mathrm{N}(0, \lambda_j)$ and $\lambda_j | \phi_j, \tau \sim \mathrm{Exp}(\phi_j\tau)$.

After integrating out λ_j, we get the marginal prior distribution of β_j given ϕ_j, τ is double exponential distribution denoted as $\text{DE}(\sqrt{\phi_j \tau/2})$. In linear regression model, however, the DL prior on coefficient is $\beta_j|\phi_j, \tau \sim \text{DE}(\phi_j \tau)$. Therefore, for this case, the original DL prior plays more mass at zero than the proposed multivariate extension with only one basis term.

2.2.2 Main and interaction effects model

Assuming the effects of covariates on response are not additive but with two-way interactions, the underlying joint function includes both main effects and interactions,

$$Y = \mu + \sum_{j=1}^{p} f_j(X_j) + \sum_{k=1}^{p-1} \sum_{l=k+1}^{p} f_{kl}(X_k, X_l) + \varepsilon, \tag{2.6}$$

where the second-order term $f_{kl}(X_k, X_l)$ represents the interaction effects on health endpoint between covariates X_k and X_l. We consider only two-way interactions in this model since higher order interactions are less interpretable and including them will make the computational complexity beyond manageable limits.

Using the B-spline basis expansion described in Chapter 2.2.1 to represent each individual main effect function, we can use the outer product of the B-spline basis terms for interaction effect functions:

$$f_{kl}(X_k, X_l) \approx \sum_{s=1}^{m^*} \sum_{t=1}^{m^*} B_s^*(X_k) B_t^*(X_l) \beta_{klst}. \tag{2.7}$$

We use a potentially different basis function number m^* for the interaction effects, and the basis functions ($B_s^*(X)$ for $s = 1, \ldots, m^*$) may also differ. We propose similar multivariate DL prior on the basis coefficients for the interaction effect function:

Normal priors are placed on the coefficients $\beta_{klst} \overset{ind}{\sim} \mathrm{N}(0, \sigma^2\lambda_{kl})$, for $s, t = 1, \ldots, m^*$. The DL prior is imposed on the local variance factors λ_{kl}:

$$\boldsymbol{\beta}_{kl}|\lambda_{kl}, \sigma^2 \overset{ind}{\sim} \mathrm{N}_{m^*\times m^*}(0, \sigma^2\lambda_{kl}\mathrm{I}_{m^*\times m^*}), l > k, k = 1, \ldots, p-1, \tag{2.8}$$

$$\lambda_{kl}|\phi^*_{kl}, \tau^* \overset{ind}{\sim} \mathrm{Exp}(\phi^*_{kl}\tau^*), l > k, k = 1, \ldots, p-1, \tag{2.9}$$

$$\boldsymbol{\phi^*} = (\phi_{12}, \ldots, \phi_{1p}, \ldots, \phi_{p-1,p})^T \sim \mathrm{Dirichlet}(\alpha, \ldots, \alpha), \tag{2.10}$$

$$\tau^* \sim \mathrm{Gamma}\left(\frac{p(p-1)}{2}\alpha, 2\right), \tag{2.11}$$

where we assume the same sparsity level α for both main effects and interactions.

2.2.3 Identifiability constraints

The additive model in (2.6) is not uniquely identified without constraints on the functions f_j and f_{kl}. For example, adding a constant to f_j and subtracting the same constant from $f_{j'}$ for $j \neq j'$ gives the same mean regression function but different individual main effect functions. While other restrictions are possible, we restrict the main effect functions to integrate to zero, $\int_0^1 f_j(x_j)dx_j = 0$, $j = 1, \ldots, p$. For the functions of interactions, we assume that the bivariate functions integrate to zero in both directions, $\int_0^1 f_{kl}(x_k, x_l)dx_k = 0$ for all x_l and $\int_0^1 f_{kl}(x_k, x_l)dx_l = 0$ for all x_k, so that the main effect functions are not in the span of the interaction functions.

The restrictions on main effect functions can be written as

$$\int_0^1 f_j(x)dx = \int_0^1 \left[\sum_{r=1}^m \beta_{jr}B_r(x)\right]dx = \sum_{r=1}^m \beta_{jr}\left[\int_0^1 B_r(x)dx\right] \overset{\text{def}}{=} \sum_{r=1}^m \beta_{jr}D_r = 0, \tag{2.12}$$

where $D_r \overset{\text{def}}{=} \int_0^1 B_r(x)dx$. Therefore, the integral restriction is equivalent to a linear

restriction on the basis coefficient $\boldsymbol{\beta}_j$, $\sum_{r=1}^{m} \beta_{jr} D_r = 0$. For the B-spline basis,

$$D_r = \begin{cases} r/[d(m-d+1)], & r = 1, \ldots, (d-1), \\ 1/(m-d+1), & r = d, \ldots, (m-d+1), \\ (m-r+1)/[d(m-d+1)], & r = (m-d+2), \ldots, m, \end{cases}$$

where m is the number of B-spline basis functions and d is the degree of B-spline basis.

The restrictions on each bivariate function of interactions are also treated as linear constraints on the coefficients for the B-spline expansion in (2.7):

$$\int_0^1 f_{kl}(x_k, x_l) dx_k = \sum_{s=1}^{m^*} \sum_{t=1}^{m^*} \left[\int_0^1 B_s^*(x_k) dx_k \right] B_t^*(x_l) \beta_{klst} \stackrel{\text{def}}{=} \sum_{t=1}^{m^*} B_t^*(x_l) \left[\sum_{s=1}^{m^*} D_s^* \beta_{klst} \right] = 0, \tag{2.13}$$

$$\int_0^1 f_{kl}(x_k, x_l) dx_l = \sum_{s=1}^{m^*} \sum_{t=1}^{m^*} B_s^*(x_k) \left[\int_0^1 B_t^*(x_l) dx_l \right] \beta_{klst} \stackrel{\text{def}}{=} \sum_{s=1}^{m^*} B_s^*(x_k) \left[\sum_{t=1}^{m^*} D_t^* \beta_{klst} \right] = 0, \tag{2.14}$$

where we assume $D_s^* \stackrel{\text{def}}{=} \int_0^1 B_s^*(x) dx$ for $s = 1, \ldots, m^*$. The restrictions in (2.13) and (2.14) hold for all x_k and x_l if an only if

$$\begin{aligned} \sum_{s=1}^{m^*} D_s^* \beta_{klst} &= 0, \text{ for all } t = 1, \ldots, m^*; \\ \sum_{t=1}^{m^*} D_t^* \beta_{klst} &= 0, \text{ for all } s = 1, \ldots, m^*. \end{aligned}$$

The restrictions on interaction functions $f_{kl}(x_k, x_l)$ are composed of $2m^*$ linear functions on $\boldsymbol{\beta}_{kl}$, for $k < l$ and $k = 1, \ldots, p-1$.

2.3 Posterior Sampling

We now describe the computational algorithm for the main-effect-only model. The regression model with both main and interaction effects in the mean function is very similar. We sample the parameters using the combination of Gibbs sampling and direct sampling. The sampler cycles through (i) $\boldsymbol{\beta}|\lambda, \phi, \tau, Y, X$, (ii) $\lambda|\boldsymbol{\beta}, \phi, \tau, Y, X$ and (iii) $\phi, \tau|\lambda, \boldsymbol{\beta}, Y, X$. Within step (iii), it follows direct sampling of (iiia) $\tau|\phi, \lambda$ and (iiib)$\phi|\lambda$.

(i) Given λ_j, Y and X_j, the conditional posterior distribution for $\boldsymbol{\beta}_j$ is m-dimensional multivariate normal with mean $\boldsymbol{\mu}_{\beta_j}$ and variance matrix Σ_{β_j}, where

$$\boldsymbol{\mu}_{\beta_j} = \left(B(X_j)^T B(X_j) + \frac{\mathrm{I}_m}{\lambda_j} \right)^{-1} B(X_j)^T \left(Y - \sum_{l=1, l\neq j}^{p} B(X_l)\beta_l \right), \tag{2.15}$$

$$\Sigma_{\beta_j} = \left(B(X_j)^T B(X_j) + \frac{\mathrm{I}_m}{\lambda_j} \right)^{-1} \sigma^2. \tag{2.16}$$

For the restrictions on basis coefficients in (2.12), we sample $\boldsymbol{\beta}_j \big| \sum_{r=1}^{m} \beta_{jr} D_r = 0$ from conditional multivariate normal $\mathrm{N}_m(\boldsymbol{\mu}^*_{\beta_j}, \Sigma^*_{\beta_j})$, where

$$\boldsymbol{\mu}^*_{\beta_j} = \boldsymbol{\mu}_{\beta_j} - \Sigma_{\beta_j} D (D^T \Sigma_{\beta_j} D)^{-1} D^T \boldsymbol{\mu}_{\beta_j}, \tag{2.17}$$

$$\Sigma^*_{\beta_j} = \Sigma_{\beta_j} - \Sigma_{\beta_j} D (D^T \Sigma_{\beta_j} D)^{-1} D^T \Sigma_{\beta_j}, \tag{2.18}$$

where $D = (D_1, \ldots, D_m)^T$.

(ii) Given ϕ_j, τ and $\boldsymbol{\beta}_j$, the variance component λ_j is sampled from generalized inverse Gaussian distribution $\mathrm{GiG}\Big(1 - \frac{m}{2},\ \frac{2}{\phi_j \tau},\ \frac{\boldsymbol{\beta}_j^T \boldsymbol{\beta}_j}{\sigma^2}\Big)$.

(iiia) Given $\phi_1, \ldots, \phi_p$ and $\lambda_1, \ldots, \lambda_p$, the global parameter τ in DL prior is sampled from generalized inverse Gaussian distribution $\mathrm{GiG}\Big(p(\alpha - 1),\ 1,\ 2\sum_{j=1}^{p} \frac{\lambda_j}{\phi_j}\Big)$.

(iiib) Now given $\lambda_1, \ldots, \lambda_p$, first sample T_j, $j = 1, \ldots, p$, independently from general-

ized inverse Gaussian distribution $\text{GiG}(\alpha - 1,\ 1,\ 2\lambda_j)$ and then let $\phi_j = T_j / \sum_{l=1}^{p} T_l$.

2.4 Asymptotic Properties

In this section, we will study the asymptotic properties of the additive nonparametric regression model for individual main effects with the multivariate DL prior. Since predictors are considered deterministic in our setting, extension to include interactions is similar except that the model will be more complicated. For asymptotic properties, we denote p and m as p_n and m_n respectively to reflect that they are increasing in n.

2.4.1 Notations and assumptions

For a fixed $n \times p_n$ covariate matrix $X = (X, \ldots, X_{p_n})$, we consider the additive nonparametric model $Y = \sum_{j=1}^{p_n} f_j(X_j) + \sigma\varepsilon$, where each additive function is approximated by B-spline basis expansion: $f_j(X_j) = B(X_j)\boldsymbol{\beta}_j + \sigma\delta$, where $B(X_j) = (B_1(X_j), \ldots, B_{m_n}(X_j))$ is the $n \times m_n$ matrix of basis terms for X_j and δ denotes the bias induced from basis expansion approximation. Therefore, the true additive regression model is represented as

$$Y = \sum_{j=1}^{p_n} B(X_j)\boldsymbol{\beta}_j + \sigma\delta + \sigma\varepsilon \stackrel{\text{def}}{=} B(X)\boldsymbol{\beta} + \sigma\delta + \sigma\varepsilon,$$

where $B(X) = [B(X_1), \ldots, B(X_{p_n})]$ is the $n \times p_n m_n$ matrix for B-spline basis expansion terms, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_{p_n}^T)^T$ is the $p_n m_n$-vector with each m_n-dimensional components corresponding to each covariate X_j and ε is an n-dimensional standard normal vector. We study a Bayesian approach based on a continuous shrinkage prior that selects the nonzero vectors among the m_n-dimensional vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{p_n}$. After integrating out parameters ϕ_j, the hierarchical Bayesian model in Chapter 2.2.1 is represented by, for

$j = 1, \ldots, p_n$,

$$\boldsymbol{\beta}_j | \lambda_j, \sigma^2 \overset{ind}{\sim} \mathrm{N}_m(0, \lambda_j \sigma^2 \mathrm{I}_m), \tag{2.19}$$

$$\lambda_j | \psi_j \overset{ind}{\sim} \mathrm{Exp}(\psi_j), \ \psi_j \sim \mathrm{Gamma}(\alpha, 2). \tag{2.20}$$

We assume the density function on $\boldsymbol{\beta}_j$ is denoted as $\pi_\alpha(\boldsymbol{\beta}_j)$ with hyper-parameter α. We let $f_{*j}(\cdot)$ be the true function for covariate X_j, $\boldsymbol{\beta}_*$ and σ_* be the true values of parameters, $\xi_* \subset \{1, \ldots, p_n\}$ be the indices of covariates with nonzero effects such that $f_{*j}(X_j) \neq 0$ for $j \in \xi_*$, and the true sparsity level is $s_* = |\xi_*|$.

Next we state some regularity conditions on the eigen structure of the B-spline basis expansion matrix $B(X)$ with respect to the sequence $\{\epsilon_n\}$, which will be defined later. These assumptions are similar as those in Song and Liang (2017) where they presented the asymptotic properties of shrinkage priors in the linear regression model. The difference lies in the fact that we are dealing with the design matrix of basis terms, and each additive nonparametric function is estimated by B-spline basis expansion so that estimation bias is introduced.

- $A_1(1)$: The data dimensionality satisfies that $m_n p_n \geq n$
- $A_1(2)$: Assume κ is the minimum smoothness parameter for all additive functions, so that the bias induced by m_n-dimensional basis expansion is $\delta \asymp m_n^{-\kappa}$
- $A_1(3)$: The rank of $B(X)$ is n and $B(X)^T B(X)$ has n positive eigenvalues denoted as $nd_1/m_n, \ldots, nd_n/m_n$, where $d_1, \ldots, d_n$ are bounded away from zero
- $A_1(4)$: There exists some integer $\bar{p}$ and a fixed constant d_0' with $\bar{p} m_n \log p_n \succ n\epsilon_n^2$, such that for any $q \leq \bar{p}$, the basis expansion sub-matrix $\tilde{B}(X)$ consisting qm_n columns of $B(X)$, $d_{\min}\left(\tilde{B}(X)^T \tilde{B}(X)\right) \geq nd_0'/m_n$

- $A_2(1)$: $\{\epsilon_n\}$ sequence is assumed to satisfy that $s_* m_n \log p_n \prec n\epsilon_n^2$ and $\epsilon_n \succ m_n^{-\kappa}$
- $A_2(2)$: $\min_{j\in\xi_*}\left\{\frac{\|\boldsymbol{\beta}_{*j}/\sigma_*\|}{\sqrt{m_n}}\right\} \succ \epsilon_n$ and $\max_j\left\{\frac{\|\boldsymbol{\beta}_{*j}/\sigma_*\|}{\sqrt{m_n}}\right\} \leq \gamma_3 E$ for fixed $\gamma_3 \in (0,1)$, and E is nondecreasing with n
- A_3: There exists an integer $\bar{p}$, which depends on n and p_n, and two constants d and d_0' such that $\bar{p} m_n \log p_n \succ n\epsilon_n^2$ and $nd_0/m_n \geq d_{\max}\left(\tilde{B}(X)^T\tilde{B}(X)\right) \geq d_{\min}\left(\tilde{B}(X)^T\tilde{B}(X)\right) \geq nd_0'/m_n$ for any $q \leq \bar{p}$ and any sub-matrix $\tilde{B}(X)$ of the basis expansion matrix $B(X)$

Note that assumption $A_1(3)$ is an extension of linear regression model in Song and Liang (2017), by replacing covariate matrix X to B-spline basis design matrix $B(X)$. From Lemma $A.9$ in Yoo and Ghosal (2016), we combine the B-spline property with the linear regression model assumption, so $A_1(3)$ is specified. The assumption in $A_2(1)$ restricts ϵ_n to be $\max(\sqrt{s_* m_n \log p_n/n}, m_n^{-\kappa})$.

2.4.2 Main results

For the nonparametric function defined in the regression model where no model parameter is assumed, we want to demonstrate that the consistency holds for the prediction on the joint effects of covariates. Therefore, the following theorem proves that, when the B-spline basis expansion is implemented, the prediction performance of $B(X)\boldsymbol{\beta}$ is asymptoticly concentrated around the additive mean function under the truth.

Theorem 2.4.1 *For the regression model* $Y = \sum_{j=1}^{m_n} f_j(X_j) + \sigma\varepsilon = B(X)\boldsymbol{\beta} + \sigma\delta + \sigma\varepsilon$, *the basis expansion bias satisfies that* $\|\delta\| \asymp \sqrt{n} m_n^{-\kappa}$, *where* κ *is the smoothness parameter. Assume the assumptions in* A_1, A_2 *and* A_3 *hold for the design matrix* $B(X)$ *and the basis coefficients* $\boldsymbol{\beta}_j$ *for covariate* X_j *follow prior density* $\pi_\alpha(\cdot)$ *as defined in (*(2.19)*) and*

((2.20)) with $\alpha \asymp p_n^{-(1+\nu)}$ for $\nu > 0$, then we conclude the posterior concentration of the predictive distribution,

$$P_* \left(P_\alpha \left(\left\| B(X)\boldsymbol{\beta} - \sum_{j=1}^{p_n} f_{*j}(X_j) \right\| \geq c_0 \sqrt{n}\epsilon_n \middle| X, Y \right) \geq e^{-c_1 n \epsilon_n^2} \right) \leq e^{-c_2 n \epsilon_n^2}, \tag{2.21}$$

for some constants c_0, c_1 and c_2.

Theorem 2.4.1 shows posterior concentration rate $\sqrt{n}\epsilon_n$ for the predictions at the observation points. Therefore, for the covariates matrix X, the prediction $B(X)\boldsymbol{\beta}$ obtained from the regression model has concentration rate close to $\sqrt{n}\max\{\sqrt{s_* m_n \log p_n / n},\ m_n^{-\kappa}\}$ in probabilities around the true mean function $\sum_{j=1}^{p_n} f_{*j}(X_j)$.

Theorem 2.4.2 *Denote the subset model by $\xi(a_n) = \{j : \|\boldsymbol{\beta}_j/\sigma\| > a_n\}$ for $na_n p_n \prec \log p_n$. Assume the conditions in Theorem 2.4.1 hold and $\min_{j\in\xi_*} \|\boldsymbol{\beta}_{*j}\|/\sqrt{m_n} \succ \epsilon_n$, then*

$$P_* \left(P_\alpha(\xi(a_n) = \xi_* | X, Y) > 1 - p_n^{-\mu''} \right) > 1 - p_n^{-\mu'}, \tag{2.22}$$

for some positive constants μ' and μ''.

Theorem 2.4.2 shows the posterior variable selection consistency given that $a_n \prec \log p_n / n p_n$ and the prior density is moderately flat at nonzero basis coefficients $\boldsymbol{\beta}_{*j}/\sigma_*$.

2.4.3 Proof of Theorem 2.4.1 and Theorem 2.4.2

The following lemmas prepare for the proof of concentration properties on prior density $\pi_\alpha(\boldsymbol{\beta}_j/\sigma)$ for $j = 1, \ldots, p_n$ on m_n-dimensional vector $\boldsymbol{\beta}_j/\sigma$. For the defined shrinkage prior (2.19) and (2.20), we let $p(\lambda_j|\psi_j)$ denote the prior density for λ_j given ψ_j and $p_\alpha(\psi)$ denote the prior density for ψ_j given the hyper-parameter α. In the following lemmas on

the multivariate Dirichlet-Laplace (DL) prior, we want to show that it concentrates within a small neighborhood around zero with a dominated and steep peak, while retaining heavy tails away from zero.

Lemma 2.4.3 *For any positive constant $\mu > 0$ and thresholding value $a_n \leq \sqrt{\frac{d_0'}{m_n}}\epsilon_n/p_n$,*

$$1 - \int_{\|\boldsymbol{\beta}|\leq a_n} \pi_\alpha(\boldsymbol{\beta})d\boldsymbol{\beta} \leq p_n^{-(1+\mu)}, \tag{2.23}$$

if the Dirichlet parameter satisfies $\alpha \asymp p_n^{-(1+\nu)}$ for $\nu > \mu$.

Proof For a vector $\boldsymbol{\beta}$ of dimension m_n, the prior density is defined in (2.19) and (2.20). When λ is given, $T \stackrel{\text{def}}{=} \|\boldsymbol{\beta}\|^2/\lambda \sim \chi^2_{m_n}$. Next, we have

$$\begin{aligned} 1 - \int_{\|\boldsymbol{\beta}\|\leq a_n} \pi_\alpha(\boldsymbol{\beta})d\boldsymbol{\beta} &= \int_0^\infty \int_0^\infty \mathrm{P}\left(T > \frac{a_n^2}{\lambda}\right) p(\lambda|\psi)p_\alpha(\psi)d\psi d\lambda \\ &= \int_0^{a_n^2/p_n} \mathrm{P}\left(T > \frac{a_n^2}{\lambda}\right) \left[\int_0^\infty p(\lambda|\psi)p_\alpha(\psi)d\psi\right] d\lambda \\ &\quad + \int_{a_n^2/p_n}^\infty \mathrm{P}\left(T > \frac{a_n^2}{\lambda}\right) \left[\int_0^\infty p(\lambda|\psi)p_\alpha(\psi)d\psi\right] d\lambda \end{aligned} \tag{2.24}$$

For any constant $\mu > 0$, we can conclude that

$$\int_0^{a_n^2/p_n} \mathrm{P}\left(T > \frac{a_n^2}{\lambda}\right) \left[\int_0^\infty p(\lambda|\psi)p_\alpha(\psi)d\psi\right] d\lambda \leq \mathrm{P}(T > p_n)\mathrm{P}_\alpha(\lambda < a_n^2/p_n) \leq e^{-p_n/2}, \tag{2.25}$$

where the last inequality is obtained from the tail property of χ^2-distribution.

Next, from Lemma 3.3 in Bhattacharya et al. (2015), it follows that

$$\int_{a_n^2/p_n}^\infty \mathrm{P}\left(T > \frac{a_n^2}{\lambda}\right) \left[\int_0^\infty p(\lambda|\psi)p_\alpha(\psi)d\psi\right] d\lambda \leq \mathrm{P}_\alpha(\lambda > a_n^2/p_n) \leq C_1\alpha\log(p_n/a_n^2) \tag{2.26}$$

for a constant $C_1 > 0$. Combining the inequalities in (2.24), (2.25) and (2.26), we have

$$1 - \int_{\|\boldsymbol{\beta}\| \le a_n} \pi_\alpha(\boldsymbol{\beta}) d\boldsymbol{\beta} \le e^{-p_n/2} + C_1 \alpha \log(p_n/a_n^2) \le p_n^{-(1+\mu)}.$$

for $\alpha \asymp p_n^{-(1+\nu)}$ where $\nu > \mu$.

Lemma 2.4.4 *For a small constant $\eta > 0$ and E defined in assumption $A_2(2)$,*

$$-\log\left(\int_{\sum_{j=1}^{p_n-s_*}\|\boldsymbol{\beta}_j\| \le \eta\sqrt{m_n}\epsilon_n} \prod_{j=1}^{p_n-s_*} \pi_\alpha(\boldsymbol{\beta}_j) d\boldsymbol{\beta}_1 \ldots d\boldsymbol{\beta}_{p_n-s_*}\right) \prec n\epsilon_n^2, \tag{2.27}$$

$$-\log\left(\inf_{\|\boldsymbol{\beta}\| \le \sqrt{m_n}E} \pi_\alpha(\boldsymbol{\beta})\right) \prec n\epsilon_n^2/s_*. \tag{2.28}$$

Proof When λ_j is given, the m_n-dimensional vector $\boldsymbol{\beta}_j$ satisfies that $\|\boldsymbol{\beta}_j\|/\sqrt{\lambda_j} \overset{\text{iid}}{\sim} \chi_{m_n}$ for $j = 1, \ldots, p_n$. Therefore, for the prior of λ_j defined in (2.20),

$$\mathrm{E}(\|\boldsymbol{\beta}_j\|) = \mathrm{E}\big[\mathrm{E}\big(\|\boldsymbol{\beta}_j\|\big|\lambda_j\big)\big] = \mathrm{E}\big[\sqrt{\lambda_j}\mathrm{E}(\|\boldsymbol{\beta}_j\|/\sqrt{\lambda_j}\big|\lambda_j)\big] = \frac{\sqrt{2}\Gamma((m_n+1)/2)}{\Gamma(m_n/2)}\mathrm{E}(\sqrt{\lambda_j}), \tag{2.29}$$

and then $\mathrm{E}(\sqrt{\lambda_j}) = \mathrm{E}\big[\mathrm{E}(\sqrt{\lambda_j}|\psi_j)\big] = \Gamma(3/2)\mathrm{E}(\sqrt{\psi_j}) = \Gamma(3/2)\frac{\sqrt{2}\Gamma(\alpha+1/2)}{\Gamma(\alpha)}$. Then for a constant $C_2 > 0$ free of α, $\mathrm{E}(\|\boldsymbol{\beta}_j\|) = C_2\sqrt{m_n}\alpha$. By the Central Limit Theorem,

$$\mathrm{P}_\alpha\left(\sum_{j=1}^{p_n-s_*}\frac{\|\boldsymbol{\beta}_j\|}{\sqrt{m_n}} \le \eta\epsilon_n\right) = \mathrm{P}_\alpha\left(\frac{1}{p_n-s_*}\sum_{j=1}^{p_n-s_*}\frac{\|\boldsymbol{\beta}_j\|}{\sqrt{m_n}} \le \frac{\eta\epsilon_n}{p_n-s_*}\right) \ge 1/2 \tag{2.30}$$

if $\alpha \le \frac{C_3\epsilon_n}{p_n-s_*}$, which can be verified since $\alpha \asymp p_n^{-(1+\nu)}$ for $\nu > 0$. Therefore, we show that

$$-\log\left(\int_{\sum_{j=1}^{p_n-s_*}\frac{\|\boldsymbol{\beta}_j\|}{\sqrt{m_n}} \le \eta\epsilon_n} \prod_{j=1}^{p_n-s_*} \pi_\alpha(\boldsymbol{\beta}_j) d\boldsymbol{\beta}_1 \ldots d\boldsymbol{\beta}_{p_n-s_*}\right) \prec n\epsilon_n^2. \tag{2.31}$$

We need to prove that $-\log\left(\inf_{\|\boldsymbol{\beta}\|\le\sqrt{m_n}E}\pi_\alpha(\boldsymbol{\beta})\right)\prec n\epsilon_n^2/s_*$. For s_1/s_2 close to zero,

$$\begin{aligned}\inf_{\|\boldsymbol{\beta}\|\le\sqrt{m_n}E}\pi_\alpha(\boldsymbol{\beta}) =& \int_0^\infty\int_0^\infty (2\pi\lambda)^{-m_n/2}e^{-m_nE^2/2\lambda}p(\lambda|\psi)p_\alpha(\psi)d\lambda d\psi\\ \ge& (2\pi s_2)^{-m_n/2}e^{-m_nE^2/2s_1}\mathrm{P}_\alpha(s_1<\lambda<s_2). \end{aligned}\tag{2.32}$$

As λ follows prior distribution in (2.20) for a given hyper-parameter α, we have

$$\begin{aligned}\mathrm{P}_\alpha(s_1<\lambda<s_2) =& \int_0^\infty (e^{-s_1/\psi}-e^{-s_2/\psi})p_\alpha(\psi)d\psi \ge \frac{e^{-s_2/2}}{\Gamma(\alpha)}(s_2-s_1)\int_0^{s_2}\psi^{\alpha-2}e^{-s_2/\psi}d\psi\\ \ge& \alpha e^{-s_2/2}(s_2-s_1)s_2^{-(1-\alpha)}\int_1^\infty t^{-\alpha}e^{-t}dt\\ \ge& C_4\alpha(1-s_1/s_2)e^{-s_2/2}\end{aligned}$$

for a constant $C_4>0$. Therefore, when $m_n/s_1\prec n\epsilon_n^2/s_*$ and $s_2\prec n\epsilon_n^2/s_*$, we have

$$-\log\left(\inf_{\|\boldsymbol{\beta}\|\le\sqrt{m_n}E}\pi_\alpha(\boldsymbol{\beta})\right)\le\frac{m_n}{2}\log(2\pi s_2)+\frac{m_nE^2}{2s_1}+(1+\nu)\log p_n+s_2/2\prec n\epsilon_n^2/s_*.\tag{2.33}$$

Lemma 2.4.5 *The prior density $\pi_\alpha(\boldsymbol{\beta}_j/\sigma)$ satisfies*

$$\max_{j\in\xi_*}\sup_{\|\boldsymbol{x}_1\|,\|\boldsymbol{x}_2\|\in[\|\boldsymbol{\beta}_{*j}/\sigma\|-\sqrt{m_n}C'\epsilon_n,\|\boldsymbol{\beta}_{*j}/\sigma\|+\sqrt{m_n}C'\epsilon_n]}\frac{\pi_\alpha(\boldsymbol{x}_1)}{\pi_\alpha(\boldsymbol{x}_2)}<l_n,\tag{2.34}$$

for a constant $C'>0$ and $s_*\log l_n\prec\log p_n$.

Proof By assumption, $\min_{j\in\xi_*}\frac{\|\boldsymbol{\beta}_{*j}/\sigma_*\|}{\sqrt{m_n}}>\epsilon_n$, $\max_{j\in\xi_*}\frac{\|\boldsymbol{\beta}_{*j}/\sigma_*\|}{\sqrt{m_n}}<E$ and

$$\begin{aligned}\pi_\alpha(\boldsymbol{x}) =& \int_0^\infty\int_0^\infty \phi(\boldsymbol{x}|\lambda)p(\lambda|\psi)p_\alpha(\psi)d\psi d\lambda\\ =& \int_0^\infty (2\pi\lambda)^{-m_n/2}e^{-\|\boldsymbol{x}\|^2/2\lambda}\int_0^\infty p(\lambda|\psi)p_\alpha(\psi)d\psi d\lambda\end{aligned}$$

where the function of λ in integrand $f^*(\lambda) \stackrel{\text{def}}{=} (2\pi\lambda)^{-m_n/2} e^{-\|\boldsymbol{x}\|^2/2\lambda}$ is concave and it is increasing on $(0, \|\boldsymbol{x}\|^2/m_n]$ and decreasing on $(\|\boldsymbol{x}\|^2/m_n, \infty)$. Then we can separate the integral by $\|\boldsymbol{x}\|^2/m_n \log p_n$ so that $f^*(\lambda)$ is increasing until the point $\|\boldsymbol{x}\|^2/m_n$. Since

$$\int_0^{\|\boldsymbol{x}\|^2/m_n \log p_n} f^*(\lambda) \int_0^\infty p(\lambda|\psi) p_\alpha(\psi) d\psi d\lambda \leq \left(\frac{2\pi\|\boldsymbol{x}\|^2}{m_n \log p_n} \right)^{-m_n/2} e^{-m_n \log p_n/2} \tag{2.35}$$

$$\begin{aligned}
\int_{\|\boldsymbol{x}\|^2/m_n \log p_n}^\infty f^*(\lambda) \int_0^\infty p(\lambda|\psi) p_\alpha(\psi) d\psi d\lambda &\geq \int_{2\|\boldsymbol{x}\|^2/m_n \log p_n}^{\|\boldsymbol{x}\|^2/m_n} f^*(\lambda) \int_0^\infty p(\lambda|\psi) p_\alpha(\psi) d\psi d\lambda \\
&\geq (2\pi\|\boldsymbol{x}\|^2/m_n)^{-m_n/2} e^{-m_n \log p_n/2} \mathrm{P}_\alpha \left(\frac{2\|\boldsymbol{x}\|^2}{m_n \log p_n} < \lambda < \frac{\|\boldsymbol{x}\|^2}{m_n} \right) \\
&\leq C_5 (2\pi\|\boldsymbol{x}\|^2/m_n)^{-m_n/2} e^{-m_n \log p_n/4} \alpha e^{-\|\boldsymbol{x}\|^2/2m_n}.
\end{aligned} \tag{2.36}$$

And $C_5 > 0$ is constant free of n. The inquality in (2.36) is obtained from the conlcusion in the proof of Lemma 2.4.4. Then for $m_n > n\epsilon_n^2$, we can prove that

$$\begin{aligned}
&\frac{\int_0^{\|\boldsymbol{x}\|^2/m_n \log p_n} f^*(\lambda) \int_0^\infty p(\lambda|\psi) p_\alpha(\psi) d\psi d\lambda}{\int_{\|\boldsymbol{x}\|^2/m_n \log p_n}^\infty f^*(\lambda) \int_0^\infty p(\lambda|\psi) p_\alpha(\psi) d\psi d\lambda} \leq \frac{(\log p_n)^{m_n/2} e^{-m_n \log p_n/2}}{C_5 e^{-m_n \log p_n/4} \alpha e^{-\|\boldsymbol{x}\|^2/2m_n}} \\
=&\exp\left\{ -\log C_5 + \frac{m_n}{2} \log\log p_n - \frac{m_n \log p_n}{4} + (1+\nu) \log p_n + \frac{\|\boldsymbol{x}\|^2}{2m_n} \right\} \\
\leq&\exp\{ -C_5' n\epsilon_n^2 \}
\end{aligned}$$

for a constant $C_5' > 0$. The proved inequality on the ratio implies that, for a given $\boldsymbol{x}$ such that $\|\boldsymbol{x}\|^2/\sqrt{m_n} \in (\epsilon_n, E)$, the prior density of $\boldsymbol{x}$ has the dominated mass on $\lambda \in (\|\boldsymbol{x}\|^2/m_n \log p_n, \infty)$:

$$\frac{\pi_\alpha(\boldsymbol{x})}{\int_{\|\boldsymbol{x}\|^2/m_n \log p_n}^\infty \int_0^\infty \phi(\boldsymbol{x}|\lambda) p(\lambda|\psi) p_\alpha(\psi) d\psi d\lambda} \asymp 1. \tag{2.37}$$

From the conclusion in (2.37), we have the following results. For any $\|\boldsymbol{x}_1\|$ and $\|\boldsymbol{x}_2\|$

such that $\|\boldsymbol{x}_1\| < \|\boldsymbol{x}_2\|$, $\pi_\alpha(\boldsymbol{x}_1)$ and $\pi_\alpha(\boldsymbol{x}_2)$ have dominated mass on the interval $\lambda \in (s_3, \infty)$, where $s_3 \stackrel{\text{def}}{=} \min\left(\|\boldsymbol{x}_1\|^2/m_n \log p_n, \|\boldsymbol{x}_2\|^2/m_n \log p_n\right)$. Then we have

$$
\begin{aligned}
l_n =& \max_{j \in \xi_*} \sup_{\|\boldsymbol{x}_1\|, \|\boldsymbol{x}_2\| \in \|\boldsymbol{\beta}_{*j}/\sigma_*\| \pm C_0\sqrt{m_n}\epsilon_n} \frac{\pi_\alpha(\boldsymbol{x}_1)}{\pi_\alpha(\boldsymbol{x}_2)} \\
\asymp& \max_{j \in \xi_*} \sup_{\|\boldsymbol{x}_1\|, \|\boldsymbol{x}_2\| \in \|\boldsymbol{\beta}_{*j}/\sigma_*\| \pm C_0\sqrt{m_n}\epsilon_n} \frac{\int_{s_3}^\infty \int_0^\infty \left(\frac{\phi_\lambda(\boldsymbol{x}_1)}{\phi_\lambda(\boldsymbol{x}_2)}\right) \phi_\lambda(\boldsymbol{x}_2) p(\lambda|\psi) p_\alpha(\psi) d\psi d\lambda}{\int_{s_3}^\infty \int_0^\infty \phi_\lambda(\boldsymbol{x}_2) p(\lambda|\psi) p_\alpha(\psi) d\psi d\lambda} \\
=& \max_{j \in \xi_*} \sup_{\|\boldsymbol{x}_1\|, \|\boldsymbol{x}_2\| \in \|\boldsymbol{\beta}_{*j}/\sigma_*\| \pm C_0\sqrt{m_n}\epsilon_n} \frac{\int_{s_3}^\infty \int_0^\infty e^{\frac{1}{2\lambda}(\|\boldsymbol{x}_2\|^2 - \|\boldsymbol{x}_1\|^2)} \phi_\lambda(\boldsymbol{x}_2) p(\lambda|\psi) p_\alpha(\psi) d\psi d\lambda}{\int_{s_3}^\infty \int_0^\infty \phi_\lambda(\boldsymbol{x}_2) \pi(\lambda|\psi) \pi_\alpha(\psi) d\psi d\lambda} \\
\leq& \max_{j \in \xi_*} \sup_{\|\boldsymbol{x}_1\|, \|\boldsymbol{x}_2\| \in \|\boldsymbol{\beta}_{*j}/\sigma_*\| \pm C_o\sqrt{m_n}\epsilon_n} \exp\left\{ \frac{1}{2s_3} \left(\|\boldsymbol{x}_2\|^2 - \|\boldsymbol{x}_1\|^2 \right) \right\} \\
\leq& \max_{j \in \xi_*} \exp\Big\{ 2m_n (\|\boldsymbol{\beta}_{*j}/\sigma_*\|/\sqrt{m_n} + C_0\epsilon_n) C_0 \epsilon_n / s_3 \Big\}.
\end{aligned}
$$

When $\|\boldsymbol{x}_1\| \geq \|\boldsymbol{x}_2\|$, then we have $l_n \leq 1$, which is obvious to prove that $s_* \log l_n \prec \log p_n$. Therefore, we only consider the situation when $\|\boldsymbol{x}_1\| < \|\boldsymbol{x}_2\|$. And we have $s_* \log l_n \leq C_6 s_* m_n E\epsilon_n / s_3$ for constant $C_6 > 0$. Therefore, given that $s_* m_n \epsilon_n \prec \log p_n$, we have $s_* \log l_n \prec \log p_n$. Thus prove the Lemma.

The three lemmas above specify the conditions on prior concentration properties. Lemma 2.4.3 and Lemma 2.4.4 establish that the prior probability that the $\mathcal{L}_2$-norm of each basis coefficient larger than a_n is arbitrarily small with the constraint: $a_n \leq \sqrt{d_0'/m_n}\epsilon_n/p_n$. Therefore, the shrinkage prior on $\boldsymbol{\beta}_j/\sigma$ has a dominated and steep peak around zero. Moreover, Lemma 2.4.5 controls the variation in the prior density of basis coefficients at points not too close to zero. The shrinkage prior density concentrates in a very small area around zero but also has heavy and sufficiently flat tails. Thus the prior mimics the composition of a continuous distribution and a point mass at zero.

Proof of Theorem 2.4.1 We follow the results of Theorem A.1 and Theorem A.2 in Song and Liang (2017), but the difference is that we have additional bias,

$$Y = \sum_{j=1}^{p} f_j(X_j) + \sigma\boldsymbol{\varepsilon} = B(X)\boldsymbol{\beta} + \sigma\boldsymbol{\delta} + \sigma\boldsymbol{\varepsilon}, \tag{2.38}$$

where each additive function is estimated by m_n-dimensional basis expansion. Therefore, Y includes both bias term and linear regression part analogous to the model in Song and Liang (2017). The linear coefficient $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \dots \boldsymbol{\beta}_{p_n})^T$ is a $p_n \times m_n$-dimensional basis coefficient vector where a multivariate DL prior will be imposed for each covariate. The true model is the additive non-parametric model $\sum_{j=1}^{p} f_{*j}(X_j) + \sigma_*\boldsymbol{\varepsilon}$, where $f_{*j} \neq 0$ are $\kappa-$smooth functions for $j \in \xi_*$ and $f_{*j} = 0$ are zero effects for $j \notin \xi_*$. In our proposed model, we estimate each function by basis expansion $B(X_j)\boldsymbol{\beta}_j$, which imposes an extra bias term $\boldsymbol{\delta}$ in (2.38) due to approximation error. The magnitude of $\boldsymbol{\delta}$ is bounded by a multiplier of $m_n^{-\kappa}$.

For $B_n = \{\text{At least } \tilde{p} \text{ of } \|\boldsymbol{\beta}_j/\sigma\| \text{ is larger than } a_n\}$ for $m_n\tilde{p} \asymp n\epsilon_n^2/\log p_n$ and $m_n\tilde{p_n} \prec n\epsilon_n^2$, $C_n = \{\|B(X)\boldsymbol{\beta} - \sum_{j=1}^{p} f_{*j}(X_j)\| \geq \sigma_*\epsilon_n\} \cup \{\sigma^2/\sigma_*^2 > (1+\epsilon_n)/(1-\epsilon_n)\} \cup \{\sigma^2/\sigma_*^2 < (1-\epsilon_n)/(1+\epsilon_n)\}$ and $A_n = B_n \cup C_n$, we first consider test function $\phi_n = \max\{\phi_n', \tilde{\phi}_n\}$,

$$\phi_n' = \bigvee_{\{\xi \supset \xi_*, |\xi| \leq \tilde{p}+s_*\}} 1\left\{\left|\frac{Y^T(I-H_\xi)Y}{(n-m_n|\xi|)\sigma_*^2} - 1\right| \geq c_0'\epsilon_n\right\}$$

$$\tilde{\phi}_n = \bigvee_{\{\xi \supset \xi_*, |\xi| \leq \tilde{p}+s_*\}} 1\left\{\|B(X_\xi)\left(B(X_\xi)^TB(X_\xi)\right)^{-1}B(X_\xi)^TY - \sum_{j\in\xi} f_{*j}(X_j)\| \geq \tilde{c}_0\sigma_*\sqrt{n}\epsilon_n/5\right\}$$

for sufficiently large constants c_0' and $\tilde{c}_0$, where $H_\xi = B(X_\xi)\left(B(X_\xi)^TB(X_\xi)\right)^{-1}B(X_\xi)^T$ is the hat matrix corresponding to the subgroup ξ, and $\sum_{j\in\xi} f_{*j}(X_j)$ is the summation of true non-parametric functions for covariates X_j within subgroup ξ. For any ξ that

satisfies $\xi \supset \xi_*$, $m_n|\xi| \leq m_n(\tilde{p}+s_*) \prec n\epsilon_n^2$,

$$
\begin{aligned}
&\mathrm{E}_{(f_*,\sigma_*^2)} 1\left\{\left|\frac{Y^T(I-H_\xi)Y}{(n-m_n|\xi|)\sigma_*^2}-1\right| \geq c_0'\epsilon_n\right\} \\
=&\mathrm{E}_{(f_*,\sigma_*^2)} 1\left\{\left|\frac{\boldsymbol{\varepsilon}^T(I-H_\xi)\boldsymbol{\varepsilon}}{n-m_n|\xi|}-1+2\frac{\boldsymbol{\varepsilon}^T(I-H_\xi)\boldsymbol{\delta}}{n-m_n|\xi|}+\frac{\boldsymbol{\delta}^T(I-H_\xi)\boldsymbol{\delta}}{n-m_n|\xi|}\right| \geq c_0'\epsilon_n\right\} \\
\leq&\mathrm{P}_{(f_*,\sigma_*^2)}\left(\left|\boldsymbol{\varepsilon}^T(I-H_\xi)\boldsymbol{\varepsilon}-(n-m_n|\xi|)\right|+2\left|\boldsymbol{\varepsilon}^T(I-H_\xi)\boldsymbol{\delta}\right|+\boldsymbol{\delta}^T(I-H_\xi)\boldsymbol{\delta} \geq (n-m_n|\xi|)c_0'\epsilon_n\right).
\end{aligned}
$$

Since the magnitude of basis expansion bias is in the order of $m_n^{-\kappa}$, we have

$$
\begin{aligned}
\boldsymbol{\delta}^T(I-H_\xi)\boldsymbol{\delta} \leq&\|\boldsymbol{\delta}\|^2 \asymp nm_n^{-2\kappa} \prec (n-m_n|\xi|)\epsilon_n \\
2\boldsymbol{\varepsilon}^T(I-H_\xi)\boldsymbol{\delta} \leq&2\|\boldsymbol{\delta}\|\cdot\|\boldsymbol{\varepsilon}\| \asymp 2\sqrt{n}m_n^{-\kappa}\|\boldsymbol{\varepsilon}\|.
\end{aligned}
$$

For normally distributed error $\boldsymbol{\varepsilon}$, given the property of chi-square distribution, we have

$$
\mathrm{P}_{(f_*,\sigma_*^2)}\left(2\left|\boldsymbol{\varepsilon}^T(I-H_\xi)\boldsymbol{\delta}\right| \geq (n-m_n|\xi|)c_0'\epsilon_n\right) \leq Pr\left(\|\boldsymbol{\varepsilon}\| \geq c_0''nm_n^{2\kappa}\epsilon_n^2\right) \prec \exp\{-c_0'''n\epsilon_n^2\}.
$$

Combining with the conclusion in (A.3) of Song and Liang (2017), we can prove that

$$
\mathrm{E}_{(f_*,\sigma_*^2)} 1\left\{\left|\frac{Y^T(I-H_\xi)Y}{(n-m_n|\xi|)\sigma_*^2}-1\right| \geq c_0'\epsilon_n\right\} \leq \exp\{-c_3n\epsilon_n^2\}. \tag{2.39}
$$

Next, $f_{*j}(X_j)$ can be approximated by B-spline basis expansion $B(X_j)\boldsymbol{\beta}_{*j}$ with bias term $\boldsymbol{\delta}$, where $\boldsymbol{\beta}_{*j}$ is the basis coefficients for the true non-parametric function. We have

$$
\begin{aligned}
\left\|B(X)\beta-\sum_{j=1}^{p_n}f_{*j}(X_j)\right\| =&\left\|B(X)(\boldsymbol{\beta}-\boldsymbol{\beta}_*)-\boldsymbol{\delta}\right\| \\
\leq&\left\|B(X_\xi)(\boldsymbol{\beta}_\xi-\boldsymbol{\beta}_{*\xi})\right\|+\left\|B(X_{\xi^c})\boldsymbol{\beta}_{\xi^c}\right\|+\left\|\boldsymbol{\delta}\right\|,
\end{aligned}
$$

where $\big\|B(X_\xi)(\boldsymbol{\beta}_\xi - \boldsymbol{\beta}_{*\xi})\big\| \le \sqrt{n}\big\|\boldsymbol{\beta}_\xi - \boldsymbol{\beta}_{*\xi}\big\|$, $\big\|B(X_{\xi^c})\boldsymbol{\beta}_{\xi^c}\big\| \prec \sqrt{n}a_n p_n \prec \sqrt{n}\epsilon_n$ and $\big\|\boldsymbol{\delta}\big\| \prec \sqrt{n}\epsilon_n$. By the condition that $m_n(\tilde{p}+s_*) \prec n\epsilon_n^2$, we have, for some constant c_3',

$$
\begin{aligned}
&\mathrm{E}_{(f_*,\sigma_*^2)} 1\Big\{\big\|B(X_\xi)\left(B(X_\xi)^T B(X_\xi)\right)^{-1} B(X_\xi)^T Y - \sum_{j\in\xi} f_{*j}(X_j)\big\| \ge \tilde{c}_0\sigma_*\sqrt{n}\epsilon_n/5\Big\} \\
=&\mathrm{E}_{(f_*,\sigma_*^2)} 1\Big\{\big\|B(X_\xi)(\boldsymbol{\beta}_\xi - \boldsymbol{\beta}_{*\xi}) - \big(\sum_{j\in\xi} f_{*j}(X_j) - B(X_\xi)\boldsymbol{\beta}_\xi\big)\big\| \ge \tilde{c}_0\sigma_*\sqrt{n}\epsilon_n/5\Big\} \\
\le&\mathrm{E}_{(f_*,\sigma_*^2)} 1\Big\{\big\|\left(B(X_\xi)^T B(X_\xi)\right)^{-1} B(X_\xi)^T Y - \boldsymbol{\beta}_\xi\big\| \ge \tilde{c}_0'\sigma_*\epsilon_n/5\Big\} \\
\le&\mathrm{E}_{(f_*,\sigma_*^2)} 1\Big\{\big\|B(X_\xi)\left(B(X_\xi)^T B(X_\xi)\right)^{-1} B(X_\xi)^T\boldsymbol{\varepsilon} + \left(B(X_\xi)^T B(X_\xi)\right)^{-1} B(X_\xi)^T\boldsymbol{\delta}\big\| \ge \tilde{c}_0\epsilon_n/5\Big\} \\
\le&\exp(-c_3' n\epsilon_n^2), \qquad (2.40)
\end{aligned}
$$

where the inequality is given by (A.4) of Song and Liang (2017) and the fact that $\|\boldsymbol{\delta}\| \asymp \sqrt{n}m_n^{-\kappa}$. Therefore, combining (2.39) and (2.40), we can prove that $\mathrm{E}_{(f_*,\sigma_*^2)}\phi_n \le \exp(-\tilde{c}_3 n\epsilon_n^2)$ for some fixed $\tilde{c}_3$, according to (A.5) in Song and Liang (2017).

For the additive model, the process to prove that $\sup_{(\boldsymbol{\beta},\sigma^2)\in C_n} \mathrm{E}_{\boldsymbol{\beta}}(1-\phi_n)$ is similar as the proof of Theorem A.1 in Song and Liang (2017), by replacing their covariate matrix X with $B(X)$. Therefore, we have inequalities

$$
\mathrm{E}_{(f_*,\sigma_*^2)}\phi_n \le \exp(-cn\epsilon_n^2) \qquad (2.41)
$$

$$
\sup_{(\boldsymbol{\beta},\sigma^2)\in C_n} \mathrm{E}_{(\boldsymbol{\beta},\sigma^2)}(1-\phi_n) \le \exp(-c' n\epsilon_n^2). \qquad (2.42)
$$

The next step is to show that

$$
\lim_n \mathrm{P}_{f_*,\sigma_*^2}\Big\{\frac{m(D_n)}{L_*(D_n)} \ge \exp(-c_4 n\epsilon_n^2)\Big\} > 1 - \exp\{-c_5 n\epsilon_n^2\}, \qquad (2.43)
$$

for any positive c_4, where $m(D_n)$ is the marginal likelihood of the data after integrating

out the parameters on their priors and $L_*(D_n)$ is the likelihood under the truth so that

$$\frac{m(D_n)}{L_*(D_n)} = \int \frac{(\sigma_*)^n \exp\{-\|Y - B(X)\boldsymbol{\beta}\|^2/2\sigma^2\}}{\sigma^n \exp\{-\|Y - \sum_{j=1}^{p_n} f_{*j}(X_j)\|^2/\sigma_*^2\}} \pi(\boldsymbol{\beta}, \sigma^2) d\boldsymbol{\beta} d\sigma^2.$$

From the proof of Theorem A.1 in Song and Liang (2017), we only need to show that, conditional on $0 \leq \sigma^2 - \sigma_*^2 \leq \eta' \epsilon_n^2$, for some constants c_5 and c_6,

$$\begin{aligned}
&\mathrm{P}_{f_*, \sigma_*^2}\left(\mathrm{P}_\alpha\left(\|Y - B(X)\boldsymbol{\beta}\|^2/2\sigma^2 < \left\|Y - \sum_{j=1}^{p_n} f_{*j}(X_j)\right\|^2/\sigma_*^2 + c_5 n\epsilon_n^2\right) \geq e^{-c_6 n\epsilon_n^2}\right) \\
\geq & 1 - \exp(-c_5 n\epsilon_n^2).
\end{aligned}$$

For our proposed model, each element in $B(X)$ is smaller than 1 and $\|\boldsymbol{\delta}\|^2 \asymp n m_n^{-2\kappa} \prec n\epsilon_n^2$. By the property of chi-square distribution and normal distribution, with probability larger than $1 - \exp\{-c_5 n\epsilon_n^2\}$, we have $\|\boldsymbol{\varepsilon}\|^2 \leq n(1 + c_5')$, $\|\boldsymbol{\varepsilon}^T B(X)\|_\infty \leq c_5' n\epsilon_n$ for constant c_5'. Then for model parameters $(\boldsymbol{\beta}, \sigma^2)$, we have

$$\begin{aligned}
&\{\|Y - B(X)\boldsymbol{\beta}\|^2/2\sigma^2 < \|Y - (f_{*1}(X_1) + \ldots + f_{*p_n}(X_{p_n}))\|^2/\sigma_*^2 + c_5 n\epsilon_n^2\} \\
=&\{\|\boldsymbol{\varepsilon}\sigma_*/\sigma + B(X)(\boldsymbol{\beta}_* - \boldsymbol{\beta})/\sigma + \boldsymbol{\delta}\sigma_*/\sigma\|^2 < \|\boldsymbol{\varepsilon}\|^2 + 2c_5 n\epsilon_n^2\} \\
\supset&\{\|B(X)(\boldsymbol{\beta}_* - \boldsymbol{\beta})/\sigma\| < \sqrt{\|\boldsymbol{\varepsilon}\|^2 + 2c_5 n\epsilon_n^2} - (\sigma_*/\sigma)\|\boldsymbol{\varepsilon}\| - (\sigma_*/\sigma)\|\boldsymbol{\delta}\|\} \\
\supset&\Big\{\sum_{j=1}^{p_n} \|(\boldsymbol{\beta}_{*j} - \boldsymbol{\beta}_j)/\sigma\| \leq 2\eta\sqrt{m_n}\epsilon_n\Big\} \\
\supset&\Big\{\|\boldsymbol{\beta}_j/\sigma\| \in \left[\|\boldsymbol{\beta}_{*j}/\sigma\| - \eta\sqrt{m_n}\epsilon_n/s_*, \|\boldsymbol{\beta}_{*j}/\sigma\| + \eta\sqrt{m_n}\epsilon_n/s_*\right] \text{ for all } j \in \xi_*\Big\} \\
&\cap \Big\{\sum_{j \notin \xi_*} \|\boldsymbol{\beta}_j/\sigma\| \leq \eta\sqrt{m_n}\epsilon_n\Big\}.
\end{aligned}$$

for some constant η. The second subset relation is given by Assumption $A_1(3)$ such that the eigenvalues of $B(X)^T B(X)$ are in the same order as n/m_n. By the results of Lemma

2.4.3 and Lemma 2.4.4,

$$
\begin{aligned}
&\mathrm{P}_\alpha\Big(\sum_{j\notin\xi_*}\big\|\boldsymbol{\beta}_j/\sigma\big\|\le\eta\sqrt{m_n}\epsilon_n\big|0\le\sigma^2-\sigma_*^2\le\eta'\epsilon_n^2\Big)\\
=&\int_{\sum_{j=1}^{p_n-s_*}\|\boldsymbol{\beta}_j\|\le\eta\sqrt{m_n}\epsilon_n}\prod_{j=1}^{p_n-s_*}\pi(\boldsymbol{\beta}_j)d\boldsymbol{\beta}_1\dots d\boldsymbol{\beta}_{p_n-s_*}\ge\exp\{-c_6'n\epsilon_n^2\}\\
&\mathrm{P}_\alpha\Big(\|\boldsymbol{\beta}_j/\sigma\|\in\big[\|\boldsymbol{\beta}_{*j}/\sigma\|-\eta\sqrt{m_n}\epsilon_n/s_*,\|\boldsymbol{\beta}_{*j}/\sigma\|+\eta\sqrt{m_n}\epsilon_n/s_*\big]\,|\{0\le\sigma^2-\sigma_*^2\le\eta'\epsilon_n^2\}\Big)\\
\ge&\eta\epsilon_n\cdot\inf_{\|\boldsymbol{\beta}\|\le\sqrt{m_n}E}\pi_\alpha(\boldsymbol{\beta})/s_*\ge\exp\{-c_6''n\epsilon_n^2\}.
\end{aligned}
$$

Thus verify the inequality in (2.43).

Finally, from the proof of Theorem A.1 in Song and Liang (2017), for B_n defined above, we have $\mathrm{P}_\alpha(B_n)<e^{-c_7n\epsilon_n^2}$ for some constants c_7, given that $\int_{\|\boldsymbol{\beta}\|>a_n}\pi_\alpha(\boldsymbol{\beta})d\boldsymbol{\beta}\le p_n^{-(1+\mu)}$.

Combining the results of $\mathrm{P}_\alpha(B_n)<e^{-c_7n\epsilon_n^2}$ with (2.41), (2.42) and (2.43), we have

$$
\begin{aligned}
&\mathrm{P}_{f_*,\sigma_*^2}\left(\pi\left(\Big\|B(X)\boldsymbol{\beta}-\sum_{j=1}^{p_n}f_{*j}(X_j)\Big\|\ge\sigma_*\sqrt{n}\epsilon_n\text{ or }|\sigma^2-\sigma_*^2|>c_4\epsilon_n\big|X,Y\right)\ge e^{-c_1n\epsilon_n^2}\right)\\
\le&e^{-c_2n\epsilon_n^2}.
\end{aligned}
$$

In conclusion, the results in Theorem 2.4.1 is proved.

Proof of Theorem 2.4.2 The proof of Theorem 2.4.2 follows the proof of Theorem A.3 in Song and Liang (2017). Note that in our proposed model, we have basis expansion matrix $B(X)$ instead of covariate matrix X, the coefficients vector $\boldsymbol{\beta}=(\boldsymbol{\beta}_1^T,\dots,\boldsymbol{\beta}_{p_n}^T)^T$ are shrunk within blocks $\boldsymbol{\beta}_1,\dots,\boldsymbol{\beta}_{p_n}$. Furthermore, our model has an additional bias term $\boldsymbol{\delta}$ due to B-spline basis expansion.

We first define the set $S_1=\{\|\boldsymbol{\beta}-\boldsymbol{\beta}_*\|\le c_1\epsilon_n,\ |\sigma^2-\sigma_*^2|\le c_2\epsilon_n\}$, and $\underline{\pi}(\boldsymbol{\beta}|\sigma^2)=$

$\inf_{(\boldsymbol{\beta},\sigma^2)\in S_1} \pi(\boldsymbol{\beta},\sigma^2)/\pi(\sigma^2)$, $\overline{\pi}(\boldsymbol{\beta}|\sigma^2) = \sup_{(\boldsymbol{\beta},\sigma^2)\in S_1} \pi(\boldsymbol{\beta},\sigma^2)/\pi(\sigma^2)$. The maximum of $\mathcal{L}_2$-norm for each group within coefficient vector is defined as $\|\boldsymbol{\beta}\|_{\max} \triangleq \max_j \|\boldsymbol{\beta}_j\|$. We can conclude by proof of Theorem A.3 in Song and Liang (2017),

$$\begin{aligned}\mathrm{P}_\alpha(\xi = \xi_*|X,Y) \gtrsim & \pi\left(\left\|\frac{\boldsymbol{\beta}_{(\xi_*)^c}}{\sigma}\right\|_{\max} \le a_n\right) \underline{\pi}(\boldsymbol{\beta}_{\xi_*}|\sigma^2)\sqrt{\det(B(X_{\xi_*})^T B(X_{\xi_*}))^{-1}}\left(\sqrt{2\pi}\right)^s \\ & \times \inf_{\|\boldsymbol{\beta}_{(\xi_*)^c}\|_{\max}\le a_n(\sigma_*+c_2\epsilon_n)} \frac{\Gamma(a_0+(n-s)/2)}{\left(\mathrm{SSE}(\boldsymbol{\beta}_{(\xi_*)^c})/2+b_0\right)^{a_0+(n-s)/2}}, \end{aligned} \tag{2.44}$$

where $\mathrm{SSE}(\boldsymbol{\beta}_{(\xi_*)^c})$ is the SSE when $\boldsymbol{\beta}_{(\xi_*)^c}$ is given and fixed.

Next, we let $\|\boldsymbol{\beta}\|_{\min} \triangleq \min \|\boldsymbol{\beta}_j\|$ be the minimum of $\mathcal{L}_2$-norms for each predictor within coefficient vector. Similarly, we can show that

$$\begin{aligned}\mathrm{P}_\alpha(\xi = \xi'|X,Y) \lesssim & \pi\left(\left\|\frac{\boldsymbol{\beta}_{\xi'\setminus\xi_*}}{\sigma}\right\|_{\min} > a_n, \left\|\frac{\boldsymbol{\beta}_{(\xi')^c}}{\sigma}\right\|_{\max} \le a_n\right) \overline{\pi}(\boldsymbol{\beta}_{\xi_*}|\sigma^2)\sqrt{\det(B(X_{\xi_*})^T B(X_{\xi_*}))^{-1}} \\ & \times \left(\sqrt{2\pi}\right)^s \sup_{\|\boldsymbol{\beta}_{(\xi')^c}\|_{\max}\le a_n(\sigma_*+c_2\epsilon_n)} \frac{\Gamma(a_0+(n-s)/2)}{\left(\mathrm{SSE}(\boldsymbol{\beta}_{(\xi')^c})/2+b_0\right)^{a_0+(n-s)/2}}. \end{aligned}$$

where $\mathrm{SSE}(\boldsymbol{\beta}_{(\xi')^c})$ is the SSE when $\boldsymbol{\beta}_{(\xi')^c}$ is given and fixed.

Due to the basis expansion bias $\boldsymbol{\delta}$ such that $\|\boldsymbol{\delta}\|^2 \asymp n m_n^{-2\kappa}$, we can the same derivation as (A.13) and (A.14) of Song and Liang (2017). With probability larger than $1-4p_n\cdot p_n^{-c_6'}$,

$$\begin{aligned}\mathrm{SSE}(\boldsymbol{\beta}_{(\xi_*)^c}) =& (Y - B(X_{(\xi_*)^c})\boldsymbol{\beta}_{(\xi_*)^c})^T(I-H_{\xi_*})(Y - B(X_{(\xi_*)^c})\boldsymbol{\beta}_{(\xi_*)^c}) \\ \le& \sigma_*^2\boldsymbol{\varepsilon}^T(I-H_{\xi_*})\boldsymbol{\varepsilon} + \sigma_*^2\boldsymbol{\delta}^T(I-H_{\xi_*})\boldsymbol{\delta} + \left\|B(X_{(\xi_*)^c})\boldsymbol{\beta}_{(\xi_*)^c}\right\|^2 \\ & + 2\sigma_*^2\sqrt{2c_6' n\log p_n}\sum_{j\in(\xi_*)^c}\|\boldsymbol{\beta}_j\|; \end{aligned} \tag{2.45}$$

$$\mathrm{SSE}(\boldsymbol{\beta}_{(\xi')^c}) \ge \sigma_*^2\boldsymbol{\varepsilon}^T(I-H_{\xi'})\boldsymbol{\varepsilon} + \sigma_*^2\boldsymbol{\delta}^T(I-H_{\xi'})\boldsymbol{\delta} - 2\sigma_*\sqrt{2c_6' n\log p_n}\sum_{j\in(\xi')^c}\|\boldsymbol{\beta}_j\|, \tag{2.46}$$

and with probability $1-p_n^{-c_6''}$ for any constant c_6'', we have

$$\boldsymbol{\varepsilon}^T(I-H_{\xi_*})\boldsymbol{\varepsilon}-\boldsymbol{\varepsilon}^T(I-H_{\xi'})\boldsymbol{\varepsilon}\leq c_7' m_n|\xi'\setminus\xi_*|\log p_n.$$

Since $\boldsymbol{\delta}^T(I-H_{\xi_*})\boldsymbol{\delta}-\boldsymbol{\delta}^T(I-H_{\xi'})\boldsymbol{\delta}\leq c_7'' m_n|\xi'\setminus\xi_*|nm_n^{-2\kappa}$, we can conclude that

$$\begin{aligned}&\frac{\sup_{\|\boldsymbol{\beta}_{(\xi_*)^c}\|_{\max}\leq a_n(\sigma_*+c_2\epsilon_n)}(\mathrm{SSE}(\boldsymbol{\beta}_{(\xi_*)^c})/2+b_0)^{a_0+(n-s)/2}}{\inf_{\|\boldsymbol{\beta}_{(\xi')^c}\|\max\leq a_n(\sigma_*+c_2\epsilon_n)}(\mathrm{SSE}(\boldsymbol{\beta}_{(\xi')^c})/2+b_0)^{a_0+(n-s)/2}}\\ \leq& e^{\{c_8' m_n|\xi'\setminus\xi_*|\log p_n+c_8'' m_n|\xi'\setminus\xi_*|nm_n^{-2\kappa}\}}.\end{aligned}\tag{2.47}$$

If $m_n\succ(\frac{n}{\log p_n})^{1/2\kappa}$, we have $nm_n^{-2\kappa}\prec\log p_n$, so that the right hand side of (2.47) is $c_8 m_n|\xi'\setminus\xi_*|\log p_n$. Given (2.34) in Lemma 2.4.5, $\overline{\pi}(\boldsymbol{\beta}_{\xi_*}|\sigma^2)/\underline{\pi}(\boldsymbol{\beta}_{\xi_*}|\sigma^2)\leq l_n^s$. We can proceed with

$$\begin{aligned}\frac{\mathrm{P}_\alpha(\xi=\xi'|X,Y)}{\mathrm{P}_\alpha(\xi=\xi_*|X,Y)}\leq& l_n^s\frac{\sigma_*+c_2\epsilon_n}{\sigma_*-c_2\epsilon_n}[p_n^{-(1+\mu)}/(1-p_n^{-(1+\mu)})]^{|\xi'\setminus\xi_*|}e^{\{c_8 m_n|\xi'\setminus\xi_*|\log p_n\}}\\ \leq& l_n^s\frac{\sigma_*+c_2\epsilon_n}{\sigma_*-c_2\epsilon_n}p_n^{-\mu' m_n|\xi'\setminus\xi_*|},\end{aligned}$$

with proper choice of the constants.

Therefore, with $\min_{j\in\xi_*}\|\boldsymbol{\beta}_j\|\succ\sqrt{m_n}\epsilon_n$, according to the proof of Theorem A.2 in Song and Liang (2017), we can prove that

$$\sum_{\xi'\supsetneq\xi_*}\mathrm{P}_\alpha\left(\xi=\xi'|X,Y\right)\leq\mathrm{P}_\alpha\left(\|\boldsymbol{\beta}-\boldsymbol{\beta}_*\|>\left|\min_{j\in\xi_*}\|\boldsymbol{\beta}_j\|-a_n\sigma\right|\middle|X,Y\right)\leq e^{-cn\epsilon_n^2}$$

In conclusion, we have $\mathrm{P}_{f_*,\sigma_*^2}\Big\{\mathrm{P}_\alpha\big(\xi(a_n)=\xi_*|X,Y\big)>1-p_n^{-\mu''}\Big\}>1-p_n^{-\mu'}$ for some constants $\mu',\mu''>0$.

2.5 Simulation Study

In the simulation study, we evaluate model performance in terms of both prediction and variable selection. We first consider the additive nonparametric regression model that only includes individual main effects as in Chapter 2.2.1, and then expand the regression model to include two-way interaction effects as in Chapter 2.2.2.

2.5.1 Simulation description

For the simulated data, the number of covariates is $p = 50$ and the sample size is fixed at either $n = 200$ or $n = 500$. For the covariate matrix X, we first sample X_j^*, $j = 1, \ldots, p$, from Gaussian distribution with $\mathrm{E}(X_j^*) = 0$, $\mathrm{Var}(X_j^*) = 1$ and $\mathrm{Cov}(X_j^*, X_k^*) = 0$ for mutually independent case or $\mathrm{Cov}(X_j^*, X_k^*) = 0.5^{|j-k|}$ for autoregressive case. Next, the simulated random vectors are rescaled onto the unit interval by $X_j = \frac{X_j^* - \min(X_j^*)}{\max(X_j^*) - \min(X_j^*)}$. Given the covariate matrix X, the response Y is generated from normal distribution with mean $f(X) \stackrel{\text{def}}{=} f_1(X_1) + f_2(X_2) + f_3(X_3) + f_4(X_4)$ and variance $\sigma^2 = 1.5$, where

$$f_1(x) = e^{1.1x^3} - 2 \qquad f_2(x) = 2x - 1,$$
$$f_3(x) = \sin(4\pi x), \qquad f_4(x) = \log\{(e^2 - 1)x + 1\} - 1.$$

The remaining $p - 4$ predictors have no effect on the response.

Under each scenario with different sample sizes ($n = 200$ or $n = 500$) and dependence structures (independent or autoregressive), we simulate 100 data sets. For each method implemented, we compare the prediction accuracy and variable selection. Specifically, the

prediction accuracy is evaluated by mean squared error (MSE) on a testing data:

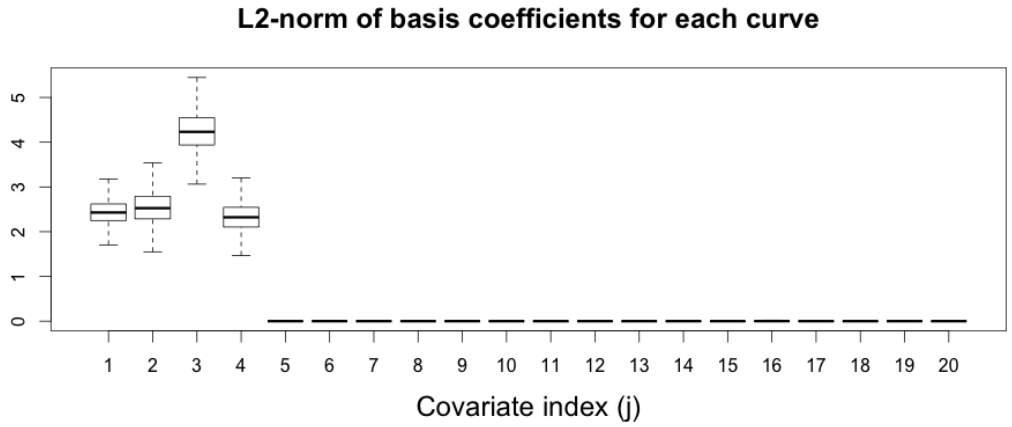
$$\text{MSE} = \frac{1}{500}\sum_{i=1}^{500}\left[f(x'_{i1},\ldots,x'_{ip}) - \hat{f}(x'_{i1},\ldots,x'_{ip})\right]^2, \tag{2.48}$$

where $f(\cdot)$ is the true mean function for covariate matrix, $\hat{f}$ is the estimated mean function, $X'_i = (x'_{i1},\ldots,x'_{ip})^T$ are a new data randomly sampled from the proposed covariate distribution, $i = 1,\ldots,500$. Note that X'_i's are not used for model fitting but for evaluating the model prediction performance. Therefore, we can compare the prediction performance of each method on future observations.

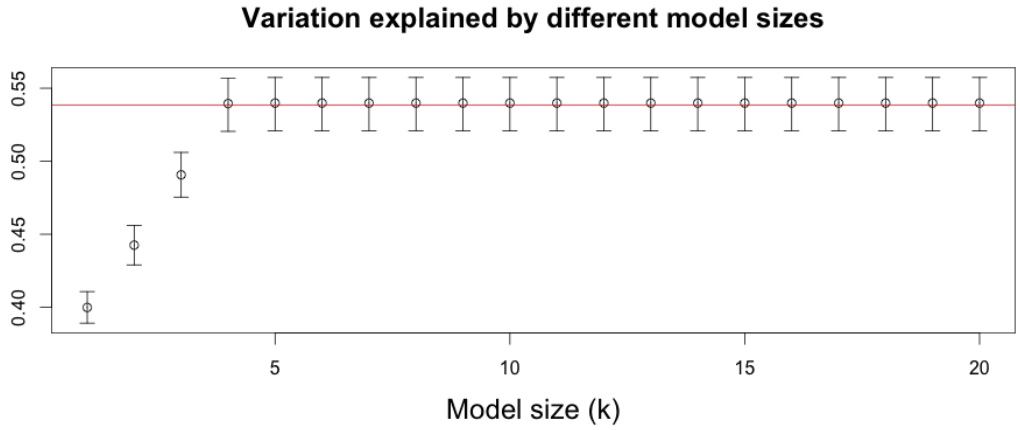
We also record the variable selection performance for correctly identifying the four significant covariates. We evaluate variable selection results in terms of the percentage of unimportant variables selected (False Positive) and the percentage of important variables excluded (False Negative), averaged over all the simulated data sets under each scenario. We also include the proportion of the simulated data sets in which the true model is selected (Truth). However, for the multivariate DL prior the factors $f_j(\cdot)$ will never equal zero exactly. Therefore, a thresholding policy is needed for variable selection. We implement a thresholding technique based on the idea that choosing a subset of predictors so that the corresponding prediction deterioration in terms of "variation-explained" can be tolerated (Hahn and Carvalho, 2015). Given the posterior samples of coefficients $\boldsymbol{\beta}_j$, the posterior samples of the "variation-explained" values for are calculated as

$$V_{(k)} = \frac{\|\sum_{j=1}^{p} B(X_j)\boldsymbol{\beta}_j\|^2}{\|\sum_{j=1}^{p} B(X_j)\boldsymbol{\beta}_j\|^2 + n\sigma^2 + \|\sum_{j=1}^{p} B(X_j)\boldsymbol{\beta}_j - \sum_{j=1}^{p} B(X_j)\boldsymbol{\beta}_j^{(k)}\|^2}, \tag{2.49}$$

where $\boldsymbol{\beta}_j^{(k)} = 0$ if $\|\boldsymbol{\beta}_j/\sigma\|$ is among the k smallest terms and $\boldsymbol{\beta}_j^{(k)} = \boldsymbol{\beta}_j$ otherwise. Figure 1(a) shows the box-plot of posterior samples of $\|\boldsymbol{\beta}_j\|$, $j = 1,\ldots,p$. Figure 1(b) shows the



(a)



(b)

Figure 2.1: (a) Box-plot of the posterior distribution of $\mathcal{L}_2$-norm $\|\boldsymbol{\beta}_j/\sigma\|$ for a single simulated data set; (b) Variation-explained plot at different model sizes for a single data sets (the horizontal red line is the full model "variation - explained" measurement). Note that only the first 20 model sizes are shown in the Figure.

posterior median of $V_{(k)}$ with 80% intervals at different model sizes $k = 1, \ldots, 20$. The shrinkage level for variable selection will be determined by the smallest k so that the 80% interval of V_k includes the posterior mean of full model V_{50} (red line in Figure 1(b)). We only include the model sizes less than 20 as model size larger than that will not affect

Table 2.1: Summary of the main-effect-only simulation study. Methods are compared in terms of mean squared errors ("MSE"), False Positive ("FP"), False Negative ("FN") and True model ("True") with their standard errors ("SE") under independent and autoregressive covariate covariance in parentheses. All values except the MSE's are given in percentages (%).

$n = 200$		MSE (SE)	FP (SE)	FN (SE)	True (SE)
Independent	DL(Oracle)	1.85(0.13)	0.00(0.00)	0.00(0.00)	100(0)
	DL(VarExp)	1.85(0.13)	0.88(0.22)	0.00(0.00)	86(3)
	ABayes	-	0.00(0.00)	15.25(1.91)	56(5)
	BSS-ANOVA	2.86(0.20)	4.88(0.58)	0.00(0.00)	47(5)
	COSSO	2.90(0.18)	2.94(0.54)	7.50(1.26)	46(5)
	MARS	2.31(0.17)	1.50(0.28)	1.00(0.49)	73(4)
AR(1)	DL(Oracle)	2.36(0.11)	4.81(0.44)	19.25(1.77)	39(5)
	DL(VarExp)	2.36(0.11)	2.50(0.38)	23.50(1.94)	25(4)
	ABayes	-	0.00(0.00)	40.25(1.77)	5(2)
	BSS-ANOVA	3.51(0.15)	3.69(0.47)	20.21(1.65)	17(4)
	COSSO	3.83(0.19)	4.87(0.87)	26.50(1.94)	10(3)
	MARS	3.90(0.14)	0.62(0.21)	38.00(1.90)	8(3)
$n = 500$		MSE (SE)	FP (SE)	FN (SE)	True (SE)
Independent	DL(Oracle)	1.52(0.09)	0.00(0.00)	0.00(0.00)	100(0)
	DL(VarExp)	1.52(0.09)	0.25(0.12)	0.00(0.00)	96(2)
	ABayes	-	0.00(0.00)	0.00(0.00)	100(0)
	BSS-ANOVA	2.17(0.10)	2.06(0.39)	0.00(0.00)	74(3)
	COSSO	2.12(0.11)	2.81(0.51)	0.25(0.25)	70(5)
	MARS	1.95(0.08)	0.44(0.18)	0.00(0.00)	94(2)
AR(1)	DL(Oracle)	2.06(0.14)	1.94(0.30)	7.75(1.21)	70(5)
	DL(VarExp)	2.06(0.14)	0.12(0.09)	13.25(1.40)	59(5)
	ABayes	-	0.00(0.00)	10.25(1.38)	60(5)
	BSS-ANOVA	2.89(0.13)	1.94(0.22)	2.00(0.68)	51(5)
	COSSO	3.01(0.10)	3.62(0.49)	7.75(1.21)	36(5)
	MARS	2.93(0.11)	0.75(0.22)	8.00(1.37)	55(5)

the value of "variation-explained".

In order to demonstrate the advantages of our proposed method over other competitors with similar model assumptions, we fit the additive nonparametric model with multi-

variate DL prior (DL method) to each data set and compare with four competing methods assuming additive nonparametric functions: approximate Bayesian method ("ABayes") by Curtis et al. (2014); the Bayesian smoothing splines ANOVA ("BSS-ANOVA") by Reich et al. (2009); component selection and smoothing operator ("COSSO") by Lin and Zhang (2006) and multivariate adaptive regression splines model ("MARS") by Friedman (1991). We also implement the "Oracle" selection in DL method which selects the four covariates with largest posterior medians of $\|\boldsymbol{\beta}_j/\sigma\|$. With the correct information on the number of significant covariates, this DL oracle method will demonstrate the ability of the general DL method to rank important covariates and formally select significant ones. For DL model and BSS-ANOVA where MCMC sampling is implemented, the total sampling iteration is $15,000$ with $5,000$ burn-in steps.

2.5.2 Simulation results: main-effect-only model

Table 2.1 summarizes the simulation results for the regression model with individual main effect functions. As for prediction performance, our proposed DL method is better than competing methods, even though all the methods make similar assumptions about the regression model. The DL method achieves smallest MSE under all the scenarios, while MARS has good prediction accuracy under the independent case. For the other methods such as BSS-ANOVA and COSSO, their performances are similar with BSS-ANOVA being slightly better than COSSO. The ABayes prediction results are not presented because the code we obtained for ABayes does not provide prediction on response variable.

As expected, the DL Oracle method performs the best under all scenarios for variable selection. The DL method with data-driven threshold, DL(VarExp), is slightly worse than DL(Oracle), but still outperforms other methods. For the thresholding policy defined in

(2.49), it might have uncertainty in determining the number of important covariates, but when the number is fixed like in the Oracle method, DL method will correctly rank the covariates so that higher variable selection accuracy will be achieved. For the others, ABayes is too conservative in choosing subset of covariates with false positive equal to zero for every case and thus large false negative proportion. In the dependent case, ABayes has perfect selection for the larger sample size, but the true model proportions drop considerably when the sample size is less. BSS-ANOVA and COSSO have similar variable selection performance. Their performance for independent data is worse than the competitors and the computation time for BSS-ANOVA is more than three times that of the DL prior method. MARS, on the other hand, performs poorly for correlated data as in the AR(1) case. Overall, the DL method outperforms the competitors, especially in more difficult setting with small sample size and correlated exposures.

2.5.3 Model with main and interaction effects

We also conduct simulation for additive models with both main effects and interactions. Since adding interactions will increase the dimensionality significantly, we reduce the number of covariates to $p = 10$ and only investigate the independent case. Therefore, there are 10 main effects and 45 interaction effects. The response variable Y is simulated from the model

$$Y = f_1(X_1) + f_2(X_3) + f_3(X_1X_3) + \varepsilon,$$

where ϵ is a random number generated from standard normal distribution and functions f_1, f_2 and f_3 are the same as in Chapter 2.5.2. As in the simulation for the main-effect-only model, we assume the sample size is $n = 200$ or $n = 500$.

Fitting the nonparametric regression model in (2.6), we approximate the main effect

function $f_j(X_j)$ with B-spline basis functions of order $m = 10$ and the two-way interaction functions $f_{kl}(X_k, X_l)$ with the outer product of basis terms of order $m^* = 5$ for each covariate. We implement the DL method as described in Chapter 2.2.2 and select the important main and interaction effects using the "variation-explained" criterion, similar to the main-effect-only model. The DL method is compared with ABayes, BSS-ANOVA, COSSO and MARS methods on both prediction performance and variable selection. The accuracy of model prediction is evaluated by computing mean squared error (MSE) for a newly generated covariate matrix, while variable selection is determined by False Negative, False Positive and the percentages of correctly identifying the main effects, interactions and the complete model. For MARS, we use the function `polymars`$(\cdot)$ in the R package `polspline`. In ABayes and COSSO where interactions are not considered in the available code, the interaction terms are represented as the product of two covariates, that is, we define 45 new covariates $X_l \cdot X_k$ and then use the main effects model with 55 additive predictors for both main effects and interaction terms. The simulation results are summarized in Table 2.2.

From Table 2.2, the simulation results show that our proposed method improves the prediction on the newly-generated data set compared to other nonparametric models. In particular, by correctly addressing the joint effect through decomposition of main effect functions and interaction effect functions, our method improves the prediction accuracy by more than 50% in MSE. Furthermore, the inclusion of shrinkage prior on basis expansion coefficients addresses the model uncertainty and improve the performance in identifying the correct sub-model. The $n = 200$ scenario is challenging for all methods, especially for selecting the correct model that includes both nonzero main effects and nonzero interactions. The DL method has the highest true model proportion among all methods. When $n = 500$, the DL method successfully selects the true interaction effects

Table 2.2: Summary of simulation study with main and interaction effects. Methods are compared in terms of mean squared errors ("MSE"), False Positive ("FP"), False Negative ("FN") and correct selection of main effects, interactions, as well as complete model with their standard errors ("SE") in parentheses. All values except the MSE's are given in percentages (%).

				Correct selection (SE)		
$n = 200$	MSE (SE)	FP (SE)	FN (SE)	Main	Interaction	Model
DL(Oracle)	1.21(0.08)	7.17(0.39)	28.67(1.57)	19(4)	38(5)	19(4)
DL(VarExp)	1.21(0.08)	0.08(0.08)	41.33(1.84)	46(5)	38(5)	15(4)
ABayes	-	0.17(0.12)	49.00(1.86)	3(2)	97(2)	2(1)
BSS-ANOVA	2.33(0.12)	0.17(0.12)	37.67(1.81)	53(5)	30(5)	9(3)
COSSO	2.71(0.13)	8.58(0.67)	41.67(2.70)	66(5)	11(3)	10(3)
MARS	2.39(0.16)	0.25(0.14)	38.33(1.29)	81(4)	1(1)	1(1)
				Correct selection (SE)		
$n = 500$	MSE (SE)	FP (SE)	FN (SE)	Main	Interaction	Model
DL(Oracle)	1.19(0.09)	0.58(0.21)	2.33(0.85)	93(3)	93(3)	93(3)
DL(VarExp)	1.19(0.09)	0.00(0.00)	2.33(0.85)	93(3)	100(0)	93(3)
ABayes	-	0.42(0.18)	16.67(1.87)	53(5)	95(2)	50(5)
BSS-ANOVA	2.08(0.11)	0.08(0.08)	2.33(0.85)	92(3)	90(1)	80(4)
COSSO	2.51(0.12)	3.08(0.51)	7.00(1.73)	94(2)	68(5)	68(5)
MARS	2.30(0.15)	0.17(0.12)	20.67(1.63)	98(1)	38(5)	37(5)

for all data sets. The competing methods perform worse than the DL method. Even for ABayes and COSSO are in favor of as the true format of the interaction term is specified, they cannot outperform the DL method. The second best model is the BSS-ANOVA, but this requires 5 times more computing time than the DL method.

2.6 Analysis of the AHS Neurobehavioral Dataset

2.6.1 Description of the neurobehavioral data

We demonstrate our method using data from a neurobehavioral (NB) substudy of the Agricultural Health Study (AHS; `http://aghealth.nih.gov/`). The goal of the NB

study (data version number: AHS44436) is to examine the association between pesticide exposures and neurobehavioral function of the central nervous system (CNS). From 2006 to 2008, $n = 701$ male farmers from Iowa or North Carolina took neurobehavioral tests. There are 12 response variables, among which $N = 8$ of them are for CNS tests that assess memory, motor speed, sustained attention, verbal learning and visual scanning and processing. For this implementation, we will focus on the analysis of these 8 continuous CNS response variables. We use the lifetime-specific pesticide use information for $p = 20$ pesticides from AHS questionnaires and interviews as exposure variables. Each exposure covariate is quantified as the days of applications of each pesticide over the participant's lifetime. We also include $q = 6$ confounding variables Z for age (years), testing site (1 if North Carolina, 0 if Iowa), farm size (acres), smoking status (packs per year), drinking status (drinks per year) and highest level of education (years).

The B-spline basis expansion requires $X_{ij} \in [0, 1]$ so we apply a rank transformation. For example, $X_{ij} = x$ means that subject i applied pesticide j more days than $100x\%$ of the study participants. This transformation makes the covariates uniformly distributed over unit interval, while still allowing for a wide range of regression relationships between covariates X and response Y via the additive nonparametric function. All response measurements are standardized to have mean zero and variance 1, and some response variables (continuous performance test, digit symbol latency, sequence A and sequence B latencies) are multiplied by -1 so that higher values indicated better performance.

2.6.2 Multivariate extension

Rather than analyzing the $N = 8$ neurobehavioral responses individually, we extend the model to account for multiple health responses. This multivariate analysis is preferred be-

cause we are more interested in the effects on overall neurobehavioral performance rather than the individual tests. Furthermore, borrowing strength across neurobehavioral test measurements should improve power for identifying important exposures and estimating their exposure-response curves.

For CNS response variables, we model the confounding variable age (Z_1) as having a nonparametric effect on health response but the other confounding variables ($Z_2, \ldots, Z_q$) as linear effects. The additive nonparametric model for response variable Y_b, $b = 1, \ldots, N$, on confounders Z and pesticide exposures X is

$$Y_b = \mu + g_{1,b}(Z_1) + \sum_{l=2}^{q} Z_l \gamma_{l,b} + f_b(X_1, \ldots, X_p) + \varepsilon \tag{2.50}$$

where μ is the intercept term and the ε_b is normally distributed with mean zero and variance σ^2. For the nonparametric function in confounding variable, we use B-spline basis expansion $g_{1,b}(Z_1) \approx \sum_{r=1}^{m} \gamma_{1r,b} B_r(Z_1)$. The joint effect function on the bth health reponse, $f_b(X_1, \ldots, X_p)$, is decomposed into main effect and interaction effect functions and approximated through basis expansion as in (2.6) and (2.7).

Next we propose Bayesian hierarchical model to build the connection on all CNS health responses. The model coefficients for each response variable share the common prior distribution with a global mean across $b = 1, \ldots, N$. Therefore, for each covariate index j, $\boldsymbol{\beta}_j^b \overset{ind}{\sim} \mathrm{N}_m(\boldsymbol{\mu}_j, \lambda_j \sigma^2 \mathrm{I}_m)$, where the normal mean $\boldsymbol{\mu}_j$ is treated as the basis coefficients for the nonparametric effect of covaraite X_j on the overall neurobehaviroal system (overall effect curve). The same multivaraite extension is applied on the interation effect functions so that $\boldsymbol{\mu}_{kl}$ determines the joint effect of X_l and X_k on overall neurobehaviroal functions. In the confounding effects, similar method is implemented. We assume $\boldsymbol{\gamma}_1^b \sim \mathrm{N}_m(\boldsymbol{\nu}_1, \sigma^2 \mathrm{I}_m)$ and $\gamma_l^b \overset{ind}{\sim} \mathrm{N}(\nu_l, \sigma^2)$ for $l = 2, \ldots, q$. For variable selection, we imple-

ment Bayesian hierarchical model with DL prior on both the response-specific curves and overall effect curves. However, we do not use shrinkage priors for the mean confounder coefficients ν_{1r} and ν_l to conservatively account for their effects in our study. The following prior distribution structure is assumed for the model parameters in pesticide main effects, while similar structure can be followed for the parameters in interaction effect functions:

$$\begin{aligned}
&\boldsymbol{\beta}_j^b|\boldsymbol{\mu}_j, \lambda_j, \sigma^2 \overset{ind}{\sim} \mathrm{N}_m(\boldsymbol{\mu}_j, \lambda_j\sigma^2\mathrm{I}_m), \quad \boldsymbol{\mu}_j|\omega_j \overset{ind}{\sim} \mathrm{N}_m(0, \omega_j\mathrm{I}_m),\\
&\lambda_j|\phi_j', \tau' \overset{ind}{\sim} \mathrm{Exp}(\phi_j'\tau'), \quad \boldsymbol{\phi'} \sim \mathrm{Dirichlet}(\alpha, \ldots, \alpha), \quad \tau' \sim \mathrm{Gamma}(p\alpha, 2),\\
&\omega_j|\phi_j, \tau \overset{ind}{\sim} \mathrm{Exp}(\phi_j\tau), \quad \boldsymbol{\phi} \sim \mathrm{Dirichlet}(\alpha, \ldots, \alpha), \quad \tau \sim \mathrm{Gamma}(p\alpha, 2),\\
&\sigma^2 \sim \mathrm{InvGamma}(0.01, 0.01), \quad \nu_{1r}, \nu_l \overset{iid}{\sim} \mathrm{N}(0, 10^2).
\end{aligned}$$

This hierarchical model centers all N exposure-response curves around the overall effect curve by shrinking the response-specific coefficients to μ_{jr}, and thus the response-specific curves are shrunk towards the average curve for the overall effects of exposures: $\bar{f}_j(X_j) \approx \sum_{r=1}^m \mu_{jr}B_r(X_j)$. Small λ_j shrinks all N curves towards $\bar{f}_j(X_j)$; large λ_j allows for variations among the response-specific curves. As for the overall effects curve that reflects the average main effect across health responses, small ω_j shrinks the average main effect function for exposure X_j towards zero so that jth pesticide does not influence the overall neurobehavioral system significantly; large ω_j, however, allows for a significant association between X_j and neurobehavioral system. If one pesticide is not associated with any of the response variables such that the overall effects are negligible, both λ_j and ω_j are small and all curves will be shrunk toward zero.

2.6.3 Neurobehavioral data analysis

In this section, we show the data analysis results for CNS measurements. We use 5-fold cross validation to select the number of basis functions (we compare $m_1 \in \{5, 10, 15, 20\}$) and the Dirichlet parameter (we compare $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$). The best prediction performance is achieved (MSE=0.908) with $m_1 = 10$ and $\alpha = 0.5$. We also find that this nonparametric additive model outperforms linear regression model using least squares (MSE=1.023), generalized additive model by restricted maximum likelihood method (MSE=0.993), MARS (MSE=0.973) and COSSO (MSE=0.965) when the responses are analyzed separately.

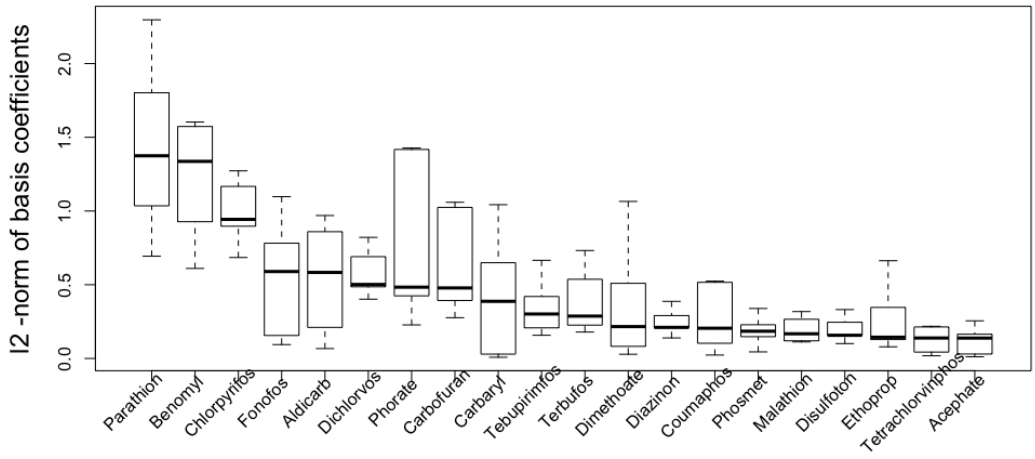


Figure 2.2: Box-plot for the $\mathcal{L}_2$-norm of mean curve basis coefficients $\|\boldsymbol{\mu}_j/\sigma\|$.

Figure 2.2 shows the posterior samples for $\mathcal{L}_2$-norm of mean curve coefficients, $\|\boldsymbol{\mu}_j/\sigma\|$. Using the variation explained measurements, the size of the final model is chosen by the smallest model for which the 80% credible interval includes the median variation

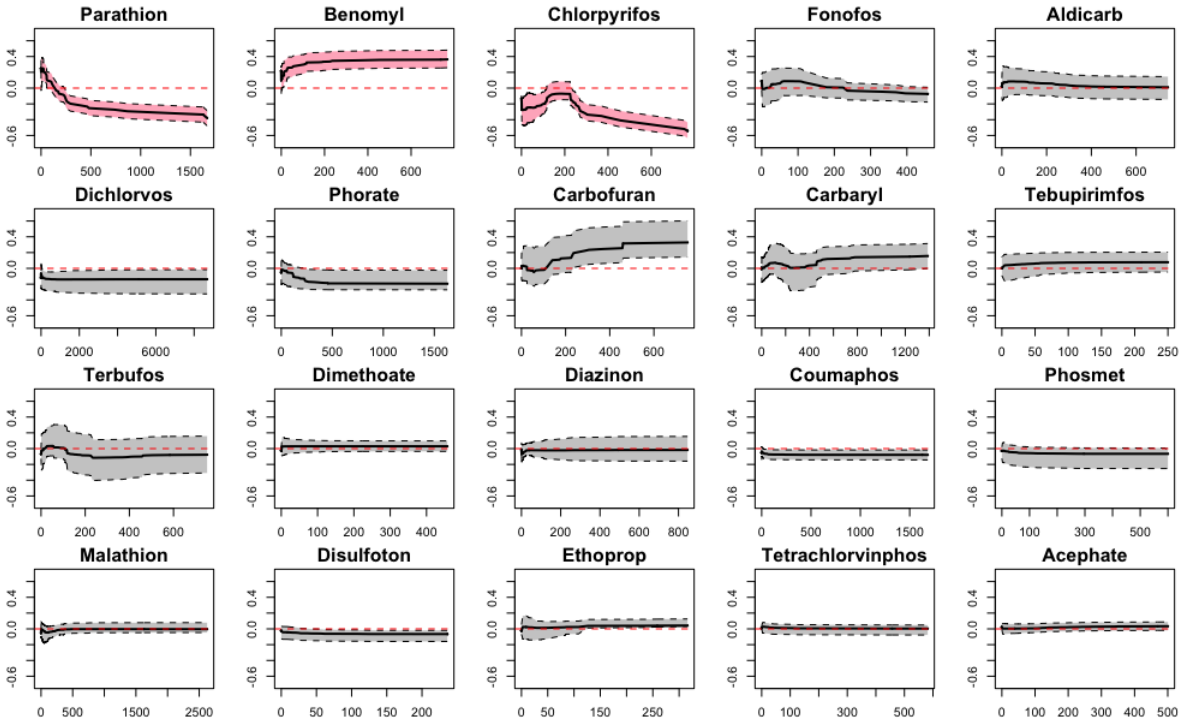


Figure 2.3: Average (over Central Nervous System tests) exposure-response curves $\bar{f}_j(X)$ for each pesticide in the Agricultural Health Study. The x-axis is the cumulative number of pesticide applications. The solid lines are the posterior means and the dashed lines are point-wise 95% credible intervals. The plots in red are the pesticide covariates selected by DL model.

explained value of full size model. Three pesticides are selected: Parathion, Benomyl and Chlorpyrifos. Note that for the additive nonparametric regression model that includes both main effects and interactions, the thresholding policy selects three main effects of pesticides and excludes the interaction effects. Therefore, we only show the results for the coefficients of individual main effects as the interaction effects are negligible.

Figure 2.3 plots the average exposure-response functions $\bar{f}(X_j)$ for each pesticide main effect. Each individual curve is a function of cumulative number of pesticide applications (i.e., the original measurement before the rank transformation). The mean curves plateau for large exposures because these exposures are rare in the sampled subjects. Among the

selected pesticides, Parathion and Chlorpyrifos show decreasing pattern in the mean curves, which implies these pesticides have negative overall effects on central nervous systems. The mean curve for Benomyl shows a positive effects, but that might because of the collinearity with other pesticides.

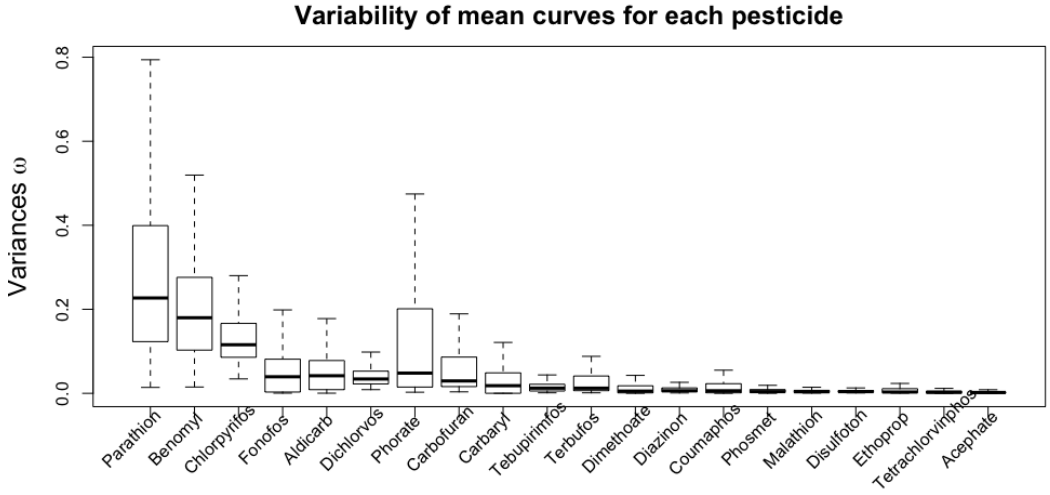


Figure 2.4: Posterior distribution of normal variance ω_j for average coefficients across CNS responses.

The variance of curves across pesticide exposure j, λ_j, and the variance of the average curve for pesticide j, ω_j, illustrate the overall importance of each covariate. The posterior of all λ_j's are concentrated very close to zero, which shows that there are no significant difference among the response-specific main effect functions for each CNS response measurement. Therefore, the average exposure-response curves $\bar{f}_j(X)$ are sufficient to delineate the associations between pesticides and the overall NB test results. For the posterior samples of ω_j, we present the box-plot in Figure 2.4. Larger values of variances ω_j indicate more variability away from the center for basis coefficients across

CNS responses.

In conclusion, we detect significant associations between three pesticide chemicals and CNS overall functions using the nonparametric model with DL prior, while the other parametric or nonparametric cannot find associations from the data. Compared with the simple linear regression analysis results in Starks et al. (2012a), our proposed method chooses a sparse model with some pesticides that affect the NB tests significantly and demonstrates the nonlinear effects on overall performance of central nervous systems. Through integrating the CNS response variables, we may have greater utility for those outcome measurements since individual NB tests might fail to capture the overall impact.

2.7 Conclusion

In this project, we propose a nonparametric regression model with additivity assumption on main effect and interaction effect functions for the motivation problem that aims to study the health effects of multiple pesticide mixture. The nonparametric functions in the additive model are approximated by B-spline basis expansion with multivariate extension of shrinkage prior for Bayesian variable selection on individual functions. Asymptotically, we show posterior consistency of model prediction and variable selection for the additive nonparametric model. The theoretical results rely on the consistency theorems for shrinkage priors in linear regression model but take into consideration both the estimation bias introduced by B-spline basis approximation and the properties of group shrinkage on basis coefficients of each covariate.

We apply the nonparametric additive regression model to the neurobehavioral (NB) data from AHS and investigate the potential associations between NB system and pesticide application frequency. With a robust assumption on the regression mean function and

computationally efficient variable selection technique, the proposed method achieves good prediction accuracy and identifies pesticides that affecting the NB responses significantly. The model extension into multivariate response variables improves the interpretability on demonstrating the overall health effects of pesticide mixture and increases the powers for valid statistical inference.

Conversely, the proposed method deals with continuous response measurements only. As there are binary or count neurobehavioral responses in the data sets, it will be interesting to consider the extension of additive nonparametric regression model into categorical data analysis. Besides that, extensions to address multiple time or spatial measurements are also desirable.

Chapter 3

Contraction Properties of Shrinkage Priors in Logistic Regression Model

3.1 Introduction

For statistical models such as linear regression, logistic regression or normal mean models, high-dimensional data analysis is challenging due to the its computational burden and the inherent limitations of any procedure due to the high complexity of the model. Bayesian variable selection procedures are more informative than penalization methods because they automatically address the model selection uncertainty. There is an extensive literature on Bayesian variable selection methods in recent years. Since the value zero of a regression coefficient is equivalent to having the corresponding predictor dropping out of the model, priors are designed to give special emphasis to the value zero. For example, a point-mass prior combines a probability mass at zero and a non-zero continuous distribution. Computation is generally carried out by Reversible Jump Markov Chain Monte Carlo methods (Green, 1995). A spike-and-slab prior (Ishwaran and Rao,

2005) is typically a mixture of two normal distributions with one highly concentrated at around zero. The Stochastic Search Variable Selection (SSVS) method (George and McCulloch, 1993) can be used to compute the posterior distribution corresponding to a spike-and-slab prior. Due to high computational costs, these methods are not scalable to very high dimensional situations commonly arising in recent applications. The Bayesian LASSO (Park and Casella, 2008) uses a double exponential prior distribution on the coefficients. Clearly, the posterior mode of the Bayesian LASSO is the LASSO estimator in linear regression model, and the posterior can be computed by a simple Gibbs sampling procedure. However, the Bayesian LASSO does not make the posterior distribution concentrate near the true value in large samples (Castillo et al., 2015), although its mode the LASSO estimator has good estimation and variable selection properties under appropriate conditions.

Earliest posterior concentration results in models of dimension increasing to infinity with the sample size are provided by Ghosal (1997, 1999, 2000), respectively for generalized linear models, regression models and exponential families. In his results, no sparsity conditions are assumed on the truth and posterior concentration at the truth and asymptotic normality of the posterior are established, provided that the growth of the dimension is sufficiently slow compared with the sample size. In more high dimensional settings, assuming sparsity conditions, Castillo and van der Vaart (2012) and Belitser and Nurushev (2015) established posterior concentration and variable selection properties for certain point-mass priors in the many normal means model. The latter paper also established coverage of Bayesian credible sets. Posterior concentration and variable selection in high dimensional linear models are obtained by Castillo et al. (2015) and Martin et al. (2017) for certain point-mass priors. Recent theoretical breakthroughs establish concentration properties of posterior distributions of continuous shrinkage priors. Armagan et

al. (2013b) showed posterior consistency in linear regression model with shrinkage priors for low-dimensional setting where the number of covariates does not exceed the number of observations. By direct calculations, van der Pas et al. (2014) showed that the posterior based on the horseshoe prior concentrates at the optimal rate for the many normal-mean problem. Bhattacharya et al. (2015) obtained an analogous result using the DL prior. Under certain restrictions on the hyper-parameter of DL prior, the posterior contraction property is applied to the coefficients of high-dimension linear regression model. Song and Liang (2017) considered a general class of continuous shrinkage priors and obtained posterior contraction rate in linear regression models depending on concentration and tail properties of the density of the continuous shrinkage prior. Essentially their conclusion may be summarized as the following statement: under appropriate conditions, the posterior contraction rates and variable selection ability of continuous shrinkage priors are close to those of the point-mass priors in high dimensional linear regression models.

Only a few papers in the literature address convergence results on Bayesian variable selection in regression beyond the linear regression model. Shen and Ghosal (2014) considered estimating conditional density in a high dimensional setting using tensor products of B-splines where the true conditional density is assumed to be a function of only a few predictors. They showed that the oracle contraction rate can be matched adaptively up to a logarithmic factor. A similar result by using Dirichlet process mixtures was obtained by Norets and Pati (2016). Yang and Tokdar (2016) obtained optimal posterior contraction rate in the setting of high dimensional additive regression models using a Gaussian process prior. The logistic regression model is another important model for practical applications and often comes with a large of predictors, among which the important ones need to be selected for precise estimation and sensible interpretation. Atchadé (2017), as a special case of his more general results on quasi-posterior distributions in possibly

nonlinear models, derived posterior contraction properties in a logistic regression model using a point-mass prior. He showed that the posterior contracts at the rate $\sqrt{s_*(\log p)/n}$, where s_* is the true number of active predictors. However the properties of posterior distributions under shrinkage priors in logistic regression model has not been studied in the literature. Borrowing a few techniques from Song and Liang (2017), we extend the results of Atchadé (2017) on logistic regression to continuous shrinkage priors. Our finding can be summarized to the statement that, provided that a prior has a sufficient concentration near zero and has sufficiently thick tails, posterior concentrates near sparse vectors and selects the correct predictors like a point-mass prior. The theoretical result is amply supported by simulations.

The remainder of this chapter is constructed as follows. Chapter 3.2 presents the logistic regression model and the assumptions on shrinkage prior densities. The main results of posterior contraction of parameter estimated is shown in Theorem 3.2.1. The proof details are also demonstrated. Several exampled of shrinkage priors and spike-and-slab prior are introduced in Chapter 3.3 to demonstrate the conditions for posterior contraction. For different types Bayesian variable selection technique, Chapter 3.4 evaluates the performance in recovering the logistic coefficients, identifying non-zero subsets of covariates and predicting on test design matrix.

3.2 Contraction Results in Logistic Regression

3.2.1 Main results

We study the concentration properties of the posterior distribution obtained from the continuous shrinkage prior in logistic regression model. Suppose that $Z \stackrel{\text{def}}{=} (Z_1, \ldots, Z_n)^T$

are independent binary random variables and $X_{n\times p} = (X_1, \ldots, X_n)^T$ is the matrix of the predictor variables for the sample size n and the number of predictors p. We consider the logistic regression model

$$\mathrm{P}(Z_i = 1|X_i) = \frac{\exp(X_i^T\beta)}{1+\exp(X_i^T\beta)}, \tag{3.1}$$

for $i = 1, \ldots, n$, where $\beta \in \mathbb{R}^p$ is the vector of regression coefficients. The likelihood function in the logistic regression model is written as

$$q_{n,\beta}(Z) = \exp\Big[\sum_{i=1}^{n}\big(Z_i \cdot X_i^T\beta - g(X_i^T\beta)\big)\Big], \tag{3.2}$$

where $g(x) \stackrel{\text{def}}{=} \log(1+e^x)$ for $x \in \mathbb{R}$. Let β_* be the true values of the coefficients of logistic regression. We denote the number of non-zero elements in β_* by $s_* \stackrel{\text{def}}{=} \|\beta_*\|_0$ and the set of non-zero elements indexes as $\xi_* = \{j : \beta_{*j} \neq 0\}$. Let E stand for a bound on the absolute value of β_{*j}: $\max_{j\in\xi_*}|\beta_j| < E$, where E may depend on n but is assumed to be non-decreasing without loss of generality.

On the p-dimensional coefficients β, we put a sequence of prior distributions, denoted as $\Pi_\alpha(\beta)$, where α is the hyper-parameter in the prior density. The prior density on β is chosen to be of the product form: $\Pi_\alpha(\beta) \stackrel{\text{def}}{=} \prod_{j=1}^{p}\pi_\alpha(\beta_j)$. For a decreasing sequence $\{\epsilon_n\}$ satisfying $\epsilon_n \prec 1$ and $n\epsilon_n^2 \succ 1$ to be chosen later, we assume

$$1 - \int_{-a_n}^{a_n}\pi_\alpha(\beta)d\beta \leq p^{-(1+\mu)} \text{ for some constant } \mu > 0, \tag{3.3}$$

$$-\log\left(\int_{\sum_{j=1}^{p-s_*}|\beta_j|\leq\eta\epsilon_n}\prod_{j=1}^{p-s_*}\pi_\alpha(\beta_j)d\beta_1\cdots d\beta_{p-s_*}\right) \prec n\epsilon_n^2, \tag{3.4}$$

$$\log s_* + 2\log(1/\epsilon_n) - \log\left(\inf_{\beta\in[-E,E]}\pi_\alpha(\beta)\right) \prec n\epsilon_n^2/s_* \tag{3.5}$$

are conditions on the prior density $\pi_\alpha(\cdot)$ for a small constant $\eta > 0$ and threshold value a_n satisfying the restriction $a_n < \epsilon_n/p$.

For simplicity of notation, we normalize the predictor to be bounded by 1, $\|X\|_\infty \stackrel{\text{def}}{=} \max_{i,j} |x_{ij}| \leq 1$. We also define the following quantity for $s \in \{1, \ldots, p\}$:

$$\kappa_{1,a_n}(s) \stackrel{\text{def}}{=} \inf \left\{ \frac{\beta^T (X^T W X) \beta}{n \|\beta\|_2^2} : 1 \leq \#\{j : |\beta_j| > a_n\} \leq s \right\} \tag{3.6}$$

for $W \in \mathbb{R}^{n \times n}$ as the diagonal matrix with the ith diagonal entry given by $W_i = g^{(2)}(X_i^T \beta)$. Since the prior on β assigns zero probability at the point zero, the exact number of nonzero elements of β is always p. A meaningful comparison with the value s_* is made by considering s the number of elements of β exceeding in absolute value the threshold a_n. In other words, only elements with absolute value larger than a_n will be treated as significant and counted towards non-zero entries. We introduce the function

$$\mathcal{L}_{n,\beta}(Z) = \log q_{n,\beta}(Z) - \log q_{n,\beta_*}(Z) - \nabla \log q_{n,\beta_*}(Z)^T (\beta - \beta_*). \tag{3.7}$$

This function plays a key role in statistical inference as it quantifies the curvature of the objective function $\beta \mapsto \log q_{n,\beta}(Z)$ around β_*.

The following theorem shows a posterior concentration property on the coefficients of the logistic regression model with shrinkage priors $\Pi_\alpha(\beta) = \prod_{j=1}^p \pi_\alpha(\beta_j)$ satisfying conditions (3.3)–(3.5). The proof is demonstrated after the theorem.

Theorem 3.2.1 *Consider the logistic regression model defined in* (3.1)*, where the vector of coefficients β is given a prior distribution with density $\Pi_\alpha(\beta)$ satisfying conditions* (3.3)–(3.5)*. Let $\bar{s} = Ln\epsilon_n^2 / \log p > s_*$ for some sufficiently large $L > 0$, the sample size*

$n \geq \left(16\bar{s}\sqrt{\log p}/\kappa_{1,a_n}(\bar{s})\right)^2$ *where* $a_n < \epsilon_n/p$ *and*

$$\epsilon_n = \max\left\{\sqrt{\frac{s_* \log p}{n}}, \frac{16}{\kappa_{1,a_n}(\bar{s})}\sqrt{\frac{\bar{s}\log p}{n}}\right\}. \tag{3.8}$$

We define the events $B_n = \{$*At least* $\bar{s}$ *entries of* $|\beta|$ *is greater than* $a_n\}$ *and* $C_n = \{\|\beta - \beta_*\|_2 > M\epsilon_n\}$ *for large constant* $M > 0$ *satisfying* $\sqrt{\bar{s}}M\epsilon_n > 2$ *and* $M > \kappa_{1,a_n}(\bar{s})^2$*. Then for some constant* $c_0 > 0$*,*

$$P^*\left(P_\alpha(C_n \cup B_n | Z, X) > 2e^{-c_0 n\epsilon_n^2}\right) \leq 7/p. \tag{3.9}$$

Hence if $p \to \infty$*, then with probability tending to one, the posterior for* β *concentrates around the truth* β_* *with respect to the Euclidean norm and the number of significant coefficient is at most* $\bar{s}$*.*

Thus with the defined contraction rate in (3.8), the estimated coefficient β in logistic regression model is asymptotically close to the true coefficients β_*. The $\mathcal{L}_2$-norm contraction rate ϵ_n is given by the maximum of the contraction rate $\sqrt{\frac{s_* \log p}{n}}$ in linear regression model (Song and Liang, 2017) and an extra term $\frac{16}{\kappa_{1,a_n}(\bar{s})}\sqrt{\frac{\bar{s}\log p}{n}}$.

3.2.2 Proof of Theorem 3.2.1

In order to show Theorem 3.2.1, we first present the following general lemmas, whose proof follows from the standard arguments used in Bayesian nonparametrics; see Ghosal and van der Vaart (2017).

Lemma 3.2.2 *Let* B_n *and* C_n *be subset of the parameter space* Θ*, and* ϕ_n *be a test function* $\phi_n(Z) \in [0, 1]$ *based on the data. If* $P(B_n) \leq b_n$*,* $E_{\beta_*}(\phi_n(Z)) \leq b_n'$*,* $\sup_{\beta \in C_n} E_\beta\big(1 -$

$\phi_n(Z)\big) \leq c_n$, *and*

$$\mathrm{P}^*\bigg\{ \int_{\mathbb{R}^p} \frac{q_{n,\beta}(Z)}{q_{n,\beta_*}(Z)} \pi(\beta) d\beta \geq d_n \bigg\} \geq 1 - d'_n, \tag{3.10}$$

then for any $\delta_n \to 0$,

$$\mathrm{P}^*\bigg(P(C_n \cup B_n | Z, X) \geq \frac{b_n + c_n}{d_n \delta_n} \bigg) \leq \delta_n + b'_n + d'_n. \tag{3.11}$$

To derive posterior concentration and variable selection results, we use the above lemma with $B_n = \{\text{At least } \bar{s} \text{ entries of } |\beta| \text{ is greater than } a_n\}$ and $C_n = \{\|\beta - \beta_*\|_2 > M\epsilon_n\}$ with $\bar{s} = Ln\epsilon_n^2 / \log p$ for a sufficiently large $L > 0$. We need to show that a test function ϕ_n, and constants $b_n, c_n, d_n, b'_n, d'_n$ such that $b'_n, d'_n \to 0$ and $(b_n + c_n)/d_n \to 0$.

Proof of Theorem 3.2.1 We first show that there exists a test function $\phi_n : \mathcal{Z}^{(n)} \to [0, 1]$ such that for some positive constant $c_1 > 0$,

$$\mathrm{E}_{\beta_*}\big[\phi(Z)\big] \leq 4/p \tag{3.12}$$

$$\sup_{\beta \in C_n} \mathrm{E}_{\beta}\big[1 - \phi(Z)\big] \leq \exp(-c_1 n \epsilon_n^2). \tag{3.13}$$

The existence of test function can be verified by following the proof of Lemma 14 in Atchadé (2017) for the test of model parameter β under the defined likelihood function $q_{n,\beta}(Z)$. For a non-empty set $\bar{\Theta} \stackrel{\text{def}}{=} \big\{\beta \in \mathbb{R}^p : 1 \leq \#\{j : |\beta_j - \beta_{*j}| > a_n\} \leq \bar{s}\big\}$,

$$\bar{\mathcal{E}}_n \stackrel{\text{def}}{=} \big\{Z \in \mathcal{Z}^{(n)} : \text{for all } \beta \in \beta_* + \bar{\Theta}, |\nabla \log q_{n,\beta}(Z)^T (\beta - \beta_*)| \leq 4\sqrt{n\bar{s}\log p}\|\beta - \beta_*\|_2/2\big\},$$

and the likelihood conditional on the subset is written as $Q_{n,\beta}(Z) \stackrel{\text{def}}{=} 1\!\!1_{Z \in \bar{\mathcal{E}}_n} q_{n,\beta}(Z)$.

For $M > 2$, $B_p(\bar{\Theta}, M\epsilon_n) \stackrel{\text{def}}{=} \{\beta \in \beta_* + \bar{\Theta} : \|\beta - \beta_*\|_2 \leq M\epsilon_n\}$ can be represented as $B_p(\bar{\Theta}, M\epsilon_n) = \cup_{l \geq 1} B(l)$, where $B(l) = \{\beta \in \beta_* + \bar{\Theta} : lM\epsilon_n < \|\beta - \beta_*\|_2 < (l+1)M\epsilon_n\}$.

For each $l \geq 1$, let S_l be the set of maximal points separated by the distance $lM\epsilon_n/2$ in $B(l)$. Therefore, there is a test function $\phi_{\beta_k} : \mathcal{Z}^{(n)} \to [0,1]$ for each $\beta_k \in S_l$ such that

$$\mathrm{E}_{\beta_*}\big[\phi_{\beta_k}(Z)\big] \leq \sup_{Q(Z)\in \mathrm{conv}(\mathcal{P}_{\beta_k})} \mathcal{H}_{1/2}\big(q_{n,\beta_*}(Z), Q(Z)\big) \tag{3.14}$$

$$\sup_{Q(Z)\in\mathcal{P}_{\beta_k}} \int_{\mathcal{Z}^{(n)}} \big(1-\phi_{\beta_k}(Z)\big)Q(Z)dZ \leq \sup_{Q(Z)\in \mathrm{conv}(\mathcal{P}_{\beta_k})} \mathcal{H}_{1/2}\big(q_{n,\beta_*}(Z), Q(Z)\big), \tag{3.15}$$

where $\mathcal{H}_{1/2}(q_1,q_2) \stackrel{\text{def}}{=} \int_{\mathcal{Z}^{(n)}} q_1^{1/2}(z)q_2^{1/2}(z)dz$ is the Hellinger transform for any two integrable non-negative functions on $\mathcal{Z}^{(n)}$, $\mathcal{P}_{\beta_k} \stackrel{\text{def}}{=} \{Q_{n,\mu}(Z) : \mu \in \beta_* + \bar{\Theta} \text{ and } \|\mu-\beta_k\|_2 \leq lM\epsilon_n/2\}$ and any $Q(Z) \in \mathrm{conv}(\mathcal{P}_\beta)$ can be written as a finite convex combination $Q(Z) = \sum_m \alpha_m Q_{n,\mu_m}(Z)$, where $\alpha_m \geq 0$, $\sum_m \alpha_m = 1$, $\mu_m \in \beta_* + \bar{\Theta}$ and $\|\mu_m - \beta_k\|_2 \leq lM\epsilon_n/2$. Since for all $\beta_k \in B(l)$, $\|\beta_k - \beta_*\| > lM\epsilon_n$, we have $\|\mu_m - \beta_*\|_2 > lM\epsilon_n/2 > \epsilon_n$. Thus,

$$\mathcal{H}_{1/2}\left(q_{n,\beta_*}(Z), Q(Z)\right) = \int_{\mathcal{Z}^{(n)}} \sqrt{\sum_m \alpha_m \mathbb{1}_{\bar{\mathcal{E}}_n(Z)} \frac{q_{n,\mu_m}(Z)}{q_{n,\beta_*}(Z)}} q_{n,\beta_*}(Z)dZ. \tag{3.16}$$

Now for $Z \in \bar{\mathcal{E}}_n$ and $\mu_m \in \beta_* + \bar{\Theta}$, we have that

$$|X_i^T(\mu_m - \beta_*)| \leq \|X\|_\infty \|\mu_m - \beta_*\|_1 \leq \sqrt{\bar{s}}\|\mu_m - \beta_*\|_2 + (p-\bar{s})a_n \leq 2\sqrt{\bar{s}}\|\mu_m - \beta_*\|_2$$

since $(p-\bar{s})a_n \leq (p-\bar{s})\epsilon_n/p < \epsilon_n < \|\mu_m - \beta_*\|_2$. Hence

$$\begin{aligned}
\mathcal{L}_{n,\mu_m} &\leq -\frac{n}{2+\max_i |X_i^T(\mu_m-\beta_*)|}(\mu_m-\beta_*)^T \frac{X^TWX}{n}(\mu_m-\beta_*) \\
&\leq -\frac{n}{2+2\sqrt{\bar{s}}\|\mu_m-\beta_*\|_2}\kappa_{1,a_n}(\bar{s})\|\mu_m-\beta_*\|_2^2
\end{aligned}$$

given the definition of $\kappa_{1,a_n}(\bar{s})$ in (3.6).

Thus for $\mu_m \in \beta_* + \bar{\Theta}$, $\|\mu_m - \beta_*\|_2 > lM\epsilon_n/2$ and $Z \in \bar{\mathcal{E}}_n$, given the definition of

$\mathcal{L}_{n,\beta}(Z)$ in (3.7), we have the following inequality:

$$\begin{aligned}
\frac{q_{n,\mu_m}(Z)}{q_{n,\beta_*}(Z)} &= \exp\Big\{\nabla \log q_{n,\beta_*}(Z)^T(\mu_m-\beta_*)+\mathcal{L}_{n,\mu_m}(Z)\Big\}\\
&\le \exp\Big\{\frac{4\sqrt{n\bar{s}\log p}}{2}\|\mu_m-\beta_*\|_2-\frac{1}{2}\frac{n\kappa_{1,a_n}(\bar{s})\|\mu_m-\beta_*\|_2^2}{1+\sqrt{\bar{s}}\|\mu_m-\beta_*\|_2}\Big\}\\
&\le \exp\Big\{\frac{8\sqrt{n\bar{s}\log p}-\big(n\kappa_{1,a_n}(\bar{s})-8\bar{s}\sqrt{n\log p}\big)\epsilon_n}{4\|\mu_m-\beta_*\|_2^{-1}+4\sqrt{\bar{s}}}-\frac{1}{4}\frac{n\kappa_{1,a_n}(\bar{s})\|\mu_m-\beta_*\|_2^2}{1+\sqrt{\bar{s}}\|\mu_m-\beta_*\|_2}\Big\}\\
&\le \exp\Big\{\frac{8\sqrt{n\bar{s}\log p}-\frac{1}{2}n\kappa_{1,a_n}(\bar{s})\frac{16}{\kappa_{1,a_n}(\bar{s})}\sqrt{\frac{\bar{s}\log p}{n}}}{4\|\mu_m-\beta_*\|_2^{-1}+4\sqrt{\bar{s}}}-\frac{1}{4}\frac{n\kappa_{1,a_n}(\bar{s})\|\mu_m-\beta_*\|_2^2}{1+\sqrt{\bar{s}}\|\mu_m-\beta_*\|_2}\Big\}\\
&\le \exp\Big\{-\frac{1}{4}\frac{n\kappa_{1,a_n}(\bar{s})(lM\epsilon_n/2)^2}{1+\sqrt{\bar{s}}(lM\epsilon_n/2)}\Big\},
\end{aligned}$$

under the conditions of $\epsilon_n \ge \frac{16}{\kappa_{1,a_n}(\bar{s})}\sqrt{\frac{\bar{s}\log p}{n}}$ and $n\kappa_{1,a_n}(\bar{s})-16\bar{s}\sqrt{n\log p}>0$. The later one imposes a restriction on the sample size $n \ge \big(16\bar{s}\sqrt{\log p}/\kappa_{1,a_n}(\bar{s})\big)^2$. Since $\bar{s} \overset{\text{def}}{=} Ln\epsilon_n^2/\log p$ and $\sqrt{\bar{s}}lM\epsilon_n > \sqrt{\bar{s}}M\epsilon_n > 2$, we have

$$\begin{aligned}
\frac{n\kappa_{1,a_n}(\bar{s})(lM\epsilon_n/2)^2}{1+\sqrt{\bar{s}}\,(lM\epsilon_n/2)} &\ge \frac{8\sqrt{n\bar{s}\log p}\,lM\epsilon_n(\sqrt{\bar{s}}lM\epsilon_n/2)}{1+\sqrt{\bar{s}}(lM\epsilon_n/2)}\\
&\ge \frac{8\sqrt{n\bar{s}\log p}\,lM\epsilon_n}{(\sqrt{\bar{s}}\,lM\epsilon_n/2)^{-1}+1}\\
&\ge 4\sqrt{n\bar{s}\log p}\,lM\epsilon_n = c_4 lMn\epsilon_n^2.
\end{aligned}$$

for $c_4 = 4\sqrt{L}$. Therefore, the right hand side of (3.16) can be bounded by

$$\begin{aligned}
\mathcal{H}_{1/2}\big(q_{n,\beta_*}(Z),Q(Z)\big)^2 &\le \sum_m \alpha_m \exp\Big\{-\frac{1}{4}\frac{n\kappa_{1,a_n}(\bar{s})(lM\epsilon_n/2)^2}{1+\sqrt{\bar{s}}(lM\epsilon_n/2)}\Big\}\\
&\le \sum_m \alpha_m \exp\{-c_4 lMn\epsilon_n^2\}.
\end{aligned}$$

Thus we conclude that

$$\mathrm{E}_{\beta_*}\big[\phi_{\beta_k}(Z)\big] \le \exp\Big\{ -\frac{1}{2}c_4 lMn\epsilon_n^2\Big\} \tag{3.17}$$

$$\sup_{Q(Z)\in\mathcal{P}_\beta}\int_{\mathcal{Z}^{(n)}}\big(1-\phi_{\beta_k}(Z)\big)Q(Z)dZ \le \exp\Big\{ -\frac{1}{2}c_4 lMn\epsilon_n^2\Big\}. \tag{3.18}$$

Now we set the test function to be $\phi(Z) = 1\!\!1_{Z\in\bar{\mathcal{E}}_n}\sup_{l\ge1}\max_{\beta_k\in S_l}\phi_{\beta_k}(Z) + 1\!\!1_{Z\in\bar{\mathcal{E}}_n^c}$. For any $\beta\in\beta_*+\bar{\Theta}$ such that $\|\beta-\beta_*\|_2 > M\epsilon_n$, it will be within the distance of $M\epsilon_n/2$ of a point in S_l for some $l\ge1$. Hence by (3.18),

$$\begin{aligned}\sup_{\beta\in C_n}\mathrm{E}_\beta\big[1-\phi(Z)\big] \le& \int_{\mathcal{Z}^{(n)}}\big(1-\phi_\beta(Z)\big)1\!\!1_{Z\in\bar{\mathcal{E}}_n}q_{n,\beta}(Z)dZ\\ \le&\exp\Big\{ -\frac{1}{2}c_4 Mn\epsilon_n^2\Big\}.\end{aligned} \tag{3.19}$$

Since the size of the defined set S_l is upper bounded by $D_l \stackrel{\text{def}}{=} D(lM\epsilon_n/2, B_p\big(\bar{\Theta},(l+1)M\epsilon_n\big)\big)$, the number of points within $B_p\big(\bar{\Theta},(l+1)M\epsilon_n\big)$ that are separated by $\mathcal{L}_2$ distance of $lM\epsilon_n/2$. Then by (3.17) and the test function $\phi(Z)$,

$$\mathrm{E}_{\beta_*}\big[\phi(Z)\big] \le \mathrm{P}^*\big(Z\in\bar{\mathcal{E}}_n^c\big) + \sum_{l\ge1}D_l\exp\Big\{ -\frac{1}{2}c_4 lMn\epsilon_n^2\Big\}. \tag{3.20}$$

By Hoeffding's inequality and $|\triangledown\log q_{n,\beta_*}(Z)^T(\beta-\beta_*)| \le \|\log q_{n,\beta_*}(Z)\|_\infty\|\beta-\beta_*\|_1$,

$$\begin{aligned}\mathrm{P}^*\big(Z\in\bar{\mathcal{E}}_n^c\big) &\le \mathrm{P}^*\bigg(\|\log q_{n,\beta_*}(Z)\|_\infty > \frac{4\sqrt{n\log p}}{2}\bigg)\\ &\le \mathrm{P}^*\bigg(\max_{1\le j\le p}\Big|\sum_{i=1}^n\big(z_i - g'(X_i^T(\beta-\beta_*))\big)x_{ij}\Big| > \frac{4\sqrt{n\log p}}{2}\bigg)\\ &\le 2\exp\bigg(\log p - \frac{16n\log p}{8n}\bigg) \le 2/p.\end{aligned} \tag{3.21}$$

For continuous shrinkage prior, no exact zero will be generated on β. Therefore, a thresholding policy is applied. When $|\beta_j| < a_n$ where $a_n < \epsilon_n/p$, the coefficient is treated as zero. For all the coordinates $j \notin \xi_*$ such that $\beta_{*j} = 0$, the set $B_p\big(\bar{\Theta}, (l+1)M\epsilon_n\big)$ will cover all the points that $|\beta_j| < a_n < (l+1)M\epsilon_n$ around $\beta_{*j} = 0$ and the distance between these points and β_{*j} will be smaller than $lM\epsilon_n/2$. By the definition of $\bar{\Theta}$, although the shrinkage prior does not produce a strict sparsity on β, the $p - \bar{s}$ coordinates where $|\beta_j| < a_n$ are treated like zero in the point-mass prior. Using the arguments given in Example 7.1 of Ghosal et al. (2000) for a k-dimensional ball, $k \leq \bar{s}$, we obtain the bound $\sup_{l\geq 1} D_l \leq (24)^{\bar{s}} e^{\bar{s}\log(pe)}$. Therefore, the last term on the right hand side of (3.20) can be bounded by

$$\begin{aligned}
\sum_{l\geq 1} D_l \exp\Big\{ -\frac{1}{2} c_4 lMn\epsilon_n^2 \Big\} \leq & (24)^{\bar{s}} e^{\bar{s}\log(pe)} \sum_{l\geq 1} \exp\Big\{ -\frac{1}{2} c_4 lMn\epsilon_n^2 \Big\} \\
= & (24)^{\bar{s}} e^{\bar{s}\log(pe)} \frac{\exp\{-\frac{1}{2}c_4 Mn\epsilon_n^2\}}{1 - \exp\{-\frac{1}{2}c_4 Mn\epsilon_n^2\}} \\
\leq & 2\exp\Big\{ \bar{s}\log(24e) + \bar{s}\log p - 2\sqrt{L}Mn\epsilon_n^2 \Big\} \\
\leq & 2\exp\Big\{ \bar{s}\log(24e) + \bar{s}\log p - 2\times 16^2\sqrt{L}M\bar{s}\log p/\kappa_{1,a_n}(\bar{s})^2 \Big\}.
\end{aligned}$$

If M is large enough such that $\frac{2\times 16^2\sqrt{L}M}{\kappa_{1,a_n}(\bar{s})^2} > 1 + \log(24e)$, we can conlude that

$$\sum_{l\geq 1} D_l \exp\Big\{ -\frac{1}{2} c_4 lMn\epsilon_n^2 \Big\} \leq 2/p. \tag{3.22}$$

The condition can be verified since $M > L'\kappa_{1,a_n}(\bar{s})^2$ for a sufficiently large constant $L' > 0$.

Combining (3.21), (3.22) and (3.20), we obtain (3.12). Setting $c_1 = \frac{1}{2}c_4 M = 2\sqrt{L}M$ and using the bound in (3.19), we show that (3.13) holds. Thus we verify the existence

of a test function with the stated properties.

Next we show that for sufficiently large n,

$$\mathrm{P}^*\left\{\int_{\mathbb{R}^p}\frac{q_{n,\beta}(Z)}{q_{n,\beta_*}(Z)}\Pi_\alpha(\beta)d\beta \geq e^{-c_2 n\epsilon_n^2}\right\} > 1-2/p \tag{3.23}$$

for some positive constant $c_2 < c_1$. This can be achieved by setting $M > c_2/2\sqrt{L}$. Now for all the $\beta \in \mathbb{R}^p$ and $Z \in \mathcal{E}_{n,0} \stackrel{\text{def}}{=} \left\{Z \in \mathcal{Z}^{(n)} : \|\triangledown \log q_{n,\beta_*}(Z)\|_\infty \leq 2\sqrt{n\log p}\right\}$, using $g^{(2)}(z) \leq 1/4$, we have that

$$\begin{aligned}
\int_{\mathbb{R}^p}\frac{q_{n,\beta}(Z)}{q_{n,\beta_*}(Z)}\Pi_\alpha(\beta)d\beta &= \int_{\mathbb{R}^p}\exp\left\{\triangledown \log q_{n,\beta}(Z)^T(\beta-\beta_*)+\mathcal{L}_{n,\beta}(Z)\right\}\Pi_\alpha(\beta)d\beta \\
&\geq \int_{\mathbb{R}^p}\exp\left\{-\|\triangledown \log q_{n,\beta_*}(Z)\|_\infty\|\beta-\beta_*\|_1-\frac{n}{8}\|X(\beta-\beta_*)/\sqrt{n}\|_2^2\right\}\Pi_\alpha(\beta)d\beta \\
&\geq \int_{\mathbb{R}^p}\exp\left\{-2\sqrt{n\log p}\|\beta-\beta_*\|_1-\frac{n}{8}\|\beta-\beta_*\|_1^2\right\}\Pi_\alpha(\beta)d\beta,
\end{aligned}$$

since $\left|\triangledown \log q_{n,\beta_*}(Z)^T(\beta-\beta_*)\right| \leq \|\triangledown \log q_{n,\beta_*}(Z)\|_\infty\|\beta-\beta_*\|_1$ and

$$\|X(\beta-\beta_*)\|_2^2 = \sum_{i=1}^n\left(X_i^T(\beta-\beta_*)\right)^2 \leq n\big(\|X\|_\infty\|\beta-\beta_*\|_1\big)^2.$$

Therefore, for a sufficiently small constant $\eta > 0$,

$$\int_{\mathbb{R}^p}\frac{q_{n,\beta}(Z)}{q_{n,\beta_*}(Z)}\Pi_\alpha(\beta)d\beta \geq \exp\big\{-4\eta\sqrt{n\log p}\epsilon_n-\eta^2 n\epsilon_n^2/2\big\}\mathrm{P}_\alpha\big(\{\|\beta-\beta_*\|_1<2\eta\epsilon_n\}\big). \tag{3.24}$$

From the proof of Theorem A.2 in Song and Liang (2017), the event $\{\|\beta-\beta_*\|_1 < 2\eta\epsilon_n\}$ contains the event

$$\left\{\beta_j \in \left[\beta_{*j}-\frac{\eta\epsilon_n}{s_*},\beta_{*j}+\frac{\eta\epsilon_n}{s_*}\right] \text{ for all } j \in \xi_*\right\} \cap \left\{\sum_{j\notin\xi_*}|\beta_j| \leq \eta\epsilon_n\right\}.$$

For the prior distribution $\Pi_\alpha(\beta) = \prod_{j=1}^{p} \pi_\alpha(\beta_j)$ satisfying condtions (3.3)–(3.5),

$$\mathrm{P}_\alpha\left(\sum_{j\notin\xi_*}|\beta_j| \le \eta\epsilon_n\right) = \int_{\|\beta\|_1\le\eta\epsilon_n}\prod_{j=1}^{p-s_*}\pi_\alpha(\beta_j)d\beta_1\cdots d\beta_{p-s_*} \ge e^{-c_5 n\epsilon_n^2}, \tag{3.25}$$

and since $|\beta_{*j}| - \eta\epsilon_n/s_* \ge -E$, $|\beta_{*j}| + \eta\epsilon_n/s_* \le E$,

$$\mathrm{P}_\alpha\left(\left\{\beta_j \in \left[\beta_{*j} - \frac{\eta\epsilon_n}{s_*}, \beta_{*j} + \frac{\eta\epsilon_n}{s_*}\right] \text{ for all } j \in \xi_*\right\}\right) \ge \frac{2\eta\epsilon_n}{s_*}\cdot\inf_{\beta\in[-E,E]}\pi_\alpha(\beta) \ge e^{-c_6 n\epsilon_n^2} \tag{3.26}$$

for some constants $c_5 > 0$ and $c_6 > 0$. Therefore, given (3.25) and (3.26),

$$\begin{aligned}\mathrm{P}_\alpha\left(\{\|\beta-\beta_*\|_1 < 2\eta\epsilon_n\}\right) \ge& \mathrm{P}_\alpha\left(\left\{\beta_j \in \left[\beta_{*j} - \frac{\eta\epsilon_n}{s_*}, \beta_{*j} + \frac{\eta\epsilon_n}{s_*}\right] \text{ for all } j \in \xi_*\right\}\right)\\ &\times \mathrm{P}_\alpha\left(\sum_{j\notin\xi_*}|\beta_j| \le \eta\epsilon_n\right) \ge e^{-(c_5+c_6)n\epsilon_n^2}.\end{aligned} \tag{3.27}$$

Combining the results in (3.24) and (3.27) with the constraints $s_*\log p \prec n\epsilon_n^2$, we conclude that, for $c_2 \stackrel{\text{def}}{=} 4\eta/\sqrt{s_*} + \eta^2 + c_5 + c_6$ and $\epsilon_n \le \sqrt{\frac{s_*\log p}{n}}$,

$$\int_{\mathbb{R}^p}\frac{q_{n,\beta}(Z)}{q_{n,\beta_*}(Z)}\Pi_\alpha(\beta)d\beta \ge \exp\left\{-\left(4\eta/\sqrt{s_*} + \eta^2/2\right)n\epsilon_n^2 - c_2' n\epsilon_n^2\right\} = e^{-c_2 n\epsilon_n^2}. \tag{3.28}$$

Since (3.28) holds for all $Z \in \mathcal{E}_{n,0}$, we have that

$$\begin{aligned}&\mathrm{P}^*\left\{\int_{\mathbb{R}^p}\frac{q_{n,\beta}(Z)}{q_{n,\beta_*}(Z)}\Pi_\alpha(\beta)d\beta \ge e^{-c_2 n\epsilon_n^2}\right\}\\ &\ge 1 - \mathrm{P}^*(Z\notin\mathcal{E}_{n,0})\\ &\ge 1 - \mathrm{P}^*\left(\max_{1\le j\le p}\Big|\sum_{i=1}^{n}\big(Z_i - g'(X_i^T(\beta-\beta_*))\big)x_{ij}\Big| > 2\sqrt{n\log p}\right)\\ &\ge 1 - 2\exp\left(\log p - \frac{16n\log p}{8n}\right) = 1 - 2/p,\end{aligned}$$

leading to (3.23).

Finally, for $B_n = \{\text{At least } \bar{s} \text{ entries of } |\beta| \text{ is greater than } a_n\}$ with $\bar{s} = Ln\epsilon_n^2/\log p$, we show that the probability $\mathrm{P}_\alpha(B_n) \leq \exp(-c_3 n\epsilon_n^2)$ for constant $c_3 > 0$. Let $N = \#\big\{j : |\beta_j| \geq a_n\big\}$. Note that $N \sim \mathrm{Bin}(p, \nu_n)$, where $\nu_n = \int_{|\beta|\geq a_n} \pi_\alpha(\beta)d\beta$, so $\mathrm{P}_\alpha(B_n) = \mathrm{P}_\alpha\big(N \geq \bar{s}\big)$. By the condition (3.3) on a shrinkage prior, $\nu_n \asymp p^{-(1+\mu)}$ for $\mu > 0$ and Lemma A.3 of Song and Liang (2017), we have

$$\mathrm{P}_\alpha(B_n) \leq 1 - \Phi\big(\sqrt{2p \cdot H(\nu_n, (\bar{s}-1)/p)}\big) \leq \frac{\exp\big\{-p \cdot H(\nu, (\bar{s}-1)/p)\big\}}{\sqrt{2\pi}\sqrt{2p \cdot H(\nu_n, (\bar{s}-1)/p)}},$$

where $p \cdot H(\nu_n, (\bar{s}-1)/p) = (\bar{s}-1)\log\big(\frac{\bar{s}-1}{p\nu_n}\big) + (p-\bar{s}+1)\log\big(\frac{p-\bar{s}+1}{p-p\nu_n}\big)$ and it is sufficient to show that $p \cdot H(\nu_n, (\bar{s}-1)/p) \gtrsim n\epsilon_n^2$. Since $1/(p\nu_n) \geq p^\nu$ and $\bar{s} = Ln\epsilon_n^2/\log(p)$, then we have $(\bar{s}-1)\log\big(\frac{\bar{s}-1}{p\nu_n}\big) \asymp n\epsilon_n^2$ and $(p-\bar{s}+1)\log\big(\frac{p-\bar{s}+1}{p-p\nu_n}\big) \prec n\epsilon_n^2$. Therefore, for a constant c_3, $\mathrm{P}_\alpha(B_n) < e^{-c_3 n\epsilon_n^2}$.

Combining (3.12), (3.13), (3.23) and $\mathrm{P}_\alpha(B_n) < e^{-c_3 n\epsilon_n^2}$ and using Lemma 3.2.2, we conclude that there exist constants c_1, c_2 and c_3, such that

$$\mathrm{P}^*\left(\mathrm{P}_\alpha\big(C_n \cup B_n | Z, X\big) \geq \frac{e^{-c_3 n\epsilon_n^2} + e^{-c_1 n\epsilon_n^2}}{\delta_n e^{-c_2 n\epsilon_n^2}}\right) \leq \delta_n + 6/p,$$

where B_n and C_n are as defined in the theorem, δ_n is any positive sequence and the constants c_1, c_2, c_3 are such that $c_2 < \min(c_1, c_3)$. This can be achieved by setting η small so that $c_2 = 4\eta/\sqrt{s_*} + \eta^2/2 + c_5 + c_6$ is smaller than c_3 and by setting M large so that $M > c_2/2\sqrt{L}$. Then with $c_0 \stackrel{\text{def}}{=} \frac{1}{2}(c_2 - \min(c_1, c_3))$ and $\delta_n \stackrel{\text{def}}{=} e^{-c_0 n\epsilon_n^2}$, we have

$$\mathrm{P}^*\big(\mathrm{P}_\alpha(C_n \cup B_n | Z, X) \geq \exp\{-c_0 n\epsilon_n^2\}\big) \leq \exp\{-c_0 n\epsilon_n^2\} + 6/p \leq 7/p,$$

thus establishing the result.

3.3 Examples of Shrinkage Priors

The conditions imposed on the prior through (3.3)–(3.5) are somewhat abstract. We now give examples of prior densities which satisfy the required conditions and hence lead to posterior concentration near the truth.

3.3.1 Spike-and-slab prior

We assume that the prior density of β_j, $j = 1, \ldots, p$ has the following form:

$$\beta_j|\rho_j, \sigma_0^2, \sigma_1^2 \sim (1-\rho_j)\mathrm{N}(0, \sigma_0^2) + \rho_j \mathrm{N}(0, \sigma_1^2), \quad \rho_j|b \sim \mathrm{Bin}(1, b), \tag{3.29}$$

where $\mathrm{N}(\mu, \sigma^2)$ stands for the normal distribution with mean μ and variance σ^2, and Bin for the binomial distribution. After integrating out ρ_j, the prior density for β_j, $j = 1, \ldots, p$, can be written as $\pi_\alpha(\beta_j) = (1-b)\phi(\beta_j; 0, \sigma_0^2) + b\phi(\beta_j; 0, \sigma_1^2)$, where $\phi(\cdot; \mu, \sigma^2)$ denotes the density for normal distribution with mean μ and variance σ^2. Here the role of the hyper-parameter α is $(b, \sigma_0^2, \sigma_1^2)$. The prior approximates a point-mass prior if the value of σ_0^2 is very small. The following theorem presents the posterior concentration of β in logistic regression model with spike-and-slab prior.

Theorem 3.3.1 *Consider the spike-and-slab prior* (3.29) *on the coefficients* β *of logistic regression model* (3.1). *If* $\log \sigma_1 + E^2/2\sigma_1^2 \prec n\epsilon_n^2/s_*$, $\sigma_0 \leq a_n/\sqrt{(1+\mu')\log p} < \epsilon_n/\left(p\sqrt{(1+\mu')\log p}\right)$ *and* $b = p^{-(1+\mu')}$ *for a constant* $\mu' > 0$, *then the posterior concentration property in* (3.9) *holds.*

Proof In order to show the posterior concentration results with the defined contraction rate ϵ_n, we need to show that the conditions (3.3)–(3.5) on the prior density hold for the

spike-and-slab prior. First for $a_n < \epsilon_n/p$,

$$\begin{aligned}1 - \int_{-a_n}^{a_n} \pi_\alpha(\beta)d\beta =& 2\int_{a_n}^{\infty} \big[(1-b)\phi(\beta;0,\sigma_0^2) + b\phi(\beta;0,\sigma_1^2)\big]d\beta \\ =& 2(1-b)\int_{a_n}^{\infty} \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\beta^2/2\sigma_0^2} d\beta + 2b\int_{a_n}^{\infty} \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\beta^2/2\sigma_1^2} d\beta \\ \le& \sqrt{2}\exp\{-a_n^2/2\sigma_0^2\}(a_n\sqrt{\pi}/\sigma_0)^{-1} + b. \end{aligned} \tag{3.30}$$

Since $b = p^{-(1+\mu')}$ and $a_n/\sigma_0 \ge \sqrt{(1+\mu')\log p}$ for constant $\mu' > 0$, we can show that for any constant $0 < \mu < \mu'$, $1 - \int_{-a_n}^{a_n} \pi_\alpha(\beta)d\beta \le p^{-(1+\mu)}$. Therefore, (3.3) holds. Next,

$$\int_{\sum_{j=1}^{p-s_*}|\beta_j| \le \eta\epsilon_n} \prod_{j=1}^{p-s_*} \pi_\alpha(\beta_j)d\beta_1 \cdots d\beta_{p-s_*} \le \big(1 - p^{-(1+\mu')}\big)^{p-s_*} \mathrm{P}_\alpha\bigg(\sum_{j=1}^{p-s_*} |T_j| \le \eta\epsilon_n\bigg) \tag{3.31}$$

where T_j, $j = 1, \ldots, p - s_*$, are independently and identically distributed as $\mathrm{N}(0, \sigma_0^2)$. Since $\mathrm{E}_\alpha(|T_j|) = \sigma_0\sqrt{2/\pi}$, we have $\mathrm{E}_\alpha\big(\sum_{j=1}^{p-s_*}|T_j|\big) = (p-s_*)\sigma_0\sqrt{2/\pi}$. Therefore by the central limit theorem, for $\sigma_0 < \eta\epsilon_n/p$, we have $\mathrm{P}_\alpha\big(\sum_{j=1}^{p-s_*}|T_j| \le \eta\epsilon_n\big) \ge \mathrm{P}_\alpha\big(\sum_{j=1}^{p-s_*}|T_j| \le (p-s_*)\sigma_0\sqrt{2/\pi}\big) \to 1/2$. Thus for a constant $c'' > 0$, we have

$$\int_{\sum_{j=1}^{p-s_*}|\beta_j| \le \eta\epsilon_n} \prod_{j=1}^{p-s_*} \pi_\alpha(\beta_j)d\beta_1 \cdots d\beta_{p-s_*} \ge \frac{1}{4}\big(1 - p^{-(1+\mu')}\big)^p \ge \exp(-c''n\epsilon_n^2). \tag{3.32}$$

Lastly for $E > 0$ nondecreasing in n, we obtain

$$\begin{aligned} & -\log\left(\inf_{\beta\in[-E,E]}\big[(1-b)\phi(\beta;0,\sigma_0^2) + b\phi(\beta;0,\sigma_1^2)\big]\right) \\ \le& -\log\left(p^{-(1+\mu')}\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\{-E^2/2\sigma_1^2\}\right) \\ =& \frac{1}{2}\log(2\pi) + \log\sigma_1 + (1+\mu')\log p + E^2/2\sigma_1^2 \prec n\epsilon_n^2/s_* \end{aligned} \tag{3.33}$$

since $\log \sigma_1 + E^2/2\sigma_1^2 \prec n\epsilon_n^2/s_*$. Then (3.5) holds. Now, with the conditions on prior density, Theorem 3.2.1 can be applied.

3.3.2 Dirichlet-Laplace prior

In Bhattacharya et al. (2015), a global-local shrinkage prior is proposed on β. Specifically, the prior distribution has the following structure:

$$\beta_j|\phi_j, \tau \sim \text{DE}(\phi_j\tau);\ (\phi_1, \ldots, \phi_p)|\alpha \sim \text{Dirichlet}(\alpha, \ldots, \alpha);\ \tau|\alpha \sim \text{Gamma}(p\alpha, 2), \quad (3.34)$$

where $\text{DE}(\lambda)$ stands for double exponential distribution with scale λ and $\text{Gamma}(a, b)$ for the gamma distribution with shape parameter a and scale parameter b. The hyper-parameter α controls the level of shrinkage such that smaller α means more shrinkage and sparser model. For convenience, we equivalently represent the prior distribution through the following hierarchical structure:

$$\beta_j|\psi_j \sim \text{DE}(\psi_j),\ \psi_j|\alpha \sim \text{Gamma}(\alpha, 2),\ j = 1, \ldots, p. \quad (3.35)$$

The following theorem shows the consistency of β in logistic regression model with the Dirichlet-Laplace prior.

Theorem 3.3.2 *Consider the Dirichlet-Laplace prior in* (3.35) *for the logistic regression model. Assume that the true regression coefficients satisfy the condition that* $\max_{j\in\xi_*} |\beta_{*j}| < E$ *with* $E \prec n\epsilon_n^2/s_*$. *If* $\alpha \prec \epsilon_n/p$ *and* $\alpha \succeq p^{-(1+\nu)}$ *for some constant* $\nu > 0$, *then the posterior concentration property in* (3.9) *holds.*

Proof We need to verify the prior conditions (3.3)–(3.5) hold for the Dirichlet-Laplace prior density π_α. First, for $a_n < \epsilon_n/p$, following the proof of Lemma 3.3 in Bhattacharya

et al. (2015), we obtain

$$\begin{aligned}1-\int_{-a_n}^{a_n}\pi_\alpha(\beta)d\beta=&\int_0^\infty e^{-a_n/\psi}\frac{1}{\Gamma(\alpha)2^\alpha}\psi^{\alpha-1}e^{-\psi/2}d\psi\\ =&\frac{1}{\Gamma(\alpha)2^\alpha}\left[\int_0^{4a_n}\psi^{\alpha-1}e^{-a_n/\psi-\psi/2}d\psi+\int_{4a_n}^\infty\psi^{\alpha-1}e^{-a_n/\psi-\psi/2}d\psi\right]\\ \le&\frac{1}{\Gamma(\alpha)2^\alpha}\left[C+\int_{4a_n}^\infty\psi^{-1}e^{-\psi/2}d\psi\right],\end{aligned}\tag{3.36}$$

where $C>0$ is a constant independent of a_n. Since $\int_{4a_n}^\infty\psi^{-1}e^{-\psi/2}d\psi\le\int_{2a_n}^\infty t^{-1}e^{-t}dt\le-\log a_n$, we conclude that for $a_n<\epsilon_n/p$ and a constant $C'>0$ free of a_n,

$$1-\int_{-a_n}^{a_n}\pi_\alpha(\beta)d\beta\le C'\alpha\log(1/a_n)=C'p^{-(1+\nu)}\log(1/a_n)\le p^{-(1+\mu)},\tag{3.37}$$

where $0<\mu<\nu$ is a constant. Next, β_j, $j=1,\ldots,p-s_*$, are independently and identically distributed as in (3.35), so $\mathrm{E}_\alpha(|\beta_j|)=\mathrm{E}_\alpha(\mathrm{E}(\beta_j|\psi_j))=2\alpha$. Then, by the central limit theorem, $\mathrm{P}_\alpha\left(\sum_{j=1}^{p-s_*}|\beta_j|\le\eta\epsilon_n\right)\ge\mathrm{P}_\alpha\left(\sum_{j=1}^{p-s_*}|\beta_j|\le2(p-s_*)\alpha\right)\to1/2$ since $p\alpha\prec\epsilon_n$. Thus the second condition (3.4) on the prior density holds. Finally

$$\begin{aligned}\inf_{\beta\in[-E,E]}\pi_\alpha(\beta)=&\inf_{\beta\in[-E,E]}\int_0^\infty\frac{1}{\psi}e^{-|\beta|/\psi}\frac{1}{\Gamma(\alpha)2^\alpha}\psi^{\alpha-1}e^{-\psi/2}d\psi\\ \ge&\frac{1}{\Gamma(\alpha)2^\alpha}\int_0^{\eta'}\psi^{\alpha-2}e^{-E/\psi-\psi/2}d\psi\\ \ge&\frac{C''}{\Gamma(\alpha)2^\alpha}\int_0^{\eta'}\psi^{-2}e^{-E/\psi}d\psi\\ \ge&\frac{C''}{\Gamma(\alpha)}\int_{\eta'^{-1}}^\infty e^{-Et}dt=\frac{C''}{\Gamma(\alpha)}\frac{1}{E}e^{-E/\eta'},\end{aligned}\tag{3.38}$$

where $\eta'>0$ is a small constant and C'' is a constant independent of α. Therefore, for $E\prec n\epsilon_n^2/s_*$ and $\log p\prec n\epsilon_n^2/s_*$, we have $-\log\left(\inf_{\beta\in[-E,E]}\pi_\alpha(\beta)\right)\le-\log C''+(1+\nu)\log p+E/\eta'+\log E\prec n\epsilon_n^2/s_*$. This completes the proof of (3.5).

3.3.3 Horseshoe prior

Another shrinkage prior we consider is the horseshoe prior introduced in Carvalho, Polson and Scott (2009). The horseshoe prior assumes that each β_j is conditionally independent with density $\pi_\alpha(\beta_j)$ having the following structure:

$$\beta_j|\lambda,\alpha \sim \mathrm{N}(0,\lambda_j^2\alpha^2), \quad \lambda_j \sim \mathrm{C}^+(0,1), \tag{3.39}$$

where $\mathrm{C}^+(0,1)$ is the half-Cauchy distribution, that is, the distribution of the absolute value of a standard Cauchy distribution, and the hyperparameter $\alpha > 0$ controls the global shrinkage. Therefore, we have the density function of β_j: $\pi_\alpha(\beta_j) = \int_0^\infty \phi(\beta_j;0,\lambda^2\alpha^2) f_{\mathrm{C}^+}(\lambda)d\lambda$, where $f_{\mathrm{C}^+}(\lambda) = (2/\pi)(1+\lambda^2)^{-1}$ stands for the density of the half-Cauchy distribution.

Theorem 3.3.3 *Consider the horseshoe prior* (3.39) *on* β *in a logistic regression model* (3.1). *Assume that* $\alpha < p^{-(2+\mu'')}$ *for a constant* $\mu'' > 0$, *and* $\alpha^2/E < p^{-(1+\nu)}$ *for a constant* $\nu > 0$. *Then the posterior concentration property in* (3.9) *holds.*

Proof In order to prove the posterior concentration, we need to show that under the stated conditions of the theorem, relations (3.3)–(3.5) hold. First, for a constant $\mu'' > 0$,

$$\begin{aligned}
&1 - \int_{-a_n}^{a_n} \pi_\alpha(\beta)d\beta \\
&= 2\int_0^{p^{1+\mu''}} \left(1 - \Phi\left(\frac{a_n}{\lambda\alpha}\right)\right) f_{\mathrm{C}^+}(\lambda)d\lambda + 2\int_{p^{1+\mu''}}^\infty \left(1 - \Phi\left(\frac{a_n}{\lambda\alpha}\right)\right) f_{\mathrm{C}^+}(\lambda)d\lambda \\
&\le 2\left(1 - \Phi\left(\frac{a_n}{\alpha}\cdot p^{-(1+\mu'')}\right)\right) + \int_{p^{1+\mu''}}^\infty \frac{2}{\pi}\frac{1}{1+\lambda^2}d\lambda \\
&\le \frac{\sqrt{2/\pi}e^{-a_n^2 p^{-2(1+\mu'')}/\alpha^2}}{a_n p^{-(1+\mu'')}/\alpha} + p^{-(1+\mu'')} \le p^{-(1+\mu)}
\end{aligned} \tag{3.40}$$

for any $0 < \mu < \mu''$, given that $\alpha < p^{-(2+\mu'')}/\sqrt{n}$ and

$$\int_{p^{1+\mu''}}^{\infty} \frac{2}{\pi}\frac{1}{1+\lambda^2}d\lambda = \frac{2}{\pi}\left(\frac{\pi}{2} - \tan^{-1}(p^{1+\mu''})\right) \asymp p^{-(1+\mu'')}.$$

Since $a_n < \epsilon_n/p$, there exists $0 < \bar{\mu}'' < \mu''$, such that $\dfrac{\eta\epsilon_n/(p-s_*)}{T_p^2\alpha^2} > (1+\bar{\mu}'')\log p$. By the proof of (3.40), we have $\int_{-\eta\epsilon_n/(p-s_*)}^{\eta\epsilon_n/(p-s_*)} \pi_\alpha(\beta)d\beta \geq 1 - p^{-(1+\bar{\mu})}$ for any constant $0 < \bar{\mu} < \bar{\mu}'' < \mu''$. For each β_j following the horseshoe prior (3.39),

$$\int_{\sum_{j=1}^{p-s_*}|\beta_j| \leq \eta\epsilon_n} \prod_{j=1}^{p-s_*} \pi_\alpha(\beta_j)d\beta_1 \cdots d\beta_{p-s_*} \geq \left[\mathrm{P}(|\beta_j| \leq \eta\epsilon_n/(p-s_*))\right]^{p-s_*},$$

which is bounded below by $\left(1 - p^{-(1+\mu)}\right)^{p-s_*} \to 1$. That verifies the condition (3.4).

For (3.5), observe that by the bound (2) of Carvalho and Scott (2010),

$$\inf_{\beta\in[-E,E]} \pi_\alpha(\beta) = \int_0^\infty \frac{1}{\sqrt{2\pi}\lambda\alpha} e^{-E^2/2\lambda_j^2\alpha^2} f_{\mathrm{C}^+}(\lambda)d\lambda \geq \frac{1}{\sqrt{2\pi^3}}\log(1+4\alpha^2/E),$$

which is of the order α^2/E since $\alpha^2/E < p^{-(1+\nu)}$ is small. This gives the desired assertion $-\log\left(\inf_{\beta\in[-E,E]}\pi_\alpha(\beta)\right) \prec (1+\nu)\log p \prec n\epsilon_n^2/s_*$.

3.4 Simulation Study

We perform a simulation study to compare the different Bayesian variable selection techniques in the logistic regression model. We generate $N = 100$ datasets with each dataset consisting $n = 500$ observations and $p = 200$ predictors. The design matrices X are generated from the multivariate normal distribution $\mathrm{N}_p(0, \Sigma)$, where the following two data generating methods are considered:

(1) the covariance matrix is $\Sigma = \mathrm{I}_p$ obtained by mutually independent covariates;

(2) correlated covariates in autoregressive structure $\Sigma_{j,k} = 0.8^{|j-k|}$.

Each covariate is further standardized to have mean zero and standard deviation one. The true values of coefficients are $\beta = (1, 1.5, -2, 2.5, 0, 0, \ldots, 0)^T$ so that only the first 4 covariates are included in the model and the rest of $p - 4$ covariates have zero effects on the response variable. With each simulated design matrix X, the response variable Y is generated from Bernoulli distribution with success probability $\mathrm{P}(Y = 1|X) = \exp(X\beta)/\big(1 + \exp(X\beta)\big)$.

Under each scenario (independent covariates or correlated covariates), Bayesian logistic regression (3.1) is implemented on each simulated data set, with different priors. We examine the performance of the methods corresponding to the following priors on the coefficients β in the logistic regression model:

1. Point mass prior: With probability$1 - b$, β_j is from the point mass distribution at 0; and with probability b, β_j follows normal distribution with mean 0 and variance $\sigma_0^2 = 100^2$. The prior on b is beta distribution $b_j \sim \mathrm{Beta}(a_0, b_0)$ with $a_0 = b_0 = 1/2$.

2. Non-informative normal prior: We assume independently and identically distributed normal prior on $\beta_j \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_0^2)$, where we choose $\sigma_0^2 = 100^2$ for a non-informative distribution.

3. Bayesian LASSO: We implement the Bayesian LASSO estimator in Park and Casella (2008). Each coefficient β_j is given a double exponential prior with parameter $\lambda_j\tau$, where λ_j are independent and identically distributed standard exponential representing local shrinkage and $\tau \sim \mathrm{C}^+(0, 1)$ induces global shrinkage.

4. Dirichlet-Laplace prior: As defined in (3.34), β_j follows a double exponential prior with scale parameter $\phi_j\tau$, where $(\phi_1, \ldots, \phi_p)^T$ are given Dirichlet$(\alpha, \ldots, \alpha)$ distribution and τ is given Gamma$(p\alpha, 2)$ prior. We fix the shrinkage level at $\alpha = 0.1$.

5. Horseshoe prior: As defined in (3.39), we assume that $\beta_j \sim \mathrm{N}(0, \lambda_j^2\alpha^2)$ and λ_j follows the standard half-Cauchy distribution. The hyper-parameter α is also given the standard half-Cauchy prior.

Among the above Bayesian methods, logistic regression model with point mass prior is implemented by R function `logit.spike`$(\cdot)$ in R package `BoomSpikeSlab`. For the logistic regression model using the other priors, we sample the coefficients using the algorithm by Polson et al. (2013) that represent the logit function as a Gaussian mixture distribution. The proposed sampling algorithm introduces latent variables following a Pólya-gamma prior distribution. We use the `rpg.devroye`$(\cdot)$ function in R package `BayesLogit` on the latent variable and combine with Gibbs sampling procedures on other model parameters for each prior specification. After the burn-in period consisting of $10,000$ Markov chain Monte Carlo (MCMC) iterations, we generate $10,000$ posterior samples and save the samples of model parameters at every 5 iterations (that is, we use a thinning factor 5 to reduce autocorrelation).

For each prior, we evaluate the accuracy of the estimates in terms of the $\mathcal{L}_2$-error $\|\beta_* - \hat{\beta}\|_2 = \sqrt{\sum_{j=1}^p (\beta_{*j} - \hat{\beta}_j)^2}$, $\mathcal{L}_1$-error $\|\beta_* - \hat{\beta}\|_1 = \sum_{j=1}^p |\beta_{*j} - \hat{\beta}_j|$ and also monitor the coverage of 95% credible intervals, where the estimates $\hat{\beta}$ are calculated by the posterior means of model parameters. We also compare the variable selection performance using the average proportion of including zero effects (FP) and the average proportion of excluding non-zero effects (FN), where the subset of covariates is selected by calculating the t-statistics and performing the t-test on the posterior samples of model parameters. For

Table 3.1: Summary of the simulation study that compares the Bayesian logistic regression models with point-mass prior, non-informative normal prior (Normal), Bayesian LASSO, Dirichlet-Laplace prior (DL) and horseshoe prior for independent and correlated covariates. The values in parentheses are standard errors.

Independent	Point mass	Normal	Bayes LASSO	DL	Horseshoe
$\mathcal{L}_1$-error	0.74(0.05)	1090.21(6.02)	20.69(2.99)	4.47(0.05)	3.84(0.20)
$\mathcal{L}_2$-error	0.43(0.03)	128.28(0.18)	2.34(0.48)	0.62(0.02)	0.70(0.05)
coverage	99.89%	77.61%	99.03%	99.96%	99.83%
FP	0.00(0.00)	0.20(0.01)	0.01(0.01)	0.00(0.00)	0.01(0.01)
FN	0.01(0.01)	0.00(0.00)	0.01(0.01)	0.01(0.01)	0.00(0.00)
Brier Score	0.10(0.01)	0.19(0.01)	0.12(0.01)	0.10(0.01)	0.10(0.01)
Correlated	Point mass	Normal	Bayes LASSO	DL	Horseshoe
$\mathcal{L}_1$-error	2.24(0.23)	1307.71(13.00)	12.88(0.58)	5.73(0.19)	3.39(0.29)
$\mathcal{L}_2$-error	1.31(0.13)	126.15(0.49)	2.09(0.08)	1.28(0.10)	0.98(0.11)
coverage	99.27%	69.15%	98.95%	99.77%	99.70%
FP	0.00(0.000)	0.29(0.02)	0.00(0.00)	0.00(0.00)	0.00(0.00)
FN	0.25(0.04)	0.00(0.00)	0.26(0.04)	0.35(0.04)	0.15(0.04)
Brier Score	0.13(0.02)	0.22(0.03)	0.14(0.01)	0.13(0.01)	0.13(0.02)

the newly-generated data sets $\{X_i', Y_i'\}_{i=1}^{n'}$ with sample size $n' = 1000$, we monitor the prediction performance using the Brier score defined by $\text{BS} = \sum_{i=1}^{n'}(Y_i' - \hat{\pi}_i')^2/n'$. The estimates of success probability are the posterior means of $\pi_i' = \exp(X_i'\beta)/(1+\exp(X_i'\beta))$ at each iteration.

Table 3.1 summarizes the simulation results for different priors in the logistic regression model. As expected, the point-mass prior achieves the best overall performance in terms of estimation, variable selection and prediction accuracies. However, the disadvantage of a point-mass prior is the heavy computational cost in generating the MCMC samples. The global-local shrinkage priors such as DL and horseshoe priors mimic the composition of a point mass prior through a dominated peak at zero while retaining heavy tails. The simulation study demonstrates that DL and horseshoe priors achieve results comparable to the point-mass prior in estimation, variable selection and prediction ac-

curacies. For the simulated datasets with mutually independent predictors, all five priors identify the non-zero effects correctly in almost all replications. But the point-mass prior has better estimation performance over the others, while the two shrinkage priors (DL and horseshoe) have the average $\mathcal{L}_1$-error and $\mathcal{L}_2$-error that are only slightly worse than that of the point-mass prior. For newly generated testing data sets, the average Brier score for the point-mass prior and the two shrinkage priors are quite close. When the covariates are correlated through an autoregressive structure, similar conclusions can be drawn. The DL prior and the horseshoe prior even outperform the point-mass prior in terms of the $\mathcal{L}_2$-error for estimation accuracy, and the horseshoe prior efficiently distinguishes the significant covariates from the trivial ones so that the best variable selection performance is observed. The shrinkage priors and the point-mass prior also make most accurate predictions on the testing data sets. Therefore, we conclude that the DL prior and the horseshoe prior perform quite like the the point mass prior in the logistic regression model, under both correlated and independent structures on the predictors, with a fraction of the computational cost.

3.5 Conclusion

In this paper, we study the posterior concentration of different priors for variable selection in the logistic regression model. The theoretical results show the consistency properties of the coefficient estimates if the prior distribution satisfies certain properties such that it simultaneously has heavy-tail and significant concentration within a small neighborhood around zero. Specifically, we demonstrate the spike-and-slab prior and some examples of global-local shrinkage priors in the format of $\beta_j \sim \mathrm{N}(0, \lambda_j^2\tau^2)$ in logistic regression model that achieve the posterior consistency when the prior hyper-parameters are under

certain restrictions. The continuous shrinkage priors are more computationally efficient than the traditional point mass prior but can obtain optimal contraction rate that is comparable with that of the point-mass prior in terms of the $\mathcal{L}_2$-norm of the estimation error. The simulation study supports the theoretical results on estimation consistency, variable selection performance and prediction accuracy.

Chapter 4

Spatial Shrinkage Prior for Forecasting Calibration

4.1 Introduction

Rising incidences of large wildland fire call for improved public health risk communication to downwind communities affected by smoke. Exposures to wildland fire smoke are linked to adverse health outcomes ranging from increased rates of cardiorespiratory symptoms, itchy eyes and headaches, to hospitalizations and mortality. Among the pollutants found in smoke, fine particulate matter (PM2.5) is of primary concern to public health. Modeled forecasts of PM2.5 are commonly used to communicate expected conditions over short-term forecasting windows (usually hourly up to 48 hours) and can potentially have large impact on timing and efficacy of health risk communication to reduce public health burden (Rappold et al., 2014). Recent advances in technologies provide for near real time measurements of air quality, as well as wildland fire smoke prediction system hourly forecasts that cover high spatial resolution without missing values. Therefore, the

combination of measurement data with the forecasts from fire smoke prediction system gives an opportunity to calibrate models in real time and improve their interpretation of public health risk messages.

In order to calibrate forecasts with the sparse monitor measurements, we consider statistical downscaling techniques that learn the relationship between measurements and gridded numerical prediction outputs, and make calibrated projections at any spatial location. Over the last decade, a body of literature has emerged on applications of downscaling methods such as linear regression, clustering analysis and neural networks for forecasting calibrations, especially in weather and climate studies. The statistical literature on downscaling methodology is also extensive. One category of statistical downscaling assumes a latent spatial process so that both measurements and numerical prediction outputs are defined in terms of the latent process (Wikle and Berliner, 2005 & Fuentes and Raftery, 2005). Another category regresses the measurements onto the forecast obtained from the prediction system. To account for spatial misalignment of the numerical forecasts and true process, Berrocal, Gelfand and Holland (2010a, 2010b, 2012) and Reich, Chang and Foley (2014) implemented spatially-smoothed summaries of the forecasts as the predictor and showed that this constructed covariate leads to improved prediction accuracy. Spatial regression with smoothed covariate is similar to convolutional neural networks (Lawrence et al., 1997) that redefine the image around each pixel into a vector of multiple elements to be fed into a neural network. Recently, Vandal et al. (2017) developed DeepSD, a statistical downscaling technique that addresses the spatial temporal dependence in the regression model and calibrates the numerical outputs from climate prediction system.

In this project, we extend the work of Berrocal, Gelfand and Holland (2012) that used spatial summaries of the numerical forecasts as constructed covariates in a spatial

regression model. Our proposed approach is well-suited for wildland fire downscaling because it incorporates the PM2.5 forecasts at neighboring cells to diminish the errors from mis-specification of plume direction. If the prediction system makes wrong forecast on wind direction so that the plume will actually affect different spatial locations, the errors between PM2.5 forecasts generated from wildland fire and PM2.5 measurements cannot be addressed by simple linear regression model. Therefore, the spatial summaries of nearby grid cells may improve the downscaling performance. Other contributions of our work are that we assume a multivariate additive nonparametric mean function, spatial variation on each individual effect function, and spatio temporal dependence on the residuals.

The addition of multiple constructed covariates calls for variable selection to avoid overfitting. Spatial variable selection methods often focus on the scalar-on-image regression model that aims to select subregions of an image that are predictive of a scalar response while retaining smoothness of the coefficients across locations. Smith and Kohn (1996) and Smith and Fahrmeir (2007) first introduced the Ising prior for selecting non-zero coefficients with an application to functional Magnetic Resonance Imaging (fMRI). This idea is recently extended by Goldsmith, Huang and Crainiceanu (2014) and Li et al. (2015) to spatial variable selection in high-dimensional neuroimaging studies. Goldsmith, Huang and Crainiceanu (2014) combined the Ising prior with a Gaussian Markov random field prior to control the smoothness of the non-zero coefficients within an image, while Li et al. (2015) proposed the joint Ising-Dirichlet process prior so that spatial variable selection is performed. Using soft-thresholded Gaussian process prior, Kang, Reich and Staicu (2016) proposed the spatial variable selection in scalar-on-image regression to control the sparsity and smoothness of the spatial regression coefficients. Unlike the scalar-on-image regression, the method proposed by Gelfand et al. (2003) allows for separate regression

coefficients at each location and spatially smooth the coefficients. Within this framework, Reich et al. (2010) presented a Bayesian stochastic search algorithm to identify a subset of important variables. Boehm Vock et al. (2015) incorporated local variable selection at each spatial location, but also share information across space for a global variable inclusion. Following the work of Reich et al. (2010), the regression coefficients in our model vary spatially, but the subset of important variables is globally consistent and we use a different prior to achieve more efficient computation and better performance.

In addition to the spatially varying effects, we allow for nonlinear relationships between the constructed covariates and the response. The regression model we consider is similar to the additive model in Nandy, Lim and Maiti (2016), where the additive functions are approximated by B-spline basis expansion with a predetermined number of basis terms. We expand their work by allowing for a spatially varying mean function and using Bayesian methods instead of penalized regression. We incorporate a continuous shrinkage prior to achieve feasible computation. The shrinkage prior mimics the behavior of spike-and-slab priors that assume a non-zero probability of some coefficients being exactly zero and a normal prior distribution for non-zero elements (Mitchell and Beauchamp, 1988), but is a continuous distribution with concentration around zero to shrink trivial terms and heavy tails to allow for large signals. Some examples of shrinkage priors are the Horseshoe prior (Carvalho, Polson and Scott, 2009), generalized double Pareto prior (Armagan, Dunson and Lee, 2012), Dirichlet-Laplace prior (Bhattacharya et al. 2015) and R2D2 prior (Zhang, Reich and Bondell, 2016). Wei et al. (2017) proposed a continuous shrinkage prior in additive nonparametric regression for group shrinkage in B-spline basis coefficients. We develop a spatial extension on the shrinkage prior to allow each nonparametric main effect function vary smoothly across locations and select non-zero effects for all locations.

The remainder of the Chapter proceeds as follows. The wildland fire smoke forecasts and measured PM2.5 data are introduced in Chapter 4.2. The spatially-varying regression model with spatial shrinkage prior is described in Chapter 4.3. Detailed posterior sampling scheme is demonstrated in Chapter 4.4. Chapter 4.5 presents a simulation study that compares our proposed method with others in terms of forecasting accuracy and variable selection. Chapter 4.6 implements our model on the wildland fire smoke data for forecasting calibration.

4.2 Data Description

4.2.1 Monitor and forecast data

The wildland fire smoke prediction system is described in Larkin et al. (2010). Fire activity is based upon aggregated satellite fire detections augmented with other data such as fire size. Fuel information is obtained for each fire location from the Fuel Characteristic Classification System (FCCS, McKenzie et al., 2007 and Prichard et al., 2013) and input into the CONSUME consumption model (Prichard et al., 2005) calculating fuel combustion. Emissions are calculated from the consumed fuel and allocated temporally and vertically in the atmosphere. These emission data are input into the HYSPLIT dispersion model (Draxler and Hess, 1998 and Stein et al., 2015). This wildland fire smoke prediction system operates daily, providing hourly smoke predictions 72-hours into the future across the northwestern US (and other regions) at the resolution of 4km$\times$4km (Figure 4.1). With the forecasts obtained from the wildland fire smoke model predictions, public are informed about what to expect with regard to potential smoke impacts.

The modeled smoke predictions are compared to measured values at $n = 55$ monitors

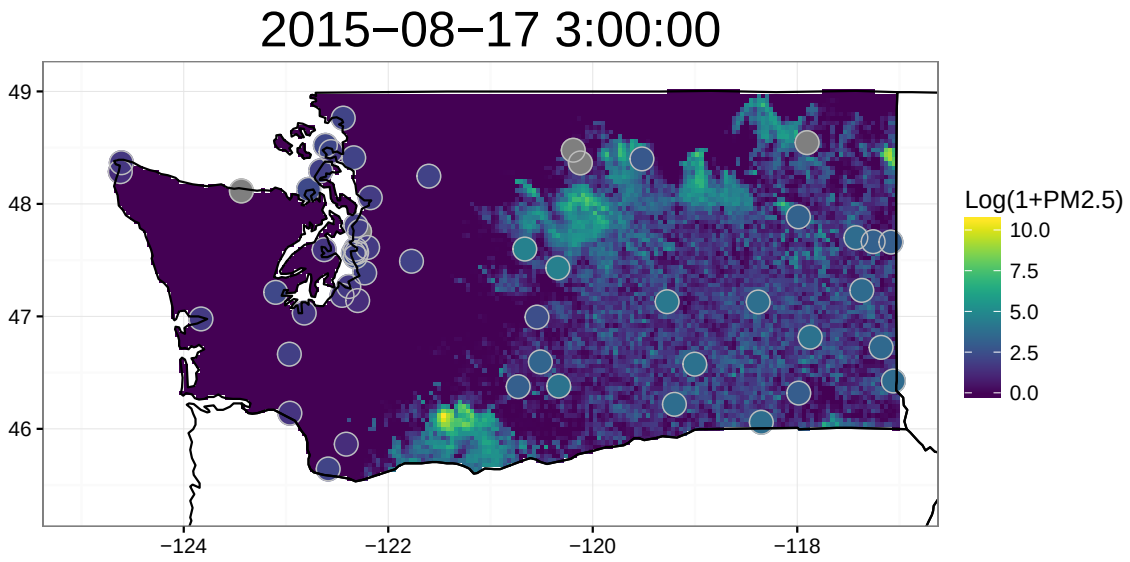


Figure 4.1: Monitor station measurements (circles) and forecasts from wildland fire smoke prediction system (background color) for $\log\left(1 + \text{PM2.5}(\mu g/m^3)\right)$ at 2015-08-17 3:00:00GMT (Greenwich Mean Time). The circles in gray are missing monitor measurements.

in Washington state that record the hourly measurements of PM2.5. For the normality assumption in the regression model, we log-transform both the PM2.5 measurements and the forecast values. We consider time period of August 13 to September 16 in 2015, which includes several large wildland fires. From the forecast data, smoke impacts were mostly east of the Cascade Mountains with the exception of a rare event on August 22-23 when east winds brought smoke into Seattle.

4.2.2 Constructing forecast covariates

Forecasting fire-related PM2.5 can be challenging because a small shift in wind can cause a large error in the forecast for a given location. Therefore, for the PM2.5 measurement at each monitor station, treating the single forecast value at the nearest grid point as

the predictor is insufficient for correcting the bias. We fully utilize the forecast information to capture the impact changes in plume direction by including several local spatial summaries of the forecasts as additional predictors in the additive regression model.

Denote the forecast from the wildland fire smoke prediction system at location $\boldsymbol{s} \in \mathbb{S}$ and time t as $Z_t(\boldsymbol{s})$, where $\mathbb{S}$ defines the entire space. We expand each PM2.5 forecast to a p-vector of constructed covariates summarizing the forecast in the neighborhood of $\boldsymbol{s}$. At location $\boldsymbol{s}_0$ and time t, we first consider the weighted average and weighted maximum over the entire space $\mathbb{S}$ with given range parameter r,

$$\begin{aligned}
\text{Mean}_t(\boldsymbol{s}_0, r) &= \sum_{\boldsymbol{s}_i \in \mathbb{S}, \boldsymbol{s}_i \neq \boldsymbol{s}_0} e^{-\frac{\|\boldsymbol{s}_i - \boldsymbol{s}_0\|}{2r}} Z_t(\boldsymbol{s}_i), \\
\text{Max}_t(\boldsymbol{s}_0, r) &= \max_{\boldsymbol{s}_i \in \mathbb{S}, \boldsymbol{s}_i \neq \boldsymbol{s}_0} e^{-\frac{\|\boldsymbol{s}_i - \boldsymbol{s}_0\|}{2r}} Z_t(\boldsymbol{s}_i).
\end{aligned}$$

Choosing radii $r_1 = 10\text{km}, r_2 = 20\text{km}, \ldots, r_5 = 50\text{km}$, the $p = 11$ covariates are: $X_{t1}(\boldsymbol{s}) = Z_t(\boldsymbol{s})$, $X_{t2}(\boldsymbol{s}) = \text{Mean}_t(\boldsymbol{s}, r_1) - Z_t(\boldsymbol{s})$, $X_{t3}(\boldsymbol{s}) = \text{Max}_t(\boldsymbol{s}, r_1)$, $X_{t4}(\boldsymbol{s}) = \text{Mean}_t(\boldsymbol{s}, r_2) - \text{Mean}_t(\boldsymbol{s}, r_1)$, $X_{t5}(\boldsymbol{s}) = \text{Max}_t(\boldsymbol{s}, r_2)$, etc. For the weighted averages of different radii, we take differences for neighboring range parameters to reduce correlations among covariates, which are between 0.12 and 0.65 after differencing. The covariates are then scaled to unit interval $[0, 1]$ through linear transformation $\frac{X - \min(X)}{\max(X) - \min(X)}$.

4.3 Model Description and Prior Specification

4.3.1 Spatially-varying nonparametric additive model

Let $\mathbb{S}_0 = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$ be the set of spatial points where data are observed. Then for any $\boldsymbol{s} \in \mathbb{S}_0$, $Y_t(\boldsymbol{s})$ is the response measurement at spatial location $\boldsymbol{s}$ and time $t = 1, \ldots, T$, and

the p-dimensional vector $\boldsymbol{X}_t(\boldsymbol{s}) = \big[X_{t1}(\boldsymbol{s}), \ldots, X_{tp}(\boldsymbol{s})\big]^T$ is the corresponding covariates. We assume a spatial regression model with spatially-varying additive main effect function:

$$Y_t(\boldsymbol{s}) = \mu(\boldsymbol{s}) + \sum_{j=1}^{p} f_j\big[X_{tj}(\boldsymbol{s}), \boldsymbol{s}\big] + \varepsilon_t(\boldsymbol{s}), \tag{4.1}$$

where $\mu(\boldsymbol{s})$ is the spatially-varying intercept, $f_j(\cdot, \boldsymbol{s})$ is the main effect function for predictor j and location $\boldsymbol{s}$, and $\varepsilon_t(\boldsymbol{s})$ is the error term. For simplicity, the predictors are all standardized to the unit interval $X_{tj}(\boldsymbol{s}) \in [0, 1]$. Each spatially-varying main effect function is approximated by the B-spline basis expansion with spatially-varying basis coefficients:

$$f_j\big[X_{tj}(\boldsymbol{s}), \boldsymbol{s}\big] \approx \sum_{r=1}^{m} \beta_{jr}(\boldsymbol{s}) B_r\big[X_{tj}(\boldsymbol{s})\big], \tag{4.2}$$

where $B_1[\cdot], \ldots, B_m[\cdot]$ are the m-dimensional B-spline basis functions. For $j = 1, \ldots, p$, the group of basis coefficients $\{\beta_{jr}(\boldsymbol{s})\}_{r=1}^{m}$ are spatially-varying coefficients that capture the shape and spatial variation of the main effect function.

4.3.2 Gaussian process prior with continuous shrinkage

In order to select covariates that are associated with the response across locations, we impose spatial dependence and sparsity in blocks on each m-dimensional basis coefficient for jth covariate at location $\boldsymbol{s}$, denoted as $\boldsymbol{\beta}_j(\boldsymbol{s}) = \big[\beta_{j1}(\boldsymbol{s}), \ldots, \beta_{jm}(\boldsymbol{s})\big]^T$. This can be achieved using a Gaussian process prior with multivariate continuous shrinkage proposed in Wei et al. (2017). We first assume that each spatial function $\beta_{jr}(\boldsymbol{s})$ follows a stationary Gaussian process with mean zero, variance $\sigma^2\lambda_{jr}$, and Matérn spatial correlation with range parameter γ_β and smoothness parameter ν_β, which is denoted as $\text{MGP}(0, \sigma^2\lambda_j, \gamma_\beta, \nu_\beta)$. The variance is a product of the global scale parameter σ^2 and

a local variance factor λ_{jr}, where the term "local" refers to a single covariate rather than a small spatial region. If $\lambda_{jr} = 0$, the basis coefficient $\beta_{jr}(\boldsymbol{s}) = 0$ across all spatial locations. However, simultaneous shrinkage on the m-dimensional basis coefficients $\{\beta_{jr}(\boldsymbol{s})\}_{r=1}^{m}$ for jth predictor is required for variable selection. Therefore, we assume that $\lambda_{j1} = \cdots = \lambda_{jm} = \lambda_j$ and λ_j determines the level of local shrinkage over space. That is, for $j = 1, \ldots, p$ and $r = 1, \ldots, m$,

$$\beta_{jr}(\boldsymbol{s})|\lambda_j, \sigma^2, \gamma_\beta, \nu_\beta \overset{\text{ind}}{\sim} \text{MGP}(0, \sigma^2\lambda_j, \gamma_\beta, \nu_\beta). \tag{4.3}$$

The local variance factor λ_j is given a positive Dirichlet-Laplace prior with hyper-parameter α, defined as in Bhattacharya et al. (2015):

$$\begin{aligned} \lambda_j|\phi_j, \tau &\overset{\text{ind}}{\sim} \text{Exp}(\phi_j\tau),\ j = 1, \ldots, p \\ \boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^T &\sim \text{Dirichlet}(\alpha, \ldots, \alpha),\ \tau \sim \text{Gamma}(p\alpha, 2). \end{aligned} \tag{4.4}$$

The Dirichlet-Laplace prior shrinks null components close to zero, while retaining the property of heavy tails for non-zero components. The hyper-parameter α controls the level of shrinkage so that smaller value of α means a sparser model and higher value of α means a model with more non-zero effects. Thus, we achieve the goal of global variable selection to choose the covariates that influence the joint effects across spatial locations while allowing for spatially-varying individual effects on the response variable.

4.3.3 Spatio temporal residual dependence

In addition to the spatial dependence in the mean function, we consider dependence in the residuals. The residuals are decomposed as $\varepsilon_t(\boldsymbol{s}) \overset{\text{def}}{=} e_t(\boldsymbol{s}) + \varepsilon^*_{\boldsymbol{s},t}$, where $e_t(\boldsymbol{s})$ characterizes

the spatio temporal dependence structure and $\varepsilon^*_{\boldsymbol{s},t} \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$ are independently and identically distributed normal residuals at observed locations and times. We incorporate an autoregressive correlation structure in spatially-correlated residuals terms $e_t(\boldsymbol{s})$. For $t = 1, \ldots, T$, we define $\tilde{e}_t(\boldsymbol{s}) \overset{iid}{\sim} \mathrm{MGP}(0, \sigma_e^2, \gamma_e, \nu_e)$. Then $e_t(\boldsymbol{s})$ is formulated as:

$$\begin{aligned} e_1(\boldsymbol{s}) =& \tilde{e}_1(\boldsymbol{s}); \\ e_t(\boldsymbol{s}) =& \rho e_{t-1}(\boldsymbol{s}) + \sqrt{1-\rho^2}\tilde{e}_t(\boldsymbol{s}), \text{ for } t > 1, \end{aligned} \tag{4.5}$$

where $\rho \in [-1, 1]$ is an unknown autoregressive parameter. Therefore, the proposed spatio temporal dependence structure for residuals will address the temporal correlation in time, as well as the spatial correlation across locations. When $\rho = 0$, the residuals are spatially correlated but independent over time.

4.4 Posterior Sampling

The MCMC sampling scheme for the additive nonparametric model with spatially-varying main effect function and spatio-temporal residuals is given below. The model is formulated as in (3.1) and (4.5), and the individual functions are approximated as B-spline basis expansion in (4.2). We further incorporate the spatially correlated multivariate Dirichlet-Laplace prior on intercept $\mu(\boldsymbol{s})$ and on basis coefficients $\{\beta_{jr}(\boldsymbol{s})\}_{r=1}^m$ for each covariate j, as described in (2.3) and (2.4). We take non-informative Gamma$(0.01, 0.01)$ priors on γ_β and γ_e, while assuming fixed values for $\nu_\beta = 2$ and $\nu_e = 2$. For temporal correlation in residuals, the autoregressive coefficient ρ is given a uniform prior on $[-1, 1]$. The global variance terms σ^2 and σ_e^2 are given InvGamma$(0.01, 0.01)$ prior. The hyper-parameter α in Dirichlet prior is set to equal to 0.5 for moderate shrinkage.

For the observed spatial locations as $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n \in \mathbb{S}_0$, then the model parameters written as a function of $\mathbb{S}_0$ indicate n-dimensional vectors with elements at each location $\boldsymbol{s}_i$. Metropolis-Hasting sampling used for γ_β, γ_e and ρ and Gibbs sampling is used for the rest of model parameters. The steps of the MCMC algorithm are:

1. Generate $\mu(\mathbb{S}_0) \stackrel{\text{def}}{=} \big(\mu(\boldsymbol{s}_1), \ldots, \mu(\boldsymbol{s}_n)\big)^T$ from multivariate normal distribution with mean $(\Sigma^{-1}_{\gamma_\beta,\nu_\beta}/\lambda_0 + \mathrm{I}_n)^{-1} \sum_{t=1}^T R_t(\mathbb{S}_0)$ and variance $\sigma^2(\Sigma^{-1}_{\gamma_\beta,\nu_\beta}/\lambda_0 + \mathrm{I}_n)^{-1}$, where we denote the residuals as $R_t(\mathbb{S}_0) = Y_t(\mathbb{S}_0) - \sum_{j=1}^p \sum_{r=1}^m \beta_{jr}(\mathbb{S}_0) B_r\big[X_{tj}(\mathbb{S}_0)\big] - e_t(\mathbb{S}_0)$.

2. Generate $\beta_{jr}(\mathbb{S}_0) \stackrel{\text{def}}{=} \big(\beta_{jr}(\boldsymbol{s}_1), \ldots \beta_{jr}(\boldsymbol{s}_n)\big)^T$, $j = 1, \ldots, p$ and $r = 1, \ldots, m$, from multivariate normal distribution with mean $(\Sigma^{-1}_{\gamma_\beta,\nu_\beta}/\lambda_j + \mathrm{I}_n)^{-1} \sum_{t=1}^T D\big[B_r(X_{tj}(\mathbb{S}_0))\big] R^*_{tj}(\mathbb{S}_0)$ and covariance matrix $\sigma^2(\Sigma^{-1}_{\gamma_\beta,\nu_\beta}/\lambda_j + \mathrm{I}_n)^{-1}$, where the residuals are denoted as $R^*_{tj}(\mathbb{S}_0) = Y_t(\mathbb{S}_0) - \mu(\mathbb{S}_0) - \sum_{j' \neq j} \sum_{r' \neq r} \beta_{j'r'}(\mathbb{S}_0) B_{r'}\big[X_{tj'}(\mathbb{S}_0)\big] - e_t(\mathbb{S}_0)$ and $D\big[B_r(X_{tj}(\mathbb{S}_0))\big]$ is the diagonal matrix for $B_r(X_{tj}(\mathbb{S}_0))$.

3. Generate the model parameters in Dirichlet-Laplace prior, following the procedures in Bhattacharya et al. (2015), with extension to spatially-correlated model:

   - For $j = 1, \ldots, p$, generate λ_j from generalized inverse Gaussian distribution $\mathrm{GiG}\big(1 - \frac{mn}{2}, \frac{2}{\phi_j \tau}, \frac{1}{2\sigma^2} \sum_{r=1}^m \beta_{jr}(\mathbb{S}_0)^T \Sigma^{-1}_{\gamma_\beta,\nu_\beta} \beta_{jr}(\mathbb{S}_0)\big)$.
   - Given λ_j's and ϕ_j's, we generate τ from generalized inverse Gaussian distribution $\mathrm{GiG}\big(p(\alpha - 1), 1, 2\sum_{j=1}^p \lambda_j/\phi_j\big)$.
   - For $j = 1, \ldots, p$, we independently generate T_j from generalized inverse Gaussian distribution $\mathrm{GiG}(\alpha - 1, 1, 2\lambda_j)$ and then $\phi_j = T_j / \sum_{l=1}^p T_l$.

4. For the autoregressive term $e_t(\mathbb{S}_0) \stackrel{\text{def}}{=} \big(e_t(\boldsymbol{s}_1), \ldots, e_t(\boldsymbol{s}_n)\big)^T$:

   - If $t = 1$, generate $\big(e_1(\boldsymbol{s}_1), \ldots, e_1(\boldsymbol{s}_n)\big)^T$ from multivariate normal distribution with mean $(\Sigma^{-1}_{\gamma_e,\nu_e}/\sigma_e^2 + \mathrm{I}_n/\sigma^2)^{-1} R'_t(\mathbb{S}_0)/\sigma^2$ and variance $(\Sigma^{-1}_{\gamma_e,\nu_e}/\sigma_e^2 + \mathrm{I}_n/\sigma^2)^{-1}$,

- If $1 < t < T$, generate $\big(e_t(\boldsymbol{s}_1), \ldots, e_t(\boldsymbol{s}_n)\big)^T$ from multivariate normal distribution with mean $\big(\frac{1+\rho^2}{(1-\rho^2)\sigma_e^2}\Sigma^{-1}_{\gamma_e,\nu_e} + \mathrm{I}_n/\sigma^2\big)^{-1}\big[R'_t(\mathbb{S}_0) + \rho\frac{\Sigma^{-1}_{\gamma_e,\nu_e}}{(1-\rho^2)\sigma_e^2}(e_{t-1}(\mathbb{S}_0) + e_{t+1}(\mathbb{S}_0))\big]/\sigma^2$ and variance $\big(\frac{1+\rho^2}{(1-\rho^2)\sigma_e^2}\Sigma^{-1}_{\gamma_e,\nu_e} + \mathrm{I}_n/\sigma^2\big)^{-1}$,
- If $t = T$, generate $\big(e_T(\boldsymbol{s}_1), \ldots, e_T(\boldsymbol{s}_n)\big)^T$ from multivariate normal distribution with mean $\big(\frac{1}{(1-\rho^2)\sigma_e^2}\Sigma^{-1}_{\gamma_e,\nu_e} + \mathrm{I}_n/\sigma^2\big)^{-1}\big[R'_t(\mathbb{S}_0) + \rho\frac{\Sigma^{-1}_{\gamma_e,\nu_e}}{(1-\rho^2)\sigma_e^2}e_{t-1}(\mathbb{S}_0)\big]/\sigma^2$ and variance $\big(\frac{1}{(1-\rho^2)\sigma_e^2}\Sigma^{-1}_{\gamma_e,\nu_e} + \mathrm{I}_n/\sigma^2\big)^{-1}$,

where $R'_t(\mathbb{S}_0) = Y_t(\mathbb{S}_0) - \mu(\mathbb{S}_0) - \sum_{j=1}^p\sum_{r=1}^m \beta_{jr}(\mathbb{S}_0)B_r\big[X_{tj}(\mathbb{S}_0)\big]$.

5. Define $R^0_t(\mathbb{S}_0) = Y_t(\mathbb{S}_0) - \mu(\mathbb{S}_0) - \sum_{j=1}^p\sum_{r=1}^m \beta_{jr}(\mathbb{S}_0)B_r\big[X_{tj}(\mathbb{S}_0)\big] - e_t(\mathbb{S}_0)$, the conditional posterior distribution for σ^{-2} given the rest of parameters is Gamma distribution with shape parameter $\frac{pmn+n+nT}{2} + 0.01$ and rate parameter $\frac{1}{2}\sum_{t=1}^T \|R^0_t(\mathbb{S}_0)\|_2^2 + \frac{1}{2}\mu(\mathbb{S}_0)^T\Sigma^{-1}_{\gamma_\beta,\nu_\beta}\mu(\mathbb{S}_0)/\lambda_0 + \frac{1}{2}\sum_{j=1}^p\sum_{r=1}^m \beta_{jr}(\mathbb{S}_0)^T\Sigma^{-1}_{\gamma_\beta,\nu_\beta}\beta_{jr}(\mathbb{S}_0)/\lambda_j + 0.01$.

6. For the residual variance σ_e^2 in autoregressive terms, the conditional posterior distribution for σ_e^{-2} is Gamma distribution with shape parameter $\frac{nT}{2} + 0.01$, rate parameter $\frac{1}{2}e_1(\mathbb{S}_0)^T\Sigma^{-1}_{\gamma_e,\nu_e}e_1(\mathbb{S}_0) + \frac{1}{2(1-\rho^2)}\sum_{t=2}^T\big(e_t(\mathbb{S}_0) - \rho e_{t-1}(\mathbb{S}_0)\big)^T\Sigma^{-1}_{\gamma_e,\nu_e}\big(e_t(\mathbb{S}_0) - \rho e_{t-1}(\mathbb{S}_0)\big) + 0.01$.

7. For the model parameters ρ, γ_β and γ_e, we generate posterior samples using Metropolis-Hasting algorithm.

4.5 Simulation Study

4.5.1 Simulation design

In the simulation study, we consider $T = 365$ observations at $n = 100$ locations $\boldsymbol{s} = \{(s_1, s_2), s_1, s_2 = 1, \ldots, 10\}$. We simulate the covariate matrix $\boldsymbol{X}$ with $p = 10$ covariates

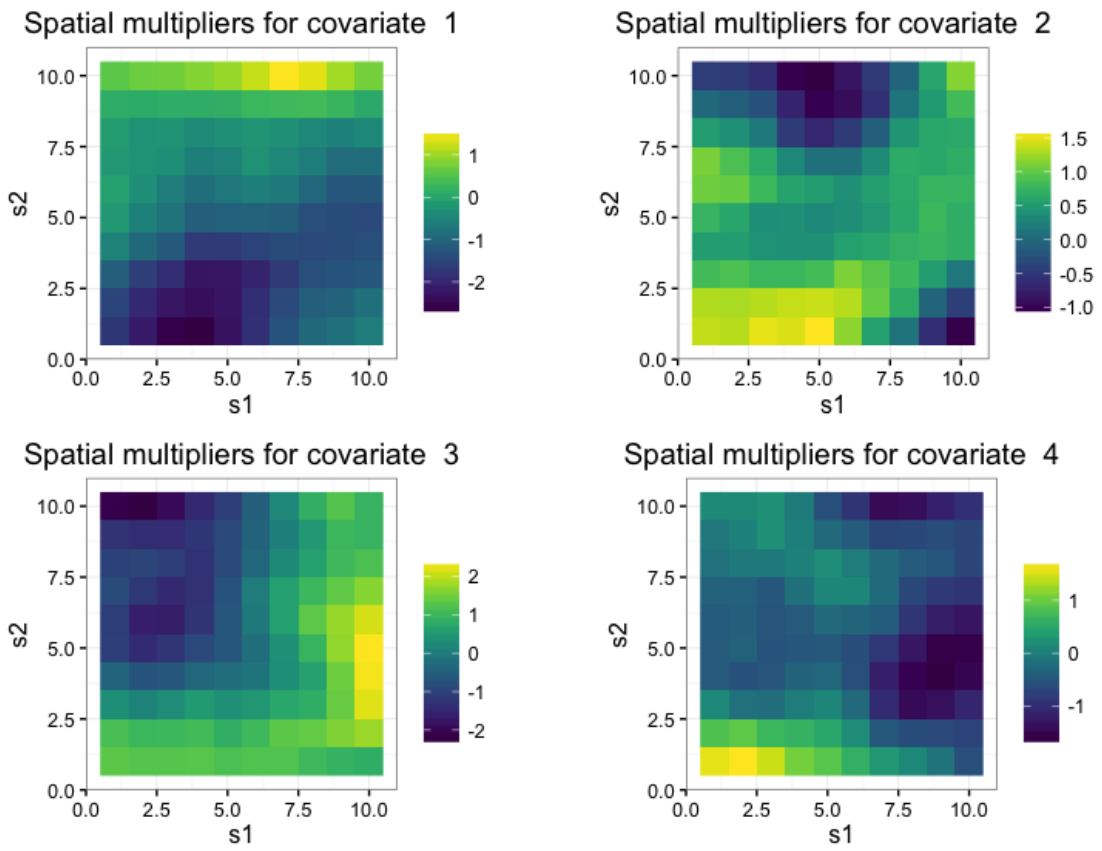


Figure 4.2: One example of the spatially-varying multipliers $\delta_j(\boldsymbol{s})$ for non-zero covariates in the simulation study.

and $T \times n$ observations using the same technique as in Nandy, Lim and Maiti (2016), such that $X_{tj}(\boldsymbol{s}) = (W_{tj}(\boldsymbol{s}) + t_0 U_t)/(1 + t_0)$, where $W_{tj}(\boldsymbol{s})$ and U_t are independently and identically distributed with a Uniform$[0, 1]$ distribution for $j = 1, \ldots, p$. Therefore, the simulated covariates $X_{tj}(\boldsymbol{s})$ and $X_{tk}(\boldsymbol{s})$ are correlated with correlation $t_0^2/(1+t_0^2)$, for any $j \neq k$. For this simulation study, we consider the case when $t_0 = 1$.

Next, the response variable $Y_t(\boldsymbol{s})$ is generated as:

$$Y_t(\boldsymbol{s}) = \sum_{j=1}^{4} \delta_j(\boldsymbol{s}) f_j\big[X_{tj}(\boldsymbol{s})\big] + \varepsilon_t(\boldsymbol{s}), \tag{4.6}$$

where the residuals $\varepsilon_t(\boldsymbol{s})$ are either independent, $\varepsilon_t(\boldsymbol{s}) \overset{iid}{\sim} \mathrm{N}(0, \sigma_0^2)$, or spatially correlated, $\varepsilon_t(\boldsymbol{s}) \overset{ind}{\sim} \mathrm{MGP}(0, 1.5^2, 1, 2)$. Following the simulation settings in Nandy, Lim and Maiti (2016), we choose the first 4 covariates as non-zero components, with $f_1(x) = 5x$, $f_2(x) = 3(2x-1)^2$, $f_3(x) = 4\sin(2\pi x)\big/\big[2 - \sin(2\pi x)\big]$ and $f_4(x) = 0.6\sin(2\pi x) + 1.2\cos(2\pi x) + 1.8\sin^2(2\pi x) + 2.4\cos^3(2\pi x) + 3.0\sin^3(2\pi x)$. The spatial variation in the additive effects is determined by the spatially-varying multipliers $\delta_j(\boldsymbol{s})$. For each simulated data set, we generate $\delta_j(\boldsymbol{s}) \sim \mathrm{MGP}(0, 1, 1, 2)$ for $j = 1, \ldots, 4$ and then generate the data sets $\{X_{t1}(\boldsymbol{s}), \ldots X_{tp}(\boldsymbol{s}), Y_t(\boldsymbol{s})\}$. Figure 4.2 shows one example of the spatially-varying multipliers $\delta_j(\boldsymbol{s})$. We generate $N = 50$ data sets under each scenario (independent or spatially-correlated residuals).

4.5.2 Methods for comparison

For each simulated data set, we separate the observations into a training set of the first 300 observations, and evaluate prediction performance using the remaining 65 observations. Under each scenario, we make different assumptions on the regression model so that we can compare the effects on model performance:

1. Non-spatial regression mean function: $f\big[X_{tj}(\boldsymbol{s}), \boldsymbol{s}\big] = f\big[X_{tj}(\boldsymbol{s})\big]$ vs spatially-varying regression mean function $f\big[X_{tj}(\boldsymbol{s}), \boldsymbol{s}\big]$;

2. Linear function: $f\big[X_{tj}(\boldsymbol{s}), \boldsymbol{s}\big] = \beta_j(\boldsymbol{s})X_{jt}(\boldsymbol{s})$ with $\beta_j(\boldsymbol{s})|\lambda_j, \sigma^2, \gamma_\beta, \nu_\beta \sim \mathrm{MGP}(0, \sigma^2\lambda_j, \gamma_\beta, \nu_\beta)$ vs nonlinear function: $f\big[X_{tj}(\boldsymbol{s}), \boldsymbol{s}\big] \approx \sum_{r=1}^{m} \beta_{jr}(\boldsymbol{s})B_r\big[X_{tj}(\boldsymbol{s})\big]$ with $m = 5$ basis functions and $\beta_{jr}(\boldsymbol{s})|\lambda_j, \sigma^2, \gamma_\beta, \nu_\beta \sim \mathrm{MGP}(0, \sigma^2\lambda_j, \gamma_\beta, \nu_\beta)$;

3. Mutually independent residuals: $\varepsilon_t(\boldsymbol{s}) \overset{\triangle}{=} \varepsilon_{\boldsymbol{s},t} \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$ vs spatial dependence in residuals as described in (4.5) with $\rho = 0$: $\varepsilon_t(\boldsymbol{s}) = e_t(\boldsymbol{s}) + \varepsilon_{\boldsymbol{s},t}$ where $e_t(\boldsymbol{s}) \overset{ind}{\sim} \mathrm{MGP}(0, \sigma^2, \gamma_e, \nu_e)$.

All eight methods include Dirichlet-Laplace prior for variable selection, adjusted to the assumptions made on the regression model. Since no exact zeros will be generated for each $f_j(\cdot, \boldsymbol{s})$, a thresholding policy is needed to distinguish the non-zero components from the zero components. We follow the method in Hahn and Carvalho (2015) and choose a subset of covariates so that the "variation-explained" in the reduced forecasting model can be controlled in a tolerated range. Specifically, within each posterior sampling step, we pick the k covariates with largest value of ϕ_j, $j = 1, \ldots, p$. Then we have $\beta_{jr}^{(k)}(\boldsymbol{s}) = \beta_{jr}(\boldsymbol{s})$ for selected non-zero j's but $\beta_{jr}^{(k)}(\boldsymbol{s}) = 0$ for all the j's not being selected. The "variation-explained" values are calculated as

$$V_{(k)} = \frac{\left\| \sum_{j=1}^{p} \sum_{r=1}^{m} B_r\big[X_j(\boldsymbol{s})\big]\beta_{jr}(\boldsymbol{s}) \right\|^2}{\left\| \sum_{j=1}^{p} \sum_{r=1}^{m} B_r\big[X_j(\boldsymbol{s})\big]\beta_{jr}(\boldsymbol{s}) \right\|^2 + n\sigma^2 + \left\| \sum_{j=1}^{p} \sum_{r=1}^{m} B_r\big[X_j(\boldsymbol{s})\big](\beta_{jr}(\boldsymbol{s}) - \beta_{jr}^{(k)}(\boldsymbol{s})) \right\|^2},$$

where $X_j(\boldsymbol{s})$ is the vector of jth covariate observations at all locations and all times. We select the number of variables to include by choosing the smallest number of selected covariates k so that the 80% interval of the "variation-explained" metric $V_{(k)}$ contains the posterior mean of "variation-explained" metric $V_{(p)}$ under the full model. This technique aims to select a parsimonious model with acceptable fidelity to the full model.

Given Dirichlet-Laplace prior on λ_j, $j = 1, \ldots, p$, as in (4.4), non-informative priors on model parameters $\sigma^{-2}, \gamma_\beta, \gamma_e \sim \text{Gamma}(0.01, 0.01)$ and fixed model parameters $\alpha = 0.5, \nu_\beta = \nu_e = 2$, the posterior samples are generated using the steps described in the Appendix. For each simulated data and each method, we generated in total of $10,000$ iterations with a burnin period of $3,000$ iterations. In addition to the spatial regression model, we also fit non-spatial component selection and smoothing operator ("COSSO") of Lin and Zhang (2006) and the multivariate adaptive regression splines model ("MARS") of Friedman (1991).

The prediction performance is evaluated on the remaining $T_1 = 65$ days in terms of mean squared error, MSE $= \frac{1}{nT_1}\sum_{t=1}^{T_1}\sum_{\boldsymbol{s}_i\in\mathbb{S}_0}\left[\hat{Y}_t(\boldsymbol{s}_i) - Y_t(\boldsymbol{s}_i)\right]^2$, and mean absolute deviation, MAD $= \frac{1}{nT_1}\sum_{t=1}^{T_1}\sum_{\boldsymbol{s}_i\in\mathbb{S}_0}\left|\hat{Y}_t(\boldsymbol{s}_i) - Y_t(\boldsymbol{s}_i)\right|$. For the test data that includes T_1 observation times, $\hat{Y}_t(\boldsymbol{s})$ is the predicted value from fitted models:

$$\hat{Y}_t(\boldsymbol{s}) = \hat{\mu}(\boldsymbol{s}) + \sum_{j=1}^{p}\sum_{r=1}^{m}\hat{\beta}_{jr}(\boldsymbol{s})B_r\big[X_{tj}(\boldsymbol{s})\big] + \hat{e}_t(\boldsymbol{s}), \tag{4.7}$$

where $\hat{\mu}(\boldsymbol{s})$, $\hat{\beta}_{jr}(\boldsymbol{s})$ and $\hat{e}_t(\boldsymbol{s})$ are the posterior means of model parameters obtained from model fitting in the training data. For selecting variables based on the null hypothesis that $H_0 : f_j\big(\cdot, \boldsymbol{s}\big) = 0$ at every location $\boldsymbol{s}$, we compare Type I and Type II error for identifying non-zero effects, where Type I error is the average percentage of zero components being selected as non-zero, Type II error is the average percentage of missing non-zero covariates in the model. We also show the percentage of times in the simulation study when the correct subset of covariates is selected, defined as "Truth".

4.5.3 Simulation results

Tables 4.1 and 4.2 summarize the simulation results for data generated with independent and spatially-correlated errors, respectively. Under both scenarios, we found that the model with nonlinear spatial mean function preforms the best in terms of both forecasting and variable selection, as the simulated data are generated from additive nonlinear functions with spatial variation.

Assuming spatially-varying coefficients in the regression model, no matter if it is linear or nonlinear, contributes to the forecasting accuracy most significantly. For the non-spatial mean function models, even the incorporation of nonlinearity and spatially-correlated residuals does not improve the forecasting performance. Furthermore, with

Table 4.1: Summary of the simulation study with independent residuals. The methods vary by whether the additive functions are linear or nonlinear, whether the coefficients $\boldsymbol{\beta}(\boldsymbol{s})$ change spatially or not, and whether the errors $\varepsilon_t(\boldsymbol{s})$ are independent or spatially-correlated. All methods are compared in terms of prediction mean squared error (MSE), prediction mean absolute deviation (MAD), Type I and Type II errors for selection important covariates, and the proportion of data sets for which true model is selected (Truth). The standard errors are in the parentheses.

Coefficients	Residuals	Function	MSE	MAD	Type I	Type II	Truth
Non-spatial	Independent	Linear	17.04(1.35)	3.23(0.71)	0.04	0.57	0.26
Non-spatial	Independent	Nonlinear	16.21(2.11)	3.13(0.62)	0.02	0.49	0.32
Non-spatial	Spatial	Linear	17.05(1.88)	3.22(0.57)	0.08	0.38	0.20
Non-spatial	Spatial	Nonlinear	16.24(1.29)	3.13(0.63)	0.06	0.47	0.30
Spatial	Independent	Linear	7.45(1.06)	2.08(0.35)	0.06	0.27	0.52
Spatial	Independent	Nonlinear	**2.68(0.53)**	**1.30(0.31)**	0.00	0.02	**0.92**
Spatial	Spatial	Linear	7.48 (1.90)	2.10(0.39)	0.06	0.13	0.40
Spatial	Spatial	Nonlinear	**2.67(0.67)**	**1.32(0.27)**	0.00	0.05	**0.88**
COSSO by Lin and Zhang (2006)			19.28(3.20)	4.011(1.27)	0.33	0.40	0.08
MARS by Friedman (1991)			20.15(4.01)	4.238(0.94)	0.42	0.39	0.06

Table 4.2: Summary of the simulation study with spatially-correlated residuals. The methods vary by whether the coefficients $\boldsymbol{\beta}(\boldsymbol{s})$ change spatially or not, whether the errors $\varepsilon_t(\boldsymbol{s})$ are independent or spatially-correlated, and whether the additive functions are linear or nonlinear. All methods are compared in terms of prediction mean squared error (MSE), prediction mean absolute deviation (MAD), Type I and Type II errors for selection important covariates, and the proportion of data sets for which true model is selected (Truth). The standard errors are in the parentheses.

Coefficients	Residuals	Function	MSE	MAD	Type I	Type II	Truth
Non-spatial	Independent	Linear	17.03(2.57)	3.25(0.68)	0.17	0.31	0.24
Non-spatial	Independent	Nonlinear	16.32(3.01)	3.15(0.61)	0.23	0.29	0.28
Non-spatial	Spatial	Linear	17.03(3.31)	3.24(0.51)	0.22	0.28	0.22
Non-spatial	Spatial	Nonlinear	16.34(2.92)	3.15(0.44)	0.18	0.26	0.24
Spatial	Independent	Linear	7.56(1.29)	2.11(0.38)	0.06	0.25	0.48
Spatial	Independent	Nonlinear	**3.87(0.52)**	**1.96(0.30)**	0.02	0.08	**0.76**
Spatial	Spatial	Linear	8.59(1.26)	2.60(0.41)	0.04	0.19	0.56
Spatial	Spatial	Nonlinear	**3.22(0.71)**	**1.43(0.33)**	0.00	0.04	**0.84**
COSSO by Lin and Zhang (2006)			20.57(4.69)	4.29(0.95)	0.55	0.33	0.06
MARS by Friedman (1991)			21.83(4.81)	4.41(1.22)	0.49	0.40	0.10

spatially-varying mean functions, nonlinearity improves the MSE and MAD by more than 50%, especially when the response variable is actually generated from spatially-correlated residuals. We also observe only small differences between the results under different residual assumptions. Thus, when the mean function includes spatial dependence through the coefficients, the assumption of independent residuals in regression model still retains good prediction performance.

For the variable selection performance, the assumption of spatially-varying coefficients is again the most important factor in identifying the correct group of non-zero covariates, while spatially-correlated residuals also improve the variable selection results when the data is generated with correlated residuals. This shows that correctly specifying the residual spatial dependence is more meaningful in determining the non-zero effects of covariates than making predictions on the joint effects. When the coefficients are correctly specified, nonlinearity benefits variable selection performance as well since the true model has nonlinear functions.

4.6 Implementation to Wildland Fire Smoke Forecasting

We implement the proposed statistical downscaling method including $p = 11$ constructed covariates to calibrate the PM2.5 forecasts with the monitoring measurements described in Chapter 4.2. Both univariate regression and multivariate regression are considered. In univariate regression model, we keep the PM2.5 forecast at closest grid cell to each monitoring station as the single predictor. For the multivariate regression model, smooth averages and maximums around each closest grid cell are combined with the forecast as

described in Chapter 4.2. The first 720 hours (from 2015-08-13 00:00:00GMT to 2015-09-11 23:00:00GMT) are used for model training and the remaining 120 hours (from 2015-09-12 00:00:00GMT to 2015-09-16 23:00:00GMT) are held out to evaluate the prediction.

4.6.1 Computational details

The downscaling process is performed on the spatial summaries of forecast data and the monitoring measurements at each monitor station. We implement the additive nonparametric model as discussed in Chapter 4.3 with Bayesian hierarchical models for spatially-varying effects. Thus, each constructed covariate is allowed to influence the PM2.5 values in additive nonlinear functions with spatial correlations. The spatial dependence structure is characterized by Gaussian Process prior on the basis coefficients $\beta_{jr}(\boldsymbol{s})$ and the spatial intercept $\mu(\boldsymbol{s})$ with Matérn correlation structure. For each basis coefficients, we incorporate an additional variance factor with Dirichlet-Laplace prior for continuous shrinkage. The residuals are defined by spatio temporal dependence as in (4.5), where the autoregressive coefficient ρ is given uniform prior on $[-1, 1]$ and Matérn correlation for spatial dependence with non-informative inverse gamma prior on σ^2. Similar as the simulation studies, we assume non-informative gamma priors on Matérn range parameters and with a fixed value of 2 on Matérn smooth parameters. Following the MCMC steps in Chapter 4.4, we generate $20,000$ posterior steps with thinning period of 10, after $20,000$ burn-in steps.

4.6.2 Model comparisons

For univariate and multivariate regression models, we compare the same eight models as the simulation study except we consider autoregressive dependence on residuals for all

Table 4.3: Calibrated forecast comparisons for univariate and multivariate regression models under different model assumptions: the mean function is linear or nonlinear; the coefficients are constant across locations or spatially-varying; the residuals are spatially independent or spatially correlated. The comparison metrics are MSE, MAD, bias, average variance and coverage of 95% credible intervals.

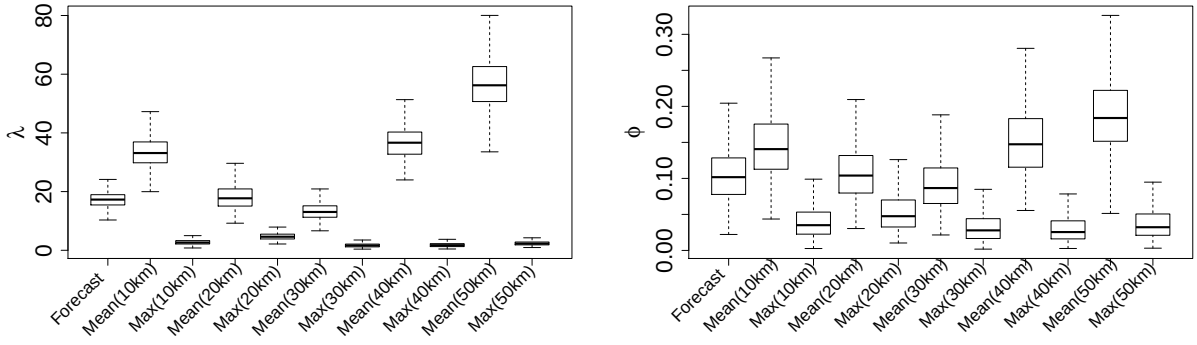
	Coefficients	Residuals	Model	MSE	MAD	Bias	Variance	Coverage
Univariate model	Non-spatial	Independent	Linear	0.295	0.416	0.064	0.036	0.787
	Non-spatial	Independent	Nonlinear	0.313	0.424	0.043	0.086	0.779
	Non-spatial	Spatial	Linear	0.290	0.421	0.056	0.033	0.761
	Non-spatial	Spatial	Nonlinear	0.307	0.430	0.137	0.072	0.783
	Spatial	Independent	Linear	0.274	0.373	-0.101	0.104	0.855
	Spatial	Independent	Nonlinear	0.280	0.357	-0.118	0.089	0.819
	Spatial	Spatial	Linear	0.271	0.369	-0.164	0.091	0.893
	Spatial	Spatial	Nonlinear	0.288	0.376	0.193	0.092	0.875
Multivariate model	Non-spatial	Independent	Linear	0.293	0.401	-0.449	0.154	0.801
	Non-spatial	Independent	Nonlinear	0.270	0.395	-0.343	0.149	0.825
	Non-spatial	Spatial	Linear	0.289	0.386	0.417	0.122	0.843
	Non-spatial	Spatial	Nonlinear	0.267	0.381	0.210	0.128	0.863
	Spatial	Independent	Linear	0.255	0.362	-0.199	0.140	0.949
	Spatial	Independent	Nonlinear	**0.248**	**0.345**	-0.200	0.107	**0.952**
	Spatial	Spatial	Linear	0.250	0.356	-0.237	0.116	0.958
	Spatial	Spatial	Nonlinear	**0.243**	**0.348**	-0.205	0.108	**0.963**

models. Table 4.3 shows the results of MSE, MAD, bias, average prediction variance and coverage of 95% credible sets averaged over the last 120 hours and all monitoring sites.

Treating the log-transformed measurements of PM2.5 as response variable, univariate linear regression model assuming independence across locations is treated as the baseline. The additive nonparametric model with spatially-varying effects and spatio temporal residuals that performs the best in the overall prediction evaluation improves the prediction accuracy by 17.7% in MSE and 16.3% in MAD. The 95% credible intervals for the full model cover around 18% more observations than the baseline. Excluding the spatial dependence on residuals in nonparametric regression model achieves comparable prediction accuracy. Overall, multivariate regression models outperform the univariate counterparts in terms of MSE, MAD and coverage. This could be because averaging over

nearby grid cells accounts for forecast errors in the plume direction. For the spatial dependence, including spatially-varying coefficients benefits the prediction accuracy more than the spatio temporal residuals, but both lead to more precise predictions. Furthermore, the assumption of nonlinearity does not improve the prediction performance in univariate regression model, which suggests linear pattern in the association between closest point forecast and the true measurement. But nonlinear mean function in multivariate regression model contributes the prediction slightly under both spatial and non-spatial cases.

4.6.3 Summarizing the calibration results



(a) (b)

Figure 4.3: Box plots of the posteriors of λ_j (left) and ϕ_j (right) $j = 1, \ldots, p$, in multivariate nonparametric regression model. "Forecast" is the forecast at the closest grid point to the monitoring station, covariates with "Mean(r)" are the differences of weighted averages in rings with current range parameter $r$km and previous range parameter $(r-10)$km, covariates with "Max(r)" are the weighted maximum in the rings with difference range parameters.

We summarize the full model with additive nonparametric mean function of multiple

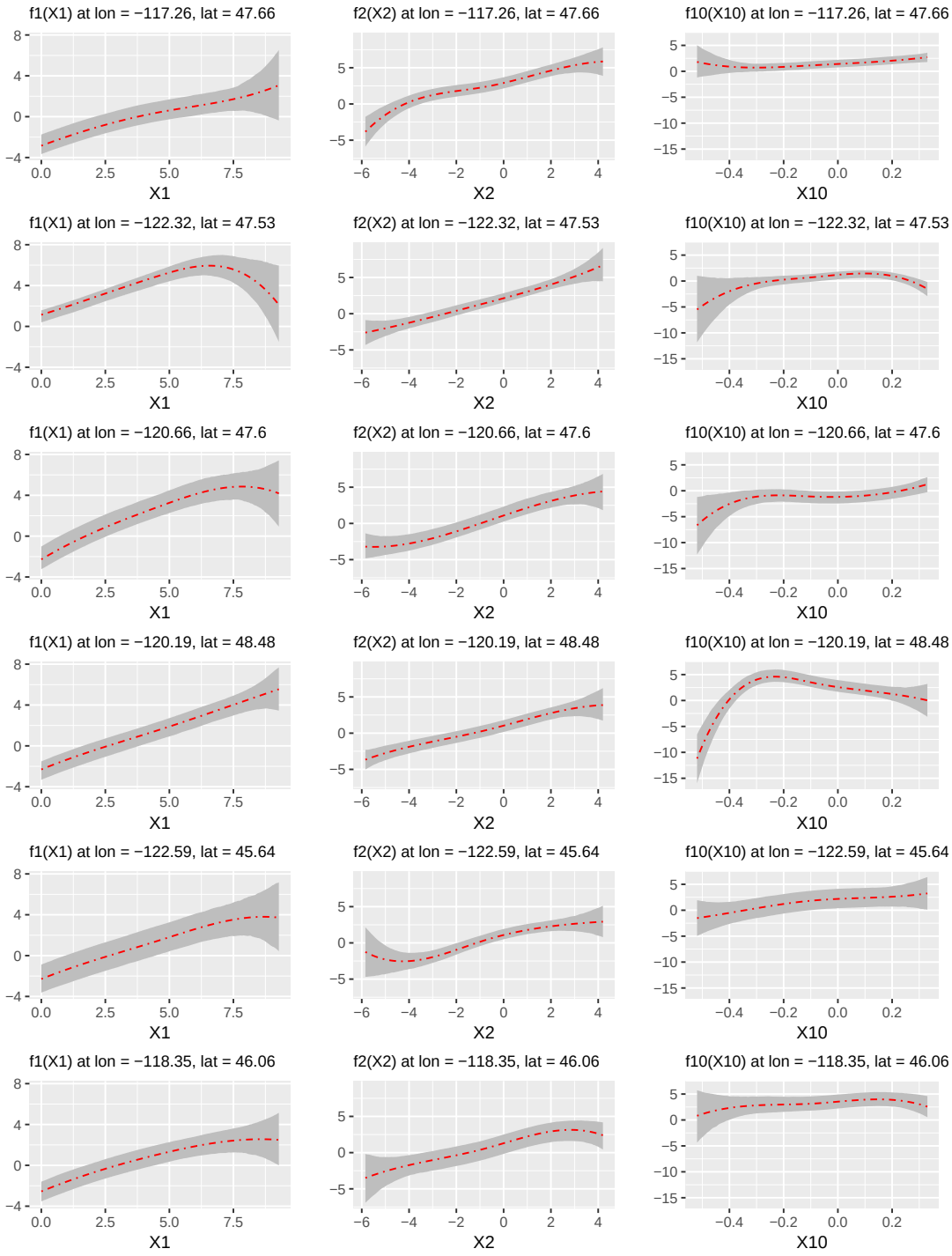


Figure 4.4: Posterior means (red) and 95% credible intervals (grey) of the main effect curves $f_j(X, \boldsymbol{s})$, for significant predictors (X_1: closest point forecast; X_2: difference of weighted average for $r_1 = 10$km and closest point forecasts; X_{10}: difference of weighted averages for $r_5 = 50$km and $r_4 = 40$km) at different locations (latitude and longitude shown in the title).

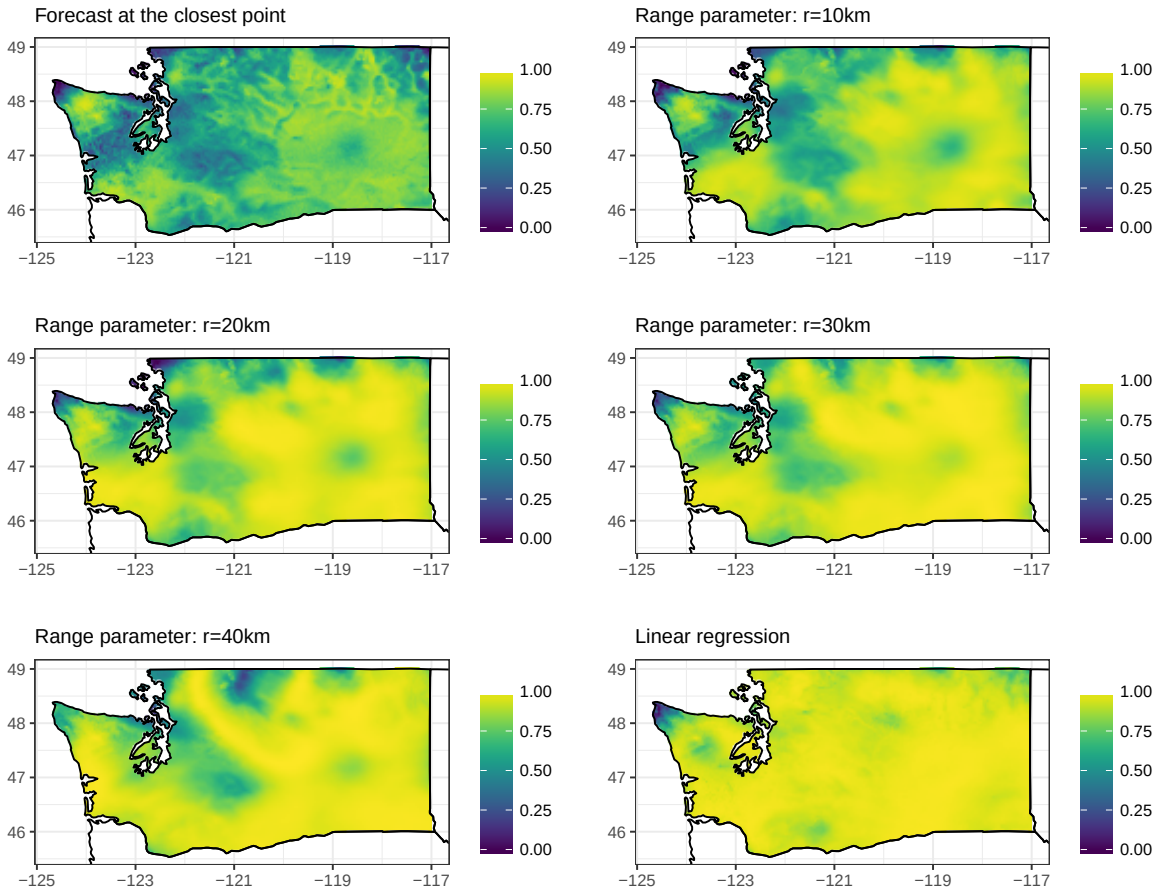


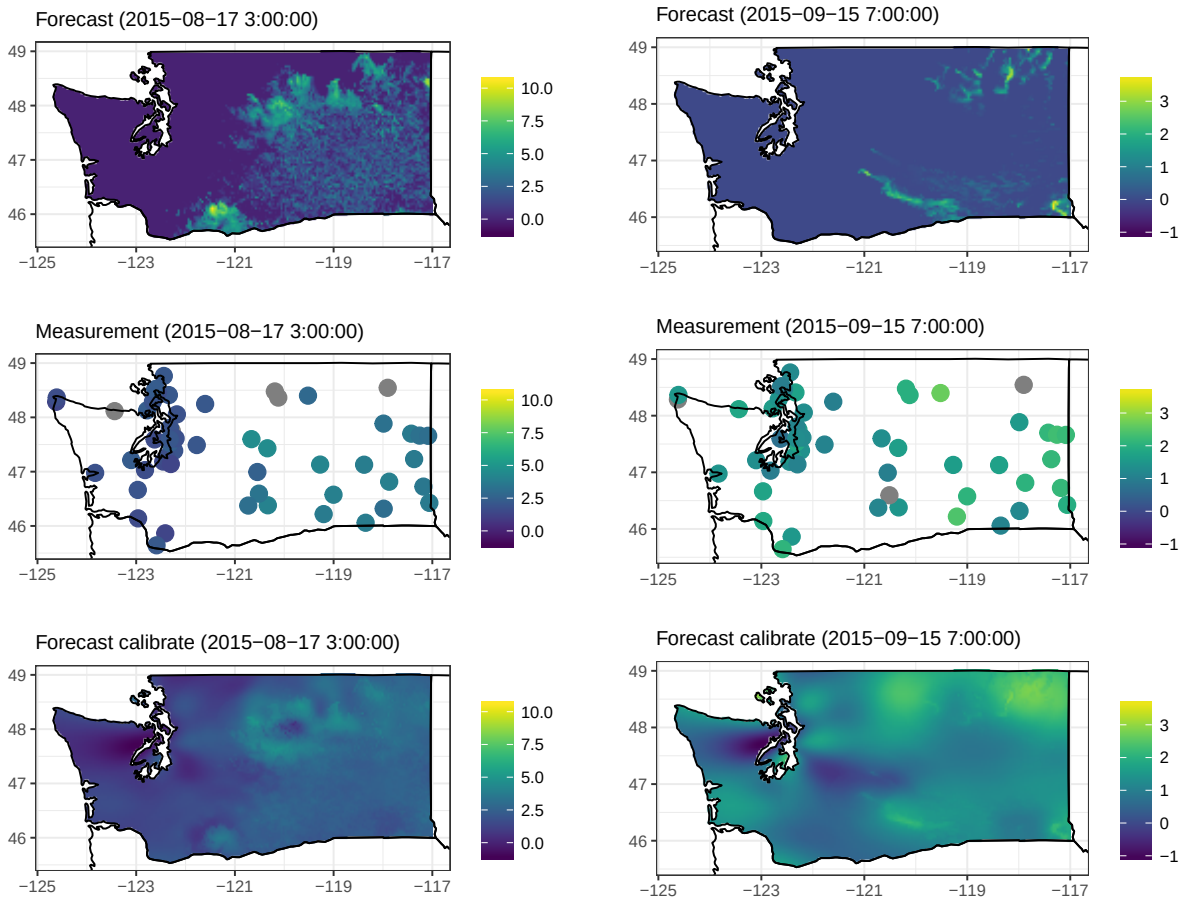
Figure 4.5: The maps of absolute correlations between calibrated forecasts under full model and the reduced model with included covariates up to the indicated ranger parameter r. The bottom right plot is the correlation map for calibrated forecasts using linear and nonlinear regression models.

covariates, spatially-varying basis coefficients and spatio temporal residual dependence. The box plots in Figure 4.3 show the posterior samples of the local variance factor λ_j defined in (4.3) and variance proportions ϕ_j. Terms with large λ_j and ϕ_j are the most influential. The box plots suggest that the forecasts at the closest grid point, as well as the difference of the weighted average of forecasting values under neighboring range parameters are the most significant covariates. Surprisingly, the difference of the weighted

averages for range parameters $r_4 = 40$km and $r_5 = 50$km contributes more to the fitted model than the forecast value at closest grid point to the monitoring station. It may be that this large-scale averaging reduces the effects of forecast error in the direction of the plume emanating from the wildland fire.

Choosing the three most influential predictors (the closest point forecast $X_{t1}(\boldsymbol{s})$, the difference, $X_{t2}(\boldsymbol{s})$, between the closest point forecast and weighted average of $r_1 = 10$km and the difference, $X_{t10}(\boldsymbol{s})$, between the weighted averages of $r_4 = 40$km and $r_5 = 50$km), Figure 4.4 shows the estimated nonlinear main effect curves with point-wise 95% credible intervals at six randomly chosen locations. In order to reflect the main effect of each covariate, the x-axis scales for the chosen three covariates are the original covariate scales before transforming to the unit interval. Thus, these functions for each covariate are not in the same scale. For these six locations, the curves are different in terms of both shape and range. The functions f_1 and f_2 follow similar pattern across locations except differences on both ends where a nonlinear trend appears at some locations. For function f_{10}, the differences in shapes across the entire space are more obvious with some locations have an increasing trend while some other locations have a decreasing trend. The plots also suggests a nonlinear relationship in the main effect curves, which supports the improved prediction performance of the nonlinear regression model.

In order to reflect the effects of adding smoothed statistics (average and maximum) at different range parameters $r = 10, 20, \ldots, 50$, Figure 4.5 plots the absolute correlations between the regression mean function of the full model and the model that includes only covariates of smooth averages and smooth maximums up to the indicated range parameter. For example, at the correlation map with $r = 20$km, we calculate the absolute correlations between the full model regression mean function and the summation of individual effect functions for the closest point forecasts, spatial smooth averages and smooth



(a) (b)

Figure 4.6: Top: Forecasted values from fire smoke prediction system at 2015-08-17 03:00:00GMT and at 2015-09-15 07:00:00GMT; middle: PM2.5 measurements from monitoring stations; bottom: Calibrated forecasts using nonparametric spatial-temporal model with smooth covariate matrix. All the values are in the scale of $\log\left(1+\text{PM2.5}(\mu g/m^3)\right)$.

maximums for range parameters $r = 10$km and 20km. The last map in Figure 4.5 is the absolute correlations at each grid cell between the regression mean functions of nonlinear model and linear model using all 11 covariates. When we only include the forecasts at closest point and spatial summary for range parameter $r = 10$km, some regions in the eastern Washington are quite well-explained, but larger radii of smoothing are needed in

the rest of the state. As more range parameters are included, the correlations are more than 0.9 in most of the Washington state, except for the south eastern part of the Puget Sound where the absolute correlations are around 0.75. According to the last plot in Firgure 4.5, the correlation between linear and nonlinear models is greater than 0.9 in most of the Washington state. This demonstrates that spatial smoothness appears to be more important than nonlinearity assumption.

Figure 4.6 shows the data and calibrated forecasts at 2015-08-17 03:00:00 (left column) and 2015-09-15 07:00:00 (right column). The first selected hour is in the training data and the second in the test data that is 80 hours separated from the last observation in the training set. The top plot is the output from wildland fire smoke prediction system, and the bottom one is the calibrated forecasts from the full regression model. We compare that with the sparsely-distributed true PM2.5 measurements at monitoring stations in the middle. Our proposed model calibrates the forecast output obtained from the wildland fire smoke prediction system with true measurements and smooth out the PM2.5 forecasts by incorporating the forecast values at nearby locations and by allowing nonlinearity and spatial temporal dependence, which differs from the sharp contrasts on the map of wildland fire smoke prediction output.

4.7 Conclusion

This project is motivated by the problem of calibrating PM2.5 forecast from wildland fire smoke in Washington state. We incorporate spatial summaries of forecasts in the neighborhood grid cells to build a multivariate covariate matrix. Our proposed model implement additive nonparametric mean functions for robust model assumption and Bayesian variable selection with spatial extension for spatially-varying effects across locations. The

novel multivariate spatial Dirichlet-Laplace prior on B-spline basis coefficients to account for both the spatial dependence and the nonlinearity on regression model. We examine the performance of different model assumptions in the simulation study by evaluating the performance in prediction accuracy and variable selection. For the wildland fire prediction system outputs at Washington State, we demonstrate the PM2.5 forecasts after calibration have smaller MSE and better coverage than competing methods. The numerical results support the advantages of our proposed model over other competing methods, while addressing the nonlinear associations and spatial dependence structure.

REFERENCES

[1] Armagan, A., Dunson, D. B., and Lee, J. Generalized double Pareto shrinkage. *Statistica Sinica 23*, 1 (2013), 119.

[2] Armagan, A., Dunson, D. B., Lee, J., Bajwa, W. U., and Strawn, N. Posterior consistency in linear models under shrinkage priors. *Biometrika 100*, 4 (2013), 1011–1018.

[3] Atchadé, Y. F. On the contraction properties of some high-dimensional quasi-posterior distributions. *arXiv preprint arXiv:1508.07929* (2017).

[4] Belitser, E., and Nurushev, N. Needles and straw in a haystack: robust empirical Bayes confidence for possibly sparse sequences. *arXiv preprint arXiv:1511.01803* (2015).

[5] Berrocal, V. J., Gelfand, A. E., and Holland, D. M. A bivariate space-time downscaler under space and time misalignment. *The annals of applied statistics 4*, 4 (2010), 1942.

[6] Berrocal, V. J., Gelfand, A. E., and Holland, D. M. A spatio-temporal downscaler for output from numerical models. *Journal of agricultural, biological, and environmental statistics 15*, 2 (2010), 176–197.

[7] Berrocal, V. J., Gelfand, A. E., and Holland, D. M. Space-time data fusion under error in computer model output: An application to modeling air quality. *Biometrics 68*, 3 (2012), 837–848.

[8] Bhadra, A., Datta, J., Polson, N. G., and Willard, B. The Horseshoe+ estimator of ultra-sparse signals. *Bayesian Analysis* (2017).

[9] Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. Dirichlet–Laplace priors for optimal shrinkage. *Journal of the American Statistical Association 110*, 512 (2015), 1479–1490.

[10] Bobb, J. F., Valeri, L., Claus Henn, B., Christiani, D. C., Wright, R. O., Mazumdar, M., Godleski, J. J., and Coull, B. A. Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics 16*, 3 (2014), 493–508.

[11] Boehm Vock, L. F., Reich, B. J., Fuentes, M., and Dominici, F. Spatial variable selection methods for investigating acute health effects of fine particulate matter components. *Biometrics 71*, 1 (2015), 167–177.

[12] Carvalho, C. M., Polson, N. G., and Scott, J. G. Handling sparsity via the Horseshoe. In *AISTATS* (2009), vol. 5, pp. 73–80.

[13] Carvalho, C. M., Polson, N. G., and Scott, J. G. The Horseshoe estimator for sparse signals. *Biometrika 97*, 2 (2010), 465–480.

[14] Castillo, I., Schmidt-Hieber, J., and Van der Vaart, A. Bayesian linear regression with sparse priors. *The Annals of Statistics 43*, 5 (2015), 1986–2018.

[15] Castillo, I., and van der Vaart, A. Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics 40*, 4 (2012), 2069–2101.

[16] Curtis, S. M., Banerjee, S., and Ghosal, S. Fast Bayesian model assessment for nonparametric additive regression. *Computational Statistics & Data Analysis 71* (2014), 347–358.

[17] Draxler, R. R., and Hess, G. An overview of the HYSPLIT_4 modelling system for trajectories. *Australian meteorological magazine 47*, 4 (1998), 295–308.

[18] Friedman, J. H. Multivariate adaptive regression splines. *The annals of statistics* (1991), 1–67.

[19] Fuentes, M., and Raftery, A. E. Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. *Biometrics 61*, 1 (2005), 36–45.

[20] Gelfand, A. E., Kim, H.-J., Sirmans, C., and Banerjee, S. Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association 98*, 462 (2003), 387–396.

[21] George, E. I., and McCulloch, R. E. Variable selection via Gibbs sampling. *Journal of the American Statistical Association 88*, 423 (1993), 881–889.

[22] Ghosal, S. Normal approximation to the posterior distribution for generalized linear models with many covariates. *Mathematical Methods of Statistics 6*, 3 (1997), 332–348.

[23] Ghosal, S. Asymptotic normality of posterior distributions in high-dimensional linear models. *Bernoulli 5*, 2 (1999), 315–331.

[24] Ghosal, S. Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity. *Journal of Multivariate Analysis 74*, 1 (2000), 49–68.

[25] Ghosal, S., Ghosh, J. K., and Van Der Vaart, A. W. Convergence rates of posterior distributions. *Annals of Statistics 28*, 2 (2000), 500–531.

[26] Ghosal, S., and van der Vaart, A. *Fundamentals of nonparametric Bayesian inference*, vol. 44. Cambridge University Press, 2017.

[27] Goldsmith, J., Huang, L., and Crainiceanu, C. M. Smooth scalar-on-image regression via spatial Bayesian variable selection. *Journal of Computational and Graphical Statistics 23*, 1 (2014), 46–64.

[28] Griffin, J. E., and Brown, P. J. Inference with Normal-Gamma prior distributions in regression problems. *Bayesian Analysis 5*, 1 (2010), 171–188.

[29] Hahn, P. R., and Carvalho, C. M. Decoupling shrinkage and selection in Bayesian linear models: a posterior summary perspective. *Journal of the American Statistical Association 110*, 509 (2015), 435–448.

[30] Ishwaran, H., and Rao, J. S. Spike and slab variable selection: frequentist and Bayesian strategies. *Annals of statistics* (2005), 730–773.

[31] Kang, J., Reich, B. J., and Staicu, A.-M. Scalar-on-image regression via the soft-thresholded Gaussian process. *arXiv preprint arXiv:1604.03192* (2016).

[32] Larkin, N. K., ONeill, S. M., Solomon, R., Raffuse, S., Strand, T., Sullivan, D. C., Krull, C., Rorig, M., Peterson, J., and Ferguson, S. A. The BlueSky smoke modeling framework. *International Journal of Wildland Fire 18*, 8 (2010), 906–920.

[33] Lawrence, S., Giles, C. L., Tsoi, A. C., and Back, A. D. Face recognition: A convolutional neural-network approach. *IEEE transactions on neural networks 8*, 1 (1997), 98–113.

[34] Lee, K.-J., Jones, G. L., Caffo, B. S., and Bassett, S. S. Spatial Bayesian variable selection models on functional magnetic resonance imaging time-series data. *Bayesian Analysis (Online) 9*, 3 (2014), 699.

[35] Li, F., Zhang, T., Wang, Q., Gonzalez, M. Z., Maresh, E. L., and Coan, J. A. Spatial Bayesian variable selection and grouping for high-dimensional scalar-on-image regression. *The Annals of Applied Statistics 9*, 2 (2015), 687–713.

[36] Lin, Y., and Zhang, H. H. Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics 34*, 5 (2006), 2272–2297.

[37] Linkletter, C., Bingham, D., Hengartner, N., Higdon, D., and Ye, K. Q. Variable selection for Gaussian process models in computer experiments. *Technometrics 48*, 4 (2006), 478–490.

[38] Makalic, E., and Schmidt, D. F. High-dimensional Bayesian regularised regression with the BayesReg package. *arXiv preprint arXiv:1611.06649* (2016).

[39] Martin, R., Mess, R., and Walker, S. G. Empirical Bayes posterior concentration in sparse high-dimensional linear models. *Bernoulli 23*, 3 (2017), 1822–1847.

[40] McKenzie, D., Raymond, C., Kellogg, L.-K., Norheim, R., Andreu, A., Bayard, A., Kopper, K., and Elman, E. Mapping fuels at multiple scales: landscape application of the Fuel Characteristic Classification System. *Canadian Journal of Forest Research 37*, 12 (2007), 2421–2437.

[41] Mitchell, T. J., and Beauchamp, J. J. Bayesian variable selection in linear regression. *Journal of the American Statistical Association 83*, 404 (1988), 1023–1032.

[42] Nandy, S., Lim, C. Y., and Maiti, T. Additive model building for spatial regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* (2016).

[43] Norets, A., and Pati, D. Adaptive Bayesian estimation of conditional densities. *Econometric Theory 33*, 4 (2017), 980–1012.

[44] Park, T., and Casella, G. The Bayesian LASSO. *Journal of the American Statistical Association 103*, 482 (2008), 681–686.

[45] Polson, N. G., Scott, J. G., and Windle, J. Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American statistical Association 108*, 504 (2013), 1339–1349.

[46] Prichard, S., Ottmar, R., and Anderson, G. Consume users guide version 3.0.. *USDA Forest Service, Pacific Wildland Fire Sciences Laboratory.(Seattle, WA)* (2005).

[47] Prichard, S. J., Sandberg, D. V., Ottmar, R. D., Eberhardt, E., Andreu, A., Eagle, P., and Swedin, K. Fuel Characteristic Classification System Version 3.0: Technical documentation.

[48] Rappold, A. G., Fann, N. L., Crooks, J., Huang, J., Cascio, W. E., Devlin, R. B., and Diaz-Sanchez, D. Forecast-based interventions can reduce the health and economic burden of wildfires. *Environmental science & technology 48*, 18 (2014), 10571–10579.

[49] Reich, B. J., Chang, H. H., and Foley, K. M. A spectral method for spatial downscaling. *Biometrics 70*, 4 (2014), 932–942.

[50] Reich, B. J., Fuentes, M., Herring, A. H., and Evenson, K. R. Bayesian variable selection for multivariate spatially varying coefficient regression. *Biometrics 66*, 3 (2010), 772–782.

[51] Reich, B. J., Storlie, C. B., and Bondell, H. D. Variable selection in Bayesian smoothing spline ANOVA models: Application to deterministic computer codes. *Technometrics 51*, 2 (2009), 110–120.

[52] Savitsky, T., Vannucci, M., and Sha, N. Variable selection for nonparametric Gaussian process priors: Models and computational strategies. *Statistical science: a review journal of the Institute of Mathematical Statistics 26*, 1 (2011), 130.

[53] Shen, W., and Ghosal, S. Adaptive Bayesian density regression for high-dimensional data. *Bernoulli 22*, 1 (2016), 396–420.

[54] Smith, M., and Fahrmeir, L. Spatial Bayesian variable selection with application to functional magnetic resonance imaging. *Journal of the American Statistical Association 102*, 478 (2007), 417–431.

[55] Smith, M., and Kohn, R. Nonparametric regression using Bayesian variable selection. *Journal of Econometrics 75*, 2 (1996), 317–343.

[56] Song, Q., and Liang, F. Nearly optimal Bayesian shrinkage for high dimensional regression. *preprint* (2017).

[57] Starks, S. E., Gerr, F., Kamel, F., Lynch, C. F., Alavanja, M. C., Sandler, D. P., and Hoppin, J. A. High pesticide exposure events and central nervous system function among pesticide applicators in the Agricultural Health Study.

International archives of occupational and environmental health 85, 5 (2012), 505–515.

[58] Starks, S. E., Hoppin, J. A., Kamel, F., Lynch, C. F., Jones, M. P., Alavanja, M. C., Sandler, D. P., and Gerr, F. Peripheral nervous system function and organophosphate pesticide use among licensed pesticide applicators in the Agricultural Health Study. *Environmental health perspectives 120*, 4 (2012), 515.

[59] Stein, A., Draxler, R. R., Rolph, G. D., Stunder, B. J., Cohen, M., and Ngan, F. NOAAs HYSPLIT atmospheric transport and dispersion modeling system. *Bulletin of the American Meteorological Society 96*, 12 (2015), 2059–2077.

[60] Van Der Pas, S., Kleijn, B., and Van Der Vaart, A. The Horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics 8*, 2 (2014), 2585–2618.

[61] van der Pas, S., Szabó, B., and van der Vaart, A. How many needles in the haystack? adaptive inference and uncertainty quantification for the Horseshoe. *arXiv preprint arXiv:1607.01892* (2016).

[62] Vandal, T., Kodra, E., Ganguly, S., Michaelis, A., Nemani, R., and Ganguly, A. R. DeepSD: Generating high resolution climate change projections through single image super-resolution. *preprint arXiv:1703.03126* (2017).

[63] Wei, R., Reich, B. J., Hoppin, J. A., and Ghosal, S. Sparse Bayesian nonparametric regression with application to health effects of pesticides mixtures. *preprint*.

[64] Wikle, C. K., and Berliner, L. M. Combining information across spatial scales. *Technometrics 47*, 1 (2005), 80–91.

[65] Wood, S., Kohn, R., Shively, T., and Jiang, W. Model selection in spline nonparametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*, 1 (2002), 119–139.

[66] Yang, Y., and Tokdar, S. T. Minimax-optimal nonparametric regression in high dimensions. *The Annals of Statistics 43*, 2 (2015), 652–674.

[67] Yoo, W. W., and Ghosal, S. Supremum norm posterior contraction and credible sets for nonparametric multivariate regression. *The Annals of Statistics 44*, 3 (2016), 1069–1102.

[68] Zhang, Y., Reich, B. J., and Bondell, H. D. High dimensional linear regression via the R2-D2 shrinkage prior. *preprint arXiv:1609.00046* (2016).