

Bayesian Variable Selection: Theory and Applications

By
Zuofeng Shang

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY
(STATISTICS)

at the
UNIVERSITY OF WISCONSIN – MADISON
2011

© Copyright by Zuofeng Shang 2011

All Rights Reserved

Abstract

My thesis contains two parts, both related to Bayesian variable selection (BVS). In part I, I introduce some theoretical results related to BVS. Linear models with a growing number of parameters have been widely used in modern statistics. One important problem about this kind of model is the variable selection issue. Bayesian approaches, which provide a stochastic search of informative variables, have gained popularity. Here, we studied the asymptotic properties related to BVS when the model dimension is growing with the sample size. We provide sufficient conditions under which the posterior probability of the true model converges to one. This will guarantee that the true model will be selected under a Bayesian framework.

In part II, I introduce the application of BVS in spatial concurrent linear models. Spatial concurrent linear models, in which the model coefficients are spatial processes varying at a local level, are flexible and useful tools for analyzing spatial data. One approach places stationary Gaussian process priors on the spatial processes, but in applications the data may display strong nonstationary patterns. Here, I propose a Bayesian variable selection approach based on wavelet tools to address this problem. The proposed approach does not involve any stationarity assumptions on the priors, and instead I impose a mixture prior distribution directly on each wavelet coefficient. I introduce an option to control the priors such that high resolution coefficients are more likely to be zero. Computationally efficient MCMC procedures are provided to address posterior sampling, and uncertainty in the estimation is assessed through posterior samples. Examples based on simulated data demonstrate the estimation accuracy and advantages of the proposed method. I also illustrate the performance of the proposed method for real data obtained through remote sensing.

Acknowledgments

I would like to express my sincere and deep thanks to my advisor Professor Murray Clayton, for his introduction of the beauty of statistics and for his encouragement, patience and inspiration on my research work. During my PhD study, he carefully supervised my research progress, and provided clear research guidance. His solid fundamental research skills and excellent personality have been deeply rooted in my heart.

I would also like to express my most sincere gratitude to Professor Kam Tsui who gave me kind encouragement and job opportunities during my PhD study, to Professor Jun Shao who kindly gave me many valuable suggestions on the theoretical parts of my thesis, to Professor Bret Larget who kindly hosted my preliminary exam, and to Professor Jun Zhu for a nice Friday study group in which I learned a lot about spatial statistics, to Professor Rick Nordheim for many valuable suggestions on my statistical consulting experience.

Finally, I wish to express my deep thanks to my parents, Zhenjia Shang and Chendan Liu, who gave me long supports before and after I came to Madison, and to my wife, Fan Yang, who supports and encourages me a lot on my research.

Contents

Abstract	i
Acknowledgments	ii
1 Overview	1
2 Consistency of Bayesian Linear Model Selection With a Growing Number of Parameters	13
2.1 Introduction	13
2.2 Preliminaries and main results	16
2.3 Generalizations to g -prior settings	28
2.4 Numerical results	29
2.5 Discussion	32
2.6 Appendix A: Proofs of the results in Section 2.2	33
2.7 Appendix B: Generalizations of Bayesian consistency to ultra-high dimensional settings	43
2.8 Appendix C: Proof of Corollaries 2.2.5 and 2.2.6	51
2.9 Appendix D: Almost Sure Consistency of $p(\gamma^0 Z)$	56
3 An Application of Bayesian Variable Selection to Spatial Concurrent Linear Models	60
3.1 Models and Algorithms	60
3.2 Numerical Results	66

	iv
3.2.1 Assessing the Performance of Models I and II	67
3.2.2 Detecting Where the Slopes Are Nonzero	71
3.2.3 Applications to Gypsy Moth Defoliation Data	75
3.3 Discussion	77
3.4 Appendix: Sampler Derivations for Algorithm I.	79
4 Conclusions And Future Work	86
Bibliography	94

List of Tables

1	Means and standard deviations of the posterior probabilities of the true model	31
2	Simulation results based on Model I with Algorithm I for case I	71
3	Simulation results based on Model I with Algorithm I for Case II	72
4	Simulation results based on Model II with Algorithm II for Case I	72
5	Simulation results based on Model II with Algorithm II for Case II	73

List of Figures

1	Gypsy moth defoliation data	2
2	One-dimensional Haar wavelet	5
3	Quad-tree structure of wavelet coefficients	8
4	Images of \hat{B} and the PSD of B	82
5	Choropleth map of B based on simulations	83
6	Some graphs on real data analysis	84
7	Choropleth map of B for defoliation rate data	85

Chapter 1

Overview

In this chapter, I briefly describe the motivation of this work, introduce the models and approaches applied in my research work, and summarize several relevant articles.

One objective in data analysis is to study the relationship between explanatory (input) and response (output) variables through building a reasonable model. A special example is remotely sensed (satellite) data in which both input and output are images. In particular, I consider the gypsy moth defoliation rate data which was obtained by satellite from a region in the Appalachian Mountains in June-July 2006. Townsend (2004) provided algorithms to generate the satellite images. In this data, the response is the defoliation rate image which records the gypsy moth defoliation rates of oak trees, and the inputs are elevation and species composition images. Here, I only focus on the elevation image. The defoliation rate and elevation images can be found in Figure 1 below. In these images, low to high data values are represented by black to white tones. Data were taken at each pixel in the image. It is of particular interest to link the information from the output image to the input images. Several authors have observed that gypsy moth defoliation rate increases with elevation (Kleiner and Montgomery 1994). Therefore, a linear model may be preferred due to its convenience in implementation and interpretation.

One simple approach assumes the regression parameters to be constant across the region,

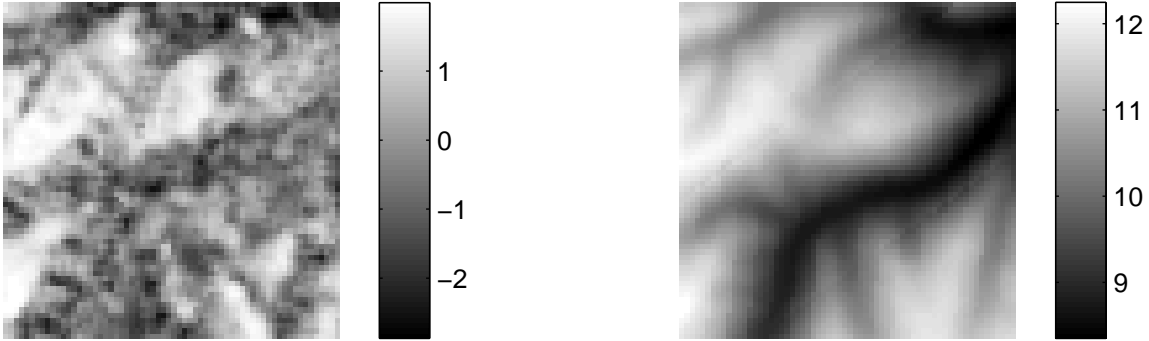


Figure 1: *Images of gypsy moth defoliation data. The left panel is the image of centered-and-scaled defoliation rate and the right panel is the image of scaled elevation (the mean and the standard deviation of the original defoliation rates are 371.9290 and 314.5869, and the elevation was scaled by its standard deviation 61.7952 m). The defoliation rate and elevation respectively represent the proportion of defoliated forest and height on a per-pixel basis, and both of the images have 30m pixel resolution. The defoliation rate data were obtained through Landsat satellite imaging and the elevation data were obtained from the National Elevation Data set of the US Geological Survey. The light color in the defoliation rate image represents high defoliation rates while the black color represents low defoliation rates.*

that is, at each location (or pixel) \mathbf{s}_i , the response y_i and input \mathbf{x}_i have a linear relationship

$$y_i = a + \mathbf{x}_i \mathbf{b} + \epsilon_i, \quad i = 1, \dots, n, \quad (1.1)$$

with ϵ_i normally distributed and a and \mathbf{b} constant. The parameters a and \mathbf{b} represents the global linear relationship between y_i and \mathbf{x}_i .

However, model (1.1) is not appropriate when there is no global linearity in the data. One example is the house price data discussed by Agarwal *et al.* (2003). In this data, the response is the house price which relates to covariates such as age of house, square feet of living area, number of bathrooms, etc. Typically, coefficients associated with these covariates in cities will be different from those in rural areas. One simple approach is to use a piecewise linear model. More explicitly, the whole region could be divided into several subregions, and

a linear model fit in each subregion, such that the model parameter vectors vary according to subregions. In fact, Agarwal et al. (2003) divides the whole region by school districts, and linear models are fitted in each of these districts.

In order to fit a piecewise linear model, we need to carefully define the subregions. However, there are no natural subregions for the defoliation rate data, and it is not obvious how to define areas in which the explanatory variables can be linearly related to the response. An extreme possibility is to fit a linear regression at each location \mathbf{s}_i , $i = 1, \dots, n$, with an intercept a_i and a slope vector \mathbf{b}_i varying by locations, i.e.,

$$y_i = a_i + \mathbf{x}_i \mathbf{b}_i + \epsilon_i, \quad i = 1, \dots, n. \quad (1.2)$$

However, model (1.2) is overparametrized in that each observation corresponds to at least two parameters, and therefore, estimation is impossible. To fix this problem, one needs to explore new modeling strategies. For example, Gelfand *et al.* (2003) proposed the following spatially varying coefficient model

$$\mathbf{y}(\mathbf{s}) = A(\mathbf{s}) + \mathbf{x}(\mathbf{s})\mathbf{B}(\mathbf{s}) + \epsilon(\mathbf{s}), \quad \mathbf{s} \in \mathbb{I}^2, \quad (1.3)$$

where $\mathbb{I} = [0, 1)$, \mathbf{s} indicates a location, $A(\mathbf{s}) \in \mathbb{R}$ and $\mathbf{B}(\mathbf{s}) \in \mathbb{R}^K$, ϵ is Gaussian process (i.e., for any finite locations $\mathbf{s}_{l_1}, \dots, \mathbf{s}_{l_k}$, the random variable $(\epsilon(\mathbf{s}_{l_1}), \dots, \epsilon(\mathbf{s}_{l_k}))$ is Gaussian), x is a K -dimensional covariate process. Note that when restricted to the discrete grid of locations (or pixels) $\mathbf{s}_1, \dots, \mathbf{s}_n$, model (1.2) is the same as model (1.3), since we may view the a_i 's and \mathbf{b}_i 's as the values of the (unknown) surfaces A and \mathbf{B} at these locations. I call A the intercept surface and \mathbf{B} the slope surface. In order to use a Bayesian approach, Gelfand *et al.* (2003) suggested a Gaussian process prior distribution on A and \mathbf{B} , which implicitly

assumes stationarity of A and \mathbf{B} .

The assumption of stationarity seems restrictive and is usually not verifiable. To relax the stationarity assumption, Zhang *et al.* (2011) utilized a wavelet approach to represent A and \mathbf{B} , so that the model (1.3) is transformed into a piecewise linear model. This transformation is used as a way to effectively reduce the number of model parameters so that the resulting model is not overparametrized. In order to better understand their modeling approach, I briefly review some background about wavelets.

Wavelets are set of functions whose shifts and scales form a set of basis functions. In particular, a bivariate wavelet consists of three functions denoted by φ^r for $r = 1, 2, 3$. When the φ^r s are chosen correctly, any two-dimensional square integrable function f can be represented by the following approximation,

$$f(\mathbf{s}) \approx \beta_0 + \sum_{r=1}^3 \sum_{j=0}^J \sum_{k \in \Lambda_j} \beta_{jk}^r \varphi_{jk}^r(\mathbf{s}), \quad \mathbf{s} \in \mathbb{I}^2, \quad (1.4)$$

where J is the maximal level of decomposition, $\varphi_{jk}^r(\mathbf{s}) := 2^j \varphi^r(2^j \mathbf{s} - k)$ is the scale-and-shift transform of function φ^r , and $\Lambda_j = \{(k_1, k_2) | k_1 = 0, 1, \dots, 2^j - 1, k_2 = 0, 1, \dots, 2^j - 1\}$ is the index set for k at level j . $\{\varphi_{jk}^r\}$ is called the wavelet basis and $\{\beta_{jk}^r, \beta_0\}$ are called the wavelet coefficients. If we want to include more details or information of the image f , a large J is preferred. Actually, when J goes to infinity, the representation (1.4) will be exact (see Daubechies 1992), which means all of the information on f is included. Through a wavelet basis, we can transform any L_2 function into a unique set of wavelet coefficients and this transformation is called a discrete wavelet transform (DWT). The DWT is invertible since any set of scalars can produce a unique function through (1.4). One feature of a DWT is that it can usually result in a very sparse coefficient set in the sense that most of the wavelet coefficients are zero while others deviate considerably from zero.

When a Haar wavelet is implemented, the corresponding wavelet basis functions become piecewise constant on their supports. A simple but illustrative example is a one dimensional Haar wavelet φ (see left panel in Figure 2). The wavelet bases satisfy $\varphi_{jk}(s) = 2^{j/2}$ when $s \in [2^{-j}k, 2^{-j}(k+1))$; and zero otherwise. An example of a piecewise constant function which can be patched up by these basis functions is also demonstrated in Figure 2 (see right panel). When generalized to a two dimensional situation, the supports of the Haar wavelet bases become square regions, and the corresponding basis functions are piecewise constant on these squares.

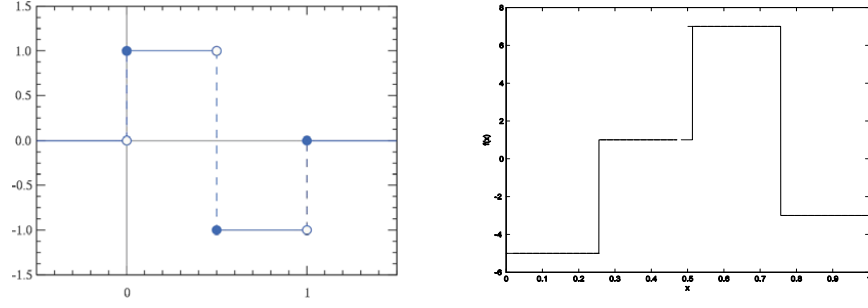


Figure 2: The left panel demonstrates the one-dimensional Haar wavelet φ . The right panel demonstrates an example piecewise constant function f . More specifically, f can be patched up by Haar wavelet bases through the representation $f(s) = 2\varphi(s) - 3\varphi(2s) + 5\varphi(2s-1)$, for $0 \leq s \leq 1$.

Zhang *et al.* (2011) utilized a DWT to express $A(\mathbf{s}) = W(\mathbf{s})\alpha$ and $B_k(\mathbf{s}) = W(\mathbf{s})\beta_k$, where α and β_k are d -dimensional vectors of wavelet coefficients, B_k is the k -th component of \mathbf{B} for $k = 1, \dots, K$, and $W(\mathbf{s})$ is a row vector of length d corresponding to the DWT at location \mathbf{s} . Note that when J -level wavelet expansions are used, then $d = 4^{J+1}$. Therefore, if p is the total number of wavelet coefficients, then $p = (K+1)d = (K+1)4^{J+1}$. If n sets of image are observed, then model (1.3) can be rewritten as

$$\mathbf{y} = X\beta + \epsilon, \quad (1.5)$$

where $\mathbf{y} = (y(\mathbf{s}_1), \dots, y(\mathbf{s}_n))'$, $X = [W, \tilde{\mathbf{x}} \circ W]$ is an $n \times p$ matrix with “ \circ ” being the Schur product, W is an $n \times d$ matrix with rows $W(\mathbf{s}_i)$'s, $\beta = [\alpha', \beta'_1, \dots, \beta'_K]'$ is a p -vector, $\tilde{\mathbf{x}} = [x_k(\mathbf{s}_i)]_{1 \leq k \leq K, 1 \leq i \leq n}$ is an $n \times K$ matrix, and x_k is the k -th component of \mathbf{x} . In order to capture fine details, the number of wavelet coefficients might be large. Consequently, fitting (1.5) becomes a large-scale problem. In general, α and the β_k 's can be expected to be sparse, but their nonzero components must be selected and estimated. For this purpose, Zhang *et al.* (2011) consider the minimization of the following LASSO criteria function

$$L_{n,\lambda}(\beta) = \|\mathbf{y} - X\beta\|^2 + \lambda\|\beta\|_1, \quad (1.6)$$

where $\|\cdot\|$ is Euclidean norm and $\|\cdot\|_1$ is l_1 norm. The l_1 penalty term in (1.6) is effective in producing a sparse estimate, which meets our goal. A large-scale l_1 least squares penalized algorithm (l_1 -ls, see Kim *et al.* 2007) was employed by Zhang *et al.* (2011) for the selection and estimation of the nonzero components in β .

The optimization of (1.6) is basically a non-Bayesian variable selection problem. There is a large amount of literature about variable selection using non-Bayesian approaches. They include a number of useful algorithms such as LASSO, adaptive LASSO and LARS. These algorithms are useful for producing point estimates of model parameters. However, if we are particularly interested in making inferences on model parameters, a Bayesian approach could be adopted. One reason is that, if the responses are non-Gaussian, a likelihood-based approach for inferences requires the construction of asymptotic confidence intervals, which might be difficult since a theoretical derivation of the asymptotic distribution might be complicated. A Bayesian approach can yield an estimate from current data without involving any asymptotics, although it might need MCMC. Another reason is that a Bayesian approach could naturally fit a hierarchical model structure through a construction of hierarchical

priors, which works effectively for a complex model. Our goal here is to perform a Bayesian variable selection (BVS) on model (1.5) to select informative wavelet coefficients, then recover the intercept and slope surfaces from these coefficients.

Variable selection from a Bayesian view is conceptually straightforward. Since each group of variables can define a candidate model, variable selection is also called model selection. By assuming priors on model coefficients and the classes of all candidate models, a posterior distribution of the models can be derived. An advantage of BVS is that, instead of merely selecting a subset of variables, it provides a complete assessment of each model in terms of posterior probability. When selecting the informative coefficients, BVS can simultaneously provide their estimates, and model uncertainty on these estimates through a posterior distribution.

The application of BVS on model (1.5) is associated with some special features of wavelet coefficients, which makes our approach different from some canonical BVS approaches. Wavelet coefficients are strongly related to their corresponding basis supports. A vanishing coefficient implies the inactivity of the corresponding basis support, and wavelet coefficients thus have clear spatial interpretations. Instead of assuming an underlying spatial structure on wavelet basis supports, I directly assume it on wavelet coefficients. Borrowing an idea from signal processing, I could assume a tree structure on the wavelet coefficients (see Crouse *et al.* 1998; Romberg *et al.* 2001). A tree structure comes from the natural nested structure of the wavelet basis supports. Thus, the wavelet coefficients naturally form a quad-tree (Figure 3). Large or small coefficients tend to propagate through quad-trees; this process is called the persistency property (Romberg *et al.* 2001). Hence, correlations among the wavelet coefficients could be modeled across quad-trees due to this persistency property. I will formulate priors on wavelet coefficients of both intercept and slope surfaces according to this tree structure. For other aspects of a model with tree structures, see Zhu and Wei

(2006).

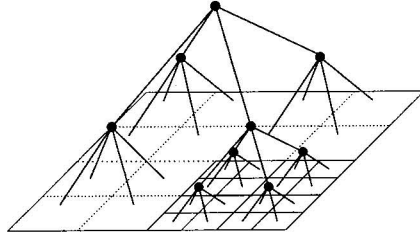


Figure 3: *The quad-tree structure of wavelet coefficients. In particular, each node represents a wavelet coefficient. The upper nodes correspond to the coefficients with lower resolution levels and lower nodes correspond to that with higher resolution levels. Each node (except for the top one) has a parent node and four child nodes.*

I briefly review several references on BVS and relevant issues. The references contain both theoretical and applied results. Both Bayesian and non-Bayesian approaches will be introduced.

The classic BVS was proposed by George and McCulloch (1993), which is based on the following model,

$$(a). \mathbf{y}|\beta, \sigma^2 \sim N(X\beta, \sigma^2 I),$$

$$(b). \beta_j|\gamma_j \sim (1 - \gamma_j)N(0, \tau_j^2) + \gamma_j N(0, c_j \tau_j^2),$$

$$(c). \gamma_j \sim_{ind.} Bernoulli(p_j).$$

where $c_j > 0$, $\tau_j^2 > 0$. Each γ_j is a Bernoulli variable and $\gamma_j = 1$ implies that β_j is included in the model. The authors give procedures to select c_j and τ_j^2 . They also define the best model to be $\hat{\gamma} = \argmax p(\gamma|\text{data})$. A Gibbs sampling procedure is provided for computations.

Subsequently, based on variations of (a)–(c), different BVS procedures were proposed. Clyde et al. (1998) provided BVS for a one dimensional wavelet regression model. They assume that a signal vector $\mathbf{f} = (f(s_1), \dots, f(s_n))'$ is observed with noise and consider the model $\mathbf{y} = \mathbf{f} + \epsilon$, where \mathbf{y} is the vector of observations and ϵ is the vector of noise. By

performing a one dimensional DWT on \mathbf{f} , they rewrite this model as $\mathbf{y} = W\beta + \epsilon$, where β is the vector of wavelet coefficients and W is an orthonormal square matrix. They considered an extreme case and proposed the following variation of (b),

$$(b)'. \quad \beta_j | \gamma_j \sim (1 - \gamma_j)\delta_0 + \gamma_j N(0, c_j \sigma^2),$$

where δ_0 is the point mass measure at zero. Then they developed BVS and relevant algorithms based on the model (a), (b)', (c).

A model similar to Clyde et al. (1998) was proposed by Clyde and George (2000), where a more general prior distribution on the model coefficients was considered. More specifically, they assume priors on the c_j 's in (b'). This prior distribution was originally proposed by Zellner (1986) and is called Zellner's g-prior. Zellner's g-prior makes a Bayesian model more general and flexible.

Another modeling strategy was implemented in Gabor regression for coefficient selection by Wolfe, Godsill and Ng (2004). One major difference between Gabor regression and wavelet regression is that the supports of wavelet bases are disjointed while those of Gabor bases overlap. Due to this feature of Gabor bases, the authors used Ising and Markov chain priors to model the relationship between the Gabor coefficients. In order to allow more flexibility, they proposed the following variations of (b)' and (c),

$$(b)'''. \quad \beta_j | \gamma_j, \tau_j^2 \sim (1 - \gamma_j)\delta_0 + \gamma_j N(0, \tau_j^2),$$

$$\tau_j^2 \sim \text{Inverse Gamma},$$

$$(c)'. \quad \gamma \sim p(\gamma),$$

where τ_j^2 may vary with β_j , and $p(\gamma)$ varies among the Bernoulli, Ising and Markov chain priors. Then, based on model (a), (b)''', (c)', the authors provide a data driven approach (Gibbs sampler) to approximate the β_j 's and τ_j^2 's.

Other relevant references include Brown *et al.* (2001) who used BVS based on a one-dimensional wavelet approach to analyze curve data over time, and proposed a Metropolis-Hasting type sampler for posterior computation. Brown *et al.* (2002) generalized the model proposed by George and McCulloch (1993) to a multi-dimensional situation, and proposed an estimation procedure based on prediction. Nott and Green (2004) discussed several computational issues related to BVS. Yuan and Lin (2005) explored the relationship between LASSO and Bayesian approaches through a variable selection view. Smith and Fahrmeir (2007) proposed a piecewise local linear model to analyze fMRI data, and performed BVS by using Ising priors on each local linear model.

Heretofore, several applied results related to BVS have been introduced. There are also several theoretical results on BVS which include (1) asymptotics of the posterior density: Jiang (2007); Jiang and Tanner (2008), in which the authors proved density consistency under some functional metric, and (2) posterior model consistency: Fernández *et al.* (2001); Casella *et al.* (2009); Liang *et al.* (2008); Moreno *et al.* (2010); and Shang and Clayton (2011), in which the authors proved that, under suitable conditions, the posterior probability of the true model converges to one as the sample size grows to infinity.

For variable selection under a frequentist setting, one interesting result is the so-called “oracle property,” which means that the estimates of the truly zero coefficients vanish with probability approaching one as n goes to infinity. Two representative articles on the oracle property in non-Bayesian variable selection are below.

Fan and Li (2001) consider the variable selection problem with a SCAD (smoothly clipped absolute deviation) penalty. The SCAD penalty is different from the usual l_1 or l_2 penalty in that it is nonconcave. The authors show that under certain regularity conditions, the SCAD penalized estimates enjoy an oracle property. Zou (2006) considers a weighted l_1 penalty estimation which he calls an adaptive LASSO, and proves that the estimates produced by

this procedure satisfy the oracle property.

There are a number of non-Bayesian approaches dealing with the high dimensional variable selection problem which includes Fan and Peng (2004), and Candes and Tao (2007). In particular, Wasserman and Roeder (2007) considered model (1.5) and showed that when the dimension of β is exponentially increasing with sample size and certain regularity conditions are satisfied, then with probability approaching one, the true model is contained in the selected model, based on a two-step selection procedure. The high dimensional problem when $p \gg n$ for a LASSO type estimate is discussed by Meinshausen and Yu (2009).

My PhD research work is concerned with two major generalizations. First, I expand on the work of Zhang *et al.* (2011) by using a Bayesian framework that allows for more direct inferences on the estimates. Second, this naturally results in a generalization of previous work on Bayesian variable selection in the wavelet-based one-dimensional time setting to a two-dimensional spatial setting. The result is an approach that is flexible and efficient for modeling the relationships between image data involving complex patterns. Furthermore, to address the large sample size and complex dependence structure of these spatial data, I implement an efficient Gibbs sampler. In addition to the applied issues, my thesis also includes some theoretical studies. Specifically, the results include the proof of PMC (posterior model consistency) and the relationship between PMC and posterior odds, under certain sufficient conditions when p grows with n at some rates. I also extend my theoretical results to an ultrahigh-dimensional setting using a dimension reduction approach and establish similar results under a g -prior setting.

The remainder of this thesis is structured as follows. Chapter 2 contains the theoretical results related to BVS. In particular, I demonstrate mild conditions under which the posterior consistency of BVS holds when p increases with n at certain rates. Chapter 3 is about the applications of BVS. I will apply a BVS approach to conduct estimation and inference related

to spatial concurrent linear models. Chapter 4 contains some discussions and future work.

Chapter 2

Consistency of Bayesian Linear Model Selection With a Growing Number of Parameters

2.1 Introduction

This work was motivated by efforts to analyze remotely sensed (satellite) data which consists of multiple spatial images. In the setting of interest, one image corresponds to a “response” while others correspond to covariates. To find the relationship between the response and covariate spatial images, Zhang *et al.* (2011) proposed a functional concurrent linear model with varying coefficients and applied a wavelet approach to transform this model into a linear model (with a particular design matrix) which contains an n -vector of responses and a sparse p -vector of wavelet coefficients. Since the images contain thousands of pixels, the model dimension p , which is determined by the maximum decomposition level in the wavelet expansion, has to be large so that sufficiently fine details in the target images can be captured. On the other hand, p has an upper bound $p \leq (K + 1)n$, where K is the total number of covariate images involved in the model. This is because each spatial image corresponds to a vector of wavelet coefficients which has dimension not exceeding n , and there are $K + 1$ images in total with one of them representing the intercept and others the

slopes. An important question is how to select the nonzero coefficients in the model, which is essentially a variable selection problem. Zhang *et al.* (2011) adopted a Lasso approach to address this.

The problem they handle relies on a specific design matrix induced by the wavelet structure. It is of interest, to frame the variable selection problem more broadly. More precisely, I suppose that data are drawn from the linear model

$$\mathbf{y} = X\beta + \epsilon, \quad (2.1)$$

where $\epsilon \sim N(\mathbf{0}, \sigma_0^2 I_n)$ is an n -vector of errors, $\mathbf{y} = (y_1, \dots, y_n)^T$ is an n -vector of responses, $\beta = (\beta_1, \dots, \beta_p)^T$ is a p -vector of parameters and $X = (X_1, \dots, X_p)$ is a $n \times p$ design matrix with X_j the j th column of X . It is also assumed that only a subset of X_1, \dots, X_p contribute to \mathbf{y} and I am interested in selecting the variables in this subset.

I consider a Bayesian variable selection (BVS) approach based on model (2.1). The Bayesian model to be considered is a variation of George and McCulloch (1993) and has been studied by Clyde *et al.* (1998), Clyde and George (2000), and Wolfe *et al.* (2004). Clearly, each subset of X_1, \dots, X_p defines a candidate model, so there are 2^p of them in total. According to George and McCulloch (1993), all the marginal posterior probabilities of these 2^p models can be calculated and the model with the largest posterior probability can be selected as the “best” model. This motivates the formal definition of posterior model consistency (PMC). I say that PMC holds if the true model, defined as the model from which samples are drawn, has a posterior probability approaching one. Since the sum of the posterior probabilities of all models equals one, when PMC holds, the posterior probability of any incorrect model will go to zero when n goes to infinity so that the true model can be correctly selected.

PMC has been theoretically verified when p is fixed (see Fernández *et al.*, 2001; Moreno and Girón, 2005; Liang *et al.*, 2008; Casella *et al.*, 2009). However, fewer results have been derived when p is growing with n , an interesting and important scenario. For increasing p , Berger *et al.* (2003), Moreno *et al.* (2010) and Girón *et al.* (2010) proved consistency for Bayes factors. Although PMC and consistency of Bayes factors are equivalent for fixed p (see Liang *et al.*, 2008; Casella *et al.*, 2009), they are different for growing p . Actually, I will see below that consistency of the Bayes factor is equivalent to consistency of the posterior odds under a general setting, but that the latter form of consistency is weaker than PMC. Therefore, it seems valuable to separately study PMC.

In this paper I will consider two classes of design matrix X , both with $p = p_n \leq n$, although our results can be generalized to $p \gg n$ when combined with certain dimension reduction approaches. In the first case, X is quite general. A representative situation is that the eigenvalues of $X^T X/n$ are uniformly bounded both above and below. Consistency is examined when p grows slower than n , say, $p \log n = o(n)$. I find that the posterior odds in favor of any incorrect model uniformly converges to zero, and the posterior probability of the true model converges to one. A second case I consider occurs when $X^T X/n$ is the identity matrix, i.e., $X^T X = nI_p$, and p grows as fast as n , say $p = n$. In that case, consistency of the posterior odds and PMC are examined, i.e., the posterior odds in favor of any incorrect model uniformly converges to zero, and the posterior probability of the true model converges to one. I also demonstrate how consistency of the posterior odds ratio can hold even though PMC fails. Finally, I generalize our results to a g -prior setting proposed firstly by Zellner (1986).

The Bayesian model structure used in this chapter is different from one that will be used in Chapter 3. However, all the theoretical results are based on an *iid*-error assumption, a scenario that will be considered in Chapter 3. Although the errors are assumed to be

independent, they will still produce some dependence structures among the responses when a suitable prior distribution is placed upon the model coefficients.

The remainder of this paper is organized as follows. In Section 2, preliminaries and main results will be provided. In Section 3, a numerical example related to the results of Section 2 is displayed. Section 4 contains the conclusion. Technical arguments are included in Section 5.

2.2 Preliminaries and main results

Suppose the n dimensional response vector $\mathbf{y} = (y_1, \dots, y_n)^T$ and the n by p covariate matrix $X = (X_1, \dots, X_p)$ are linked by the model

$$\mathbf{y} = X\beta + \epsilon, \quad (2.2)$$

where the X_j 's are n -vectors, $\beta = (\beta_1, \dots, \beta_p)^T$ is an unknown p -vector and ϵ is a vector of random errors. Here, X is allowed to be either (1) random but independent of ϵ or (2) deterministic. For $1 \leq j \leq p$, define the state variable of β_j by $\gamma_j = I(\beta_j \neq 0)$ and $\gamma = (\gamma_1, \dots, \gamma_p)^T$, where $I(\cdot)$ is the indicator function. I call γ the state vector of β and denote the number of 1's in γ by $|\gamma|$. The state vector γ completely determines the inclusion or exclusion of β_j 's in model (2.2), and therefore, can define a model $\mathbf{y} = X_\gamma \beta_\gamma + \epsilon$, where X_γ is an $n \times |\gamma|$ submatrix of X whose columns are indexed by the nonzero components of γ , and β_γ is the subvector (with size $|\gamma|$) of β indexed by the nonzero components of γ . It is natural, therefore, to call each γ a model. Note that there are 2^p such γ 's representing 2^p different models. For any state vectors γ and γ' , let $(\gamma \setminus \gamma')_j = I(\gamma_j = 1, \gamma'_j = 0)$ denote the difference (which is also a state vector) between γ and γ' , i.e., the 0-1 vector indicating

the variables that are present in γ but absent in γ' . I say that γ is nested in γ' (denoted by $\gamma \subset \gamma'$) if $\gamma \setminus \gamma' = \emptyset$. Denote the true model coefficient vector by β^0 and the corresponding state vector by γ^0 , and let $s_n = |\gamma^0|$ denote the size of the true model.

In this paper I consider the following hierarchical Bayesian model which is a variation of the model used by George and McCulloch (1993)

$$\begin{aligned} \mathbf{y}|\beta, \sigma^2 &\sim N(X\beta, \sigma^2 I_n), \\ \beta_j|\gamma_j, \sigma^2 &\sim (1 - \gamma_j)\delta_0 + \gamma_j N(0, c_j \sigma^2), \\ 1/\sigma^2 &\sim \chi_\nu^2, \\ \gamma &\sim p(\gamma), \end{aligned} \tag{2.3}$$

where δ_0 is point mass measure concentrated at zero. Hereafter, ν will be fixed a priori. Let $\Sigma = \text{diag}(\mathbf{c})$ with $\mathbf{c} = (c_j)_{1 \leq j \leq p}$ a p -vector of positive components, and let Σ_γ be the $|\gamma| \times |\gamma|$ sub-diagonal matrix of Σ corresponding to γ . Let $Z = (\mathbf{y}, X)$ denote the full data set. It follows by integrating out β and σ that the posterior distribution of γ is given by

$$p(\gamma|Z) \propto (2\pi)^{-n/2} \det(W_\gamma)^{-1/2} p(\gamma) \left\{ \frac{2}{1 + \mathbf{y}^T (I_n - X_\gamma U_\gamma^{-1} X_\gamma^T) \mathbf{y}} \right\}^{(n+\nu)/2}, \tag{2.4}$$

where $U_\gamma = \Sigma_\gamma^{-1} + X_\gamma^T X_\gamma$ and $W_\gamma = \Sigma_\gamma^{1/2} U_\gamma \Sigma_\gamma^{1/2}$. In particular, if $\gamma = \emptyset$ (the null model containing no covariate variables), (2.4) still holds if I adopt the conventions that $X_\emptyset = 0$ and $\Sigma_\emptyset = U_\emptyset = W_\emptyset = 1$.

Define $S_1 = \{\gamma | \gamma^0 \subset \gamma, \gamma \neq \gamma^0\}$ and $S_2 = \{\gamma | \gamma^0 \text{ is not nested in } \gamma\}$. It is clear that $S(n)$ defined by $S(n) = S_1 \cup S_2 \cup \{\gamma^0\}$ is the class of all state vectors. In particular, when $\gamma^0 = \emptyset$, S_2 is empty, and hence S_1 is the class of all state vectors excluding γ^0 . As was found by Liang *et al.* (2008), I will see later in this section that whether γ^0 is null or nonnull will

result in some differences in the main results (especially in the assumptions that are needed to establish our main results); thus, I will treat these cases separately. When γ^0 is nonnull, I denote $\varphi_{\min}(n) = \min_{\gamma \in S_2} \lambda_- \left(\frac{1}{n} X_{\gamma^0 \setminus \gamma}^T (I_n - P_\gamma) X_{\gamma^0 \setminus \gamma} \right)$ and $\varphi_{\max}(n) = \max_{\gamma \in S_2} \lambda_+ \left(\frac{1}{n} X_{\gamma^0 \setminus \gamma}^T X_{\gamma^0 \setminus \gamma} \right)$, where $P_\gamma = X_\gamma (X_\gamma^T X_\gamma)^{-1} X_\gamma^T$ is a projection matrix, $\lambda_-(A)$ and $\lambda_+(A)$ are the minimal and maximal eigenvalues of the square matrix A . I also adopt the convention that $P_\emptyset = 0$. For the case that $\gamma^0 = \emptyset$, both φ_{\min} and φ_{\max} are meaningless, and S_1 will be focused on in this situation.

Before proceeding further, I introduce several types of consistency central to this work. Generally speaking, to make a correct model selection

$$\max_{\gamma \neq \gamma^0} p(\gamma|Z)/p(\gamma^0|Z) \rightarrow 0 \quad (2.5)$$

should hold as $n \rightarrow \infty$, which means that the posterior probability of the true model asymptotically dominates that of any incorrect model. Following a framework similar to that of Zellner (1978), the term $p(\gamma|Z)/p(\gamma^0|Z)$, which is called the posterior odds in favor of γ , satisfies the relationship

$$p(\gamma|Z)/p(\gamma^0|Z) = BF(\gamma : \gamma^0) \frac{p(\gamma)}{p(\gamma^0)}, \quad (2.6)$$

where $BF(\gamma : \gamma^0) := p(Z|\gamma)/p(Z|\gamma^0)$ is the Bayes factor of γ versus γ^0 and $p(\gamma)/p(\gamma^0)$ is the prior odds in favor of γ . The Bayes factor is consistent if for any $\gamma \neq \gamma^0$, $BF(\gamma : \gamma^0) \rightarrow 0$. The posterior odds is consistent if for any $\gamma \neq \gamma^0$, $p(\gamma|Z)/p(\gamma^0|Z) \rightarrow 0$. It is easy to see that property (2.5) implies consistency of the posterior odds. I say that posterior model consistency (PMC) holds if $p(\gamma^0|Z) \rightarrow 1$. These types of consistency all have been useful in Bayesian model selection. Representative references include (1) assessment of posterior

odds: Jeffreys (1967), Zellner (1971, 1978); (2) performance of Bayes factor: Berger and Pericchi (1996), Moreno *et al.* (1998, 2010), Casella *et al.* (2009); (3) PMC: Fernández *et al.* (2001), Liang *et al.* (2008).

It is easy to see that when

$$\tilde{c}^{-1} \leq \min_{\gamma} p(\gamma)/p(\gamma^0) \leq \max_{\gamma} p(\gamma)/p(\gamma^0) \leq \tilde{c} \quad (2.7)$$

holds for some positive constant \tilde{c} , consistency of the Bayes factor is equivalent to consistency of the posterior odds, and that both are weaker than (2.5). A special case is that $p(\gamma) = 2^{-p}$ for all γ 's, which results in an indifference prior distribution for γ , see, e.g., Smith and Kohn (1996).

To illustrate the relationship between PMC and (2.5), note that

$$p(\gamma^0|Z) = \frac{1}{1 + \sum_{\gamma \neq \gamma^0} p(\gamma|Z)/p(\gamma^0|Z)}, \quad (2.8)$$

and thus $p(\gamma^0|Z) \rightarrow 1$ will imply (2.5). When p is fixed, it has been noted by Liang *et al.* (2008) that (2.5) implies PMC. However, when p grows with n , it will be shown later that this may not be true. This somewhat illustrates the difference between PMC and (2.5).

In what follows, I introduce some regularity conditions that are useful to establish our main results. I will also demonstrate some particular situations when these conditions are satisfied.

Assumption 2.2.1. There exists a constant $C_0 > 0$ such that for any n , $\max_{\gamma \in S(n)} p(\gamma)/p(\gamma^0) \leq C_0$.

Assumption 2.2.2. There exist positive constants C_1, C_2 such that with probability equal

to one, $\liminf_n \varphi_{\min}(n) \geq C_1$ and $\limsup_n \varphi_{\max}(n) \leq C_2$.

Assumption 2.2.3. There exists a positive sequence ψ_n such that $\min_{j \in \gamma^0} |\beta_j^0| \geq \psi_n$ and, as $n \rightarrow \infty$, $\psi_n \sqrt{n} \rightarrow \infty$.

Assumption 2.2.4. $p_n \rightarrow \infty$, $s_n \leq p_n \leq n$ and $p_n \log n = o(n \log(1 + \min\{\psi_n^2, 1\}))$.

Assumption 2.2.5. $p_n \rightarrow \infty$, $s_n \leq p_n \leq n$ and $p_n \log p_n = o(n)$.

Hereafter, unless otherwise explicitly stated, I will drop the subscript from p_n .

Assumption 2.2.6. There is a positive sequence $\bar{\phi}_n = O(n^{\delta_0})$ for some $\delta_0 > 0$ such that $\max_{1 \leq j \leq p} c_j \leq \bar{\phi}_n$, where c_j 's are the hyperparameters (in model (2.3)) controlling the prior variances of the nonzero β_j 's.

Assumption 2.2.7. There is a positive sequence $\underline{\phi}_n$ such that $k_n = O(\underline{\phi}_n)$ and $\min_{1 \leq j \leq p} c_j \geq \underline{\phi}_n$, where $k_n = \|\beta_{\gamma^0}^0\|_2^2$.

Assumption 2.2.8. There exist $C_3 > 0$ and $\delta \geq 0$ such that $n^{1-\delta} \underline{\phi}_n \rightarrow \infty$, and for any n , with probability equal to one,

$$\inf_{\gamma \in S_1} \lambda_- \left(\frac{1}{n} X'_{\gamma \setminus \gamma^0} (I_n - P_{\gamma^0}) X_{\gamma \setminus \gamma^0} \right) \geq C_3 n^{-\delta}. \quad (2.9)$$

Remark 2.1.

- (a). Assumption 2.2.1 is satisfied by some commonly used priors $p(\gamma)$, such as the flat prior distribution $p(\gamma) = 2^{-p}$ (Smith and Kohn, 1996). More generally, if $p(\gamma_j = 1) = \theta_j$ is such that both $\prod_{j \in \gamma \setminus \gamma^0} \left(\frac{\theta_j}{1-\theta_j} \right)$ and $\prod_{j \in \gamma^0 \setminus \gamma} \left(\frac{1-\theta_j}{\theta_j} \right)$ are bounded, then Assumption 2.2.1 is satisfied.

- (b). I use Assumption 2.2.3 to prove consistency for a growing p . Fan and Peng (2004) introduced a similar assumption in the framework of SCAD penalized optimization where \sqrt{n} in Assumption 2.2.3 was replaced by $1/\lambda_n$ with λ_n the penalty parameter. This condition requires the true parameters to be away from zero. Otherwise, it is impossible to distinguish between zero and nonzero parameters.
- (c). Assumptions 2.2.4 and 2.2.5 define a rate on the dimension p . In particular, when $\inf_n \psi_n > 0$, Assumption 2.2.4 is satisfied if $s_n \leq p$ and $p \log n = o(n)$. The results hold when s_n is either bounded or growing with n .
- (d). Assumption 2.2.6 excludes the possibility that $\bar{\phi}_n$ is extremely large, e.g., I exclude the situation that $\bar{\phi}_n = \exp(n^\omega)$ for some $\omega > 0$. Assumption 2.2.7 requires that $\underline{\phi}_n$ is not growing slower than $k_n = \|\beta_{\gamma_0}^0\|_2^2$. When the design matrix X is nonorthogonal, I use this assumption to facilitate the proof of consistency (see Theorem 2.2.2 below). But when X is orthogonal, this assumption is redundant and can be removed (see Corollary 2.2.5 below). \square

Assumptions 2.2.1, 2.2.3–2.2.7 are easily satisfied. The following proposition demonstrates that a broad class of design matrices X can satisfy Assumptions 2.2 and 2.8.

Proposition 2.2.1. If the $n \times p$ matrix X satisfies $\lambda_- \left(\frac{1}{n} X^T X \right) \geq c$, where $c > 0$ is constant, then for any $\gamma \subset \bar{\gamma}$ and $\gamma \neq \bar{\gamma}$,

$$\lambda_- \left(\frac{1}{n} X_{\bar{\gamma} \setminus \gamma}^T (I_n - P_\gamma) X_{\bar{\gamma} \setminus \gamma} \right) \geq c. \quad (2.10)$$

The proof of Proposition 2.2.1 can be found in Section 5 (Appendix).

Remark 2.2. Proposition 2.2.1 demonstrates that Assumptions 2.2.2 and 2.2.8 can hold

under general classes of design matrices. One such class consists of matrices X satisfying

$$1/\bar{c} \leq \lambda_- \left(\frac{1}{n} X^T X \right) \leq \lambda_+ \left(\frac{1}{n} X^T X \right) \leq \bar{c}, \quad (2.11)$$

where \bar{c} is some positive constant. For any $\gamma \in S_1$, I will have that $\gamma^0 \subset \gamma$ and $\gamma^0 \neq \gamma$. Thus, by Proposition 2.2.1, $\lambda_- \left(\frac{1}{n} X'_{\gamma \setminus \gamma^0} (I_n - P_{\gamma^0}) X_{\gamma \setminus \gamma^0} \right) \geq 1/\bar{c}$, i.e., inequality (2.9) in Assumption 2.2.8 holds. Notice that when $\gamma \in S_2$, the relationship $\gamma \subset \gamma^0 \vee \gamma$ and $\gamma \neq \gamma^0 \vee \gamma$ holds, where $\gamma^0 \vee \gamma$ denotes the p -vector with j th component the larger of $(\gamma^0)_j$ and γ_j , then Assumption 2.2.2 follows by applying Proposition 2.2.1. \square

In the following text, I assume that data are generated from the true model $\mathbf{y} = X\beta^0 + \epsilon$ with $\epsilon \sim N(0, \sigma_0^2 I_n)$. Let γ^0 be the p -dimensional state vector corresponding to β^0 . Unless otherwise stated, the limits in our main results will be taken when $n \rightarrow \infty$.

Theorem 2.2.2. Suppose that γ^0 is nonnull and Assumptions 2.2.1–2.2.4, 2.2.6–2.2.8 are satisfied. Let $\delta \geq 0$ satisfy Assumption 2.2.8. If $p^{\alpha_0} = o(n^{1-\delta} \underline{\phi}_n)$ for some $\alpha_0 > 2$, then

$\sup_{c_1, \dots, c_p \in [\underline{\phi}_n, \bar{\phi}_n]} \max_{\gamma \neq \gamma^0} p(\gamma|Z)/p(\gamma^0|Z) \rightarrow 0$ in probability. If $p^{\alpha_0+2} = o(n^{1-\delta} \underline{\phi}_n)$ for some $\alpha_0 > 2$, then $\sup_{c_1, \dots, c_p \in [\underline{\phi}_n, \bar{\phi}_n]} \sum_{\gamma \neq \gamma^0} p(\gamma|Z) \rightarrow 0$ in probability, and consequently, $\inf_{c_1, \dots, c_p \in [\underline{\phi}_n, \bar{\phi}_n]} p(\gamma^0|Z) \rightarrow 1$ in probability.

The proof of Theorem 2.2.2 follows by first deriving asymptotic approximations of the posterior odds $p(\gamma|Z)/p(\gamma^0|Z)$ for any $\gamma \neq \gamma^0$, and then using these approximations to show that $\sum_{\gamma \neq \gamma^0} p(\gamma|Z)/p(\gamma^0|Z) \rightarrow 0$ in probability. The limit $p(\gamma^0|Z) \rightarrow 1$ (in probability) thus immediately follows from (2.8). Details are in the Appendix A.

Remark 2.3. Theorem 2.2.2 provides sufficient conditions under which (2.5) and PMC are satisfied. It asserts that, with large probability, uniformly for c_j 's $\in [\underline{\phi}_n, \bar{\phi}_n]$, $p(\gamma^0|Z)$ dominates $p(\gamma|Z)$ for any $\gamma \neq \gamma^0$, and $p(\gamma^0|Z)$ approaches one in probability. Thus, with large probability, the true model γ^0 can be selected from a Bayesian perspective. \square

Remark 2.4. When combined with certain dimension reduction techniques such as sure independence screening (SIS) proposed by Fan and Lv (2008), one can generalize Theorem 2.2.2 to the ultra-high dimensional setting, i.e., $p \gg n$. This framework has been explored by many authors from non-Bayesian perspectives (see, e.g., Meinshausen and Bühlmann, 2006; Meinshausen and Yu, 2009; Zhang and Huang, 2010; Bühlmann and Kalisch, 2010). Here, I explore it by a Bayesian way. The basic idea is to first reduce the high-dimensional linear model so that the model dimension is below n , and then apply Bayesian model (2.3) to this reduced linear model. Under suitable conditions and using the arguments similar to the proof of Theorem 2.2.2, one can show that the posterior probability of the true model based on the reduced linear model converges in probability to 1. I refer to Appendix B for the description of this result and details of the proof. \square

The following result is an application of Theorem 2.2.2 in a special setting, which allows the growth rate of p to be $p \log n = o(n)$.

Corollary 2.2.3. Suppose that γ^0 is nonnull and Assumptions 2.2.1, 2.2.2 and inequality (2.9) are satisfied. Assume that $\min_{j \in \gamma^0} |\beta_j^0| \geq \psi_n$ with $\inf_n \psi_n > 0$, and p satisfies $p \log n = o(n)$. Suppose there exists a constant δ_0 with $\delta_0 > 3 + \delta$ such that $k_n = O(n^{\delta_0})$, where $\delta \geq 0$ is specified in inequality (2.9). Then with the selection $\bar{\phi}_n = O(n^{\delta_0})$ and $n^{\delta_0} = O(\underline{\phi}_n)$, I have

$$\inf_{c_1, \dots, c_p \in [\underline{\phi}_n, \bar{\phi}_n]} p(\gamma^0 | Z) \rightarrow 1 \text{ in probability.}$$

The proof of Corollary 2.2.3 can be finished by choosing $\alpha_0 \in (2, \delta_0 - \delta - 1)$ and verifying the assumptions in Theorem 2.2.2.

Theorem 2.2.2 deals with the case when the true model is nonnull. If the true model is null, then the response vector \mathbf{y} will have a zero mean. The corresponding result is summarized below.

Theorem 2.2.4. Suppose γ^0 is null, i.e., $\mathbf{y} = \epsilon \sim N(0, \sigma_0^2 I_n)$, and that Assumptions 2.2.1,

2.2.5–2.2.8 are satisfied. If $p^{\alpha_0} = o(n^{1-\delta}\underline{\phi}_n)$ for some $\alpha_0 > 2$, then $\sup_{c_1, \dots, c_p \in [\underline{\phi}_n, \bar{\phi}_n]} \max_{\gamma \neq \gamma^0} p(\gamma|Z)/p(\gamma^0|Z) \rightarrow 0$ in probability. If $p^{\alpha_0+2} = o(n^{1-\delta}\underline{\phi}_n)$ for some $\alpha_0 > 2$, then $\sup_{c_1, \dots, c_p \in [\underline{\phi}_n, \bar{\phi}_n]} \sum_{\gamma \neq \gamma^0} p(\gamma|Z) \rightarrow 0$ in probability, and consequently, $\inf_{c_1, \dots, c_p \in [\underline{\phi}_n, \bar{\phi}_n]} p(\gamma^0|Z) \rightarrow 1$ in probability.

The proof of Theorem 2.2.4 is similar to Theorem 2.2.2 and can be found in Appendix.

Although it is valid for a general type of design matrix, Theorem 2.2.2 requires that p grows slower than n . More precisely, if the Assumptions in Theorem 2.2.2 are satisfied, then $p = o(n)$. To see this, I notice that Assumptions 2.2.6, 2.2.7 and the fact that $\psi_n \leq k_n^{1/2}$ lead to $\psi_n = O(n^{\delta_0})$ for some $\delta_0 > 0$. Therefore, $p = o(n)$ follows from Assumption 2.2.4. In order to obtain consistency when p may grow as fast as n , one idea, but not the weakest possible, is to assume orthogonality of X , i.e., $X^T X = nI_p$, and to relax Assumption 2.2.7. To simplify the technical proof, I assume in the following Corollaries 2.2.5 and 2.2.6 that all c_j 's in model (2.3) are equal to ϕ_n . Moreover, I need the following assumption about the growth rates of s_n and p to replace Assumptions 2.2.4 and 2.2.5.

Assumption 2.2.9. Let $a_n = n + \sigma_0^{-2}k_n/(n^{-1} + \phi_n)$ and $\zeta \in (1, \infty)$ be a constant such that $n\psi_n^2 > \sigma_0^2\zeta a_n$ as $n \rightarrow \infty$. The numbers p and s_n with $p \rightarrow \infty$ and $s_n \leq p \leq n$ satisfy

$$(i). \quad s_n = o\left(\min\left\{\frac{(n+\nu)\log(n\psi_n^2/(\sigma_0^2\zeta a_n))}{\log(1+n\phi_n)}, n\psi_n^2, n\right\}\right).$$

$$(ii). \quad p \log p = o(a_n).$$

Assumption 2.2.9 potentially allows the case $p = n$. To see this, suppose $s_n = O(1)$ and I choose ϕ_n such that $(n + \nu)/\log(1 + n\phi_n) \rightarrow \infty$. When a_n grows faster than $n \log n$ and $n\psi_n^2/a_n \rightarrow \infty$, $p = n$ will satisfy Assumption 2.2.9. However, this requires ψ_n^2 to grow at least faster than $\log n$. This extra requirement on ψ_n^2 has not been imposed by Theorems 2.2.2 and 2.2.4, and can be treated as the price which I pay to relax the growth rate for p .

Under Assumption 2.2.9 and assuming orthogonality on X , I have the following consistency result which allows a faster growth rate for the dimension p .

Corollary 2.2.5. Assume that $X^T X = nI_p$ and $\Sigma = \phi_n I_p$ with $n\phi_n \rightarrow \infty$ and $\log \phi_n = O(\log n)$. Suppose γ^0 is nonnull and that Assumptions 2.2.1 and 2.2.9 are satisfied. If $p^{\alpha_0(n+\nu)/a_n} = o(n\phi_n)$ for some $\alpha_0 > 2$, then $\max_{\gamma \neq \gamma^0} p(\gamma|Z)/p(\gamma^0|Z) \rightarrow 0$ in probability. If $p = o\left((n+\nu) \log\left(\frac{n\psi_n^2}{\sigma_0^2 \zeta a_n}\right)\right)$ with ζ specified in Assumption 2.2.9, and $p^{2+\alpha_0(n+\nu)/a_n} = o(n\phi_n)$ for some $\alpha_0 > 2$, then $\sum_{\gamma \neq \gamma^0} p(\gamma|Z) \rightarrow 0$ in probability, and consequently, $p(\gamma^0|Z) \rightarrow 1$ in probability.

The proof of Corollary 2.2.5 is similar to those for Theorems 2.2.2 and 2.2.4 and is given in Appendix C. The following result, which requires a special model set-up, demonstrates that PMC and consistency of the posterior odds may hold in some situations but fail in others.

Corollary 2.2.6. Assume $p = n$, $X^T X = nI_n$ and $\Sigma = \phi_n I_n$. Suppose $\min_{j \in \gamma^0} |\beta_j^0| \geq \psi_n$ with $\psi_n^2 = c_1 n^{1+\delta_1} (\log n)^2$ for some constants $\delta_1 > 1$ and $c_1 > 0$, $k_n = O(\psi_n^2)$ and $p(\gamma) = \text{constant}$ for all γ . Assume that $s_n = s$ with $s > 0$ a fixed integer, i.e., the true parameter vector β^0 contains exactly s nonzero components.

(a). Suppose $\phi_n = c_2 n^{\delta_2}$ for some constants $c_2 > 0$ and δ_2 .

i. If $-1 < \delta_2 \leq 1$, then $\max_{\gamma \neq \gamma^0} p(\gamma|Z)/p(\gamma^0|Z) \rightarrow 0$ in probability, but PMC does not hold. Specifically, when $-1 < \delta_2 < 1$, $p(\gamma^0|Z) \rightarrow 0$, a.s.; when $\delta_2 = 1$, then there exists a constant c_0 with $0 < c_0 < 1$ such that $\limsup_n p(\gamma^0|Z) \leq c_0$, a.s.

ii. If $1 < \delta_2 \leq \delta_1$, then $p(\gamma^0|Z) \rightarrow 1$ in probability.

(b). If $n^{n \log n} = O(\phi_n)$, then $p(\emptyset|Z)/p(\gamma^0|Z) \rightarrow \infty$ in probability, where \emptyset represents the null model. Therefore, $p(\gamma^0|Z) \rightarrow 0$ in probability.

- (c). If $n\phi_n \rightarrow \eta \in [0, \infty)$, then almost surely, $\liminf_n \max_{\gamma \neq \gamma^0} p(\gamma|Z)/p(\gamma^0|Z) \geq (1 + \eta)^{-1/2}$ and $\lim_n p(\gamma^0|Z) = 0$.

The proof of Corollary 2.2.6 is given in Appendix C.

Remark 2.5. The main contribution of Corollary 2.2.6 is to demonstrate the difference between PMC and (2.5), and provide example growth rates for ϕ_n under which the two forms of consistency fail. Although this is obtained in a special situation, similar results should be still true under a more general setting, for instance, where $p < n$ or $X^T X$ is not diagonal, but I do not consider those circumstances here.

Corollary 2.2.6 (a) demonstrates that (2.5) does not necessarily imply PMC. This means that, although the posterior probability of the true model might not be approaching one, the ratio of the posterior probabilities of any “incorrect” model and the true model can still converge to zero. This phenomenon will not occur when p is fixed. In practice, (2.5) is sufficient to make a correct model selection even if PMC might fail.

Corollary 2.2.6 (b) and (c) demonstrate that in order to make a correct model selection, ϕ_n cannot be either too small or too large. Specifically, when $\phi_n = o(n^{-1})$, it follows by Corollary 2.2.6 (c) that almost surely $\liminf_n \max_{\gamma \neq \gamma^0} p(\gamma|Z)/p(\gamma^0|Z) \geq 1$. Thus, with probability one, for any $\varepsilon > 0$, there exists an integer N such that for any $n \geq N$

$$\max_{\gamma \neq \gamma^0} p(\gamma|Z)/p(\gamma^0|Z) \geq 1 - \varepsilon.$$

This implies that there exists a model, say γ^* , such that $p(\gamma^*|Z) \geq (1 - \varepsilon)p(\gamma^0|Z)$. Thus, when ε is small, either $p(\gamma^*|Z) > p(\gamma^0|Z)$, or $p(\gamma^*|Z)$ is very close to $p(\gamma^0|Z)$, which will both affect the selection result. On the other hand, when ϕ_n is growing faster than $n^{n \log n}$, it follows from (b) that the null model will be preferred in favor of γ^0 .

Corollary 2.2.6 (b) and (c) can be also understood intuitively. When ϕ_n is too small,

the two distribution components in the mixture prior distribution of β tend to be indistinguishable so that it is difficult to separate the true model from some incorrect model; when ϕ_n approaches infinity, by (2.4), the posterior probability of any nonnull model approaches zero, and thus, all β_j 's are forced to be zero. This conclusion has been empirically obtained by Smith and Kohn (1996) under spline regression models. \square

Remark 2.6. Using arguments similar to the proofs of Theorems 2.2.2 and 2.2.4, and by the Borel-Cantelli lemma of Shao (2003), one can show the almost sure convergence of $p(\gamma^0|Z)$. I refer to Appendix D for details. \square

To conclude this section, let us look at an example which demonstrates that, when $\bar{\phi}_n = \bar{\phi}$ and $\underline{\phi}_n = \underline{\phi}$ with $\bar{\phi}$ and $\underline{\phi}$ unrelated to n , consistency might still hold under certain circumstances. This is motivated by a full Bayesian framework which requires all hyperparameters to be fixed.

Example 2.1. If a full Bayesian approach is desired, then I have to preselect the hyperparameters c_j 's, and so $\bar{\phi}_n = \bar{\phi}$ and $\underline{\phi}_n = \underline{\phi}$ could be fixed. Assume that $k_n = O(1)$, which is a slightly weaker assumption than that in Jiang (2007). Note that Assumptions 2.2.6 and 2.2.7 follow immediately. Suppose $\min_{j \in \gamma^0} |\beta_j^0| \geq \psi_n$ with $\psi_n \propto n^{-1/4} \sqrt{\log n}$, the prior distribution of model γ satisfies Assumption 2.2.1. Assume that $s_n = s$ with $s > 0$ a fixed integer (thus, the true model is nonnull), and design matrix X satisfies (2.11). Therefore, by Proposition 2.2.1 and Remark 2.2, Assumptions 2.2.2 and 2.2.8 both hold. I also notice that Assumption 2.2.3 is well satisfied. It follows from Theorem 2.2.2 that if $p \propto n^r$ for some $0 < r < 1/2$, then with probability approaching one, (2.5) holds, i.e., the true model can be correctly selected; if $p \propto n^r$ for some $0 < r < 1/4$, then PMC holds in probability.

2.3 Generalizations to g -prior settings

In section 2, I assume in the Bayesian model (2.3) that the prior variance of a nonzero β_j is $c_j\sigma^2$ with c_j being a fixed priori. In practice, one may consider placing a prior distribution $g(c)$ on c_j 's, which reduces to the so-called g -prior setting (see Zellner 1986; Liang *et al.* 2008). In this section, I will give some asymptotic results under a g -prior setting.

I consider the following variation in model (2.3):

$$\begin{aligned}\beta_j|\gamma_j, \sigma^2, c &\sim (1 - \gamma_j)\delta_0 + \gamma_j N(0, c\sigma^2), \quad j = 1, \dots, p, \\ c &\sim g(c),\end{aligned}$$

where g is a proper prior distribution on $[0, \infty)$. I still use $p(\gamma^0|Z)$ to denote the posterior probability of the true model. Note that $p(\gamma|Z)$ is obtained by integrating $p(\gamma, \beta, \sigma^2, c|Z)$ with respect to (β, σ^2, c) .

Theorem 2.3.1. Suppose that γ^0 is nonnull and Assumption 2.2.1 holds. Furthermore, $\|\beta^0\|_2 = O(1)$, $s_n = O(1)$, $\min_{j \in \gamma^0} |\beta_j^0| \geq \psi_n$ with $\psi_n \propto n^{-1/4} \sqrt{\log n}$, and the design matrix X satisfies property (2.11).

- (i) Let g be supported on $[\underline{\phi}, \bar{\phi}]$ with $0 < \underline{\phi} < \bar{\phi} < \infty$. If $p \propto n^r$ for some $0 < r < 1/2$, then $\max_{\gamma \neq \gamma^0} p(\gamma|Z)/p(\gamma^0|Z) \rightarrow 0$ in probability.
- (ii) Let g be proper on $[0, \infty)$. If $p \propto n^r$ for some $0 < r < 1/4$, then $p(\gamma^0|Z) \rightarrow 1$ in probability.

Theorem 2.3.2. Suppose that γ^0 is nonnull and Assumptions 2.2.1, 2.2.3 and 2.2.4 are satisfied. Let X satisfy (2.11). Suppose that $\underline{\phi}_n$ and $\bar{\phi}_n$ satisfy Assumptions 2.2.6 and 2.2.7.

- (i) Let g be supported on $[\underline{\phi}_n, \bar{\phi}_n]$. If $p^{\alpha_0} = o(n\underline{\phi}_n)$ for some $\alpha_0 > 2$, then $\max_{\gamma \neq \gamma^0} p(\gamma|Z)/p(\gamma^0|Z) \rightarrow 0$ in probability.
- (ii) Let g be proper on $[0, \infty)$ such that $\int_{\underline{\phi}_n}^{\bar{\phi}_n} g(c)dc = 1 + o(1)$. If $p^{\alpha_0+2} = o(n\underline{\phi}_n)$ for some $\alpha_0 > 2$, then $p(\gamma^0|Z) \rightarrow 1$ in probability.

Remark 3.1. Theorems 2.3.1 and 2.3.2 provide sufficient conditions for (2.5) and PMC under a g -prior setting. It states that with large probability, $p(\gamma^0|Z)$ dominates $p(\gamma|Z)$ for any $\gamma \neq \gamma^0$, and $p(\gamma^0|Z)$ approaches one in probability. In particular, the prior density g in Theorem 2.3.1 does not depend on n , which corresponds to a full Bayesian framework, but I need a narrow restriction on the growth rate of p ; g in Theorem 2.3.2 might depend on n , but I allow p to grow faster with n . \square

Remark 3.2. I conjecture, although do not rigorously prove, that the ranges $0 < r < 1/2$ and $0 < r < 1/4$ in parts (a) and (b) of Theorems 2.3.1 are optimal, in the sense that for any $r > 1/2$, if $p \propto n^r$, then $\max_{\gamma \neq \gamma^0} p(\gamma|Z)/p(\gamma^0|Z)$ does not converge to zero in probability; and for any $r > 1/4$, if $p \propto n^r$, then $p(\gamma^0|Z)$ does not converge to one in probability.

Remark 3.3. Liang *et al.* (2008) obtained model consistency under a mixture g -prior setting. Their proof relies on the Laplace approximation of the integrals. While the proofs of both Theorems 2.3.1 and 2.3.2 rely on the uniform convergence in Theorem 2.2.2. \square

2.4 Numerical results

In this section, simulated examples are given to show the finite sample behaviors of the model selection procedure, and a comparison with variable selection using mixture g -priors will be also demonstrated.

To construct the random design matrix X , I generated *iid* p -dimensional row vectors

$U_1, \dots, U_n \sim N(\mathbf{0}, I_p)$ and let U be an $n \times p$ matrix with i th row U_i for $i = 1, \dots, n$. Then I let $X = \sqrt{n}U (U^T U)^{-1/2}$. Thus, $X^T X = nI_p$. (I choose X to be orthonormal for purposes of illustration, although, as I saw in the preceding section, results can be derived for general X .) To explore the dimension effect, I have considered three growth rates for p with respect to n : (1) $p = n^{1/4}$, (2) $p = n^{1/2}$ and (3) $p = n^{3/4}$. Data were simulated from model (2.2) with $\sigma = 1$, $s_n = 2$ and the true model coefficients $(\beta_1^0, \beta_2^0) = (2, 2)$ and $(\beta_3^0, \dots, \beta_p^0) = (0, \dots, 0)$. I considered sample sizes $n = 100, 200$ and 400 respectively.

The hierarchical Bayesian model (2.3) was fitted and the prior distributions on σ^2 and γ were assumed to be $1/\sigma^2 \sim \chi_4^2$ and $p(\gamma_j = 1) = w_j$, for any $j = 1, \dots, p$. I examined two cases for w_j 's, namely, Case I: $w_j = 0.5$ for $1 \leq j \leq p$; Case II: $w_1 = w_2 = 0.3$, $w_3 = \dots = w_p = 0.7$. Case I places equal prior probabilities on all the models, while Case II places larger prior probabilities on the “incorrect” models. For simplicity, I let $c_1 = \dots = c_p = \phi_n$. The values of ϕ_n were chosen to be $\phi_n = 10, 100, 1000$. After 20,000 samples of (β, γ, σ) were drawn from the posterior distribution $p(\beta, \gamma, \sigma | Z)$ using a sub-blockwise Gibbs sampler developed by Godsill and Rayner (1998), I recorded the last 10,000 samples and treated the previous 10,000 samples as burnins. Convergence has been assessed by applying Gelman-Rubin's statistic to 5 parallel Markov chains for each ϕ_n . I denote $\gamma^{(1)}, \dots, \gamma^{(10000)}$ to be the last 10,000 samples of γ . Then $p(\gamma^0 | Z)$ is approximated by $p(\gamma^0 | Z) \approx \sum_{t=1}^{10000} I(\gamma^{(t)} = \gamma^0) / 10000$.

To study the frequentist property of $p(\gamma^0 | Z)$, I have generated 100 data sets Z_1, \dots, Z_{100} independently from model (2.2), and for each ϕ_n calculated the corresponding 100 posterior probabilities $p(\gamma^0 | Z_m)$, $m = 1, \dots, 100$. This idea was inspired from Fernández *et al.* (2001) who studied the Bayesian selection problem when p is fixed.

Table 1 summaries the mean and standard deviations of the 100 $p(\gamma^0 | Z_m)$'s. I compared four methods, namely, Methods 1 to 3 correspond to $\phi_n = 10, 100, 1000$ under the Bayesian model (2.3), and Method 4 uses hyper g -prior with tuning parameter 3 (see

Liang *et al.* 2008). Method 4 was performed using the R package BAS available from <http://www.stat.duke.edu/~clyde/BAS>. I observed that when $p = n^{1/4}$, all the four methods select the true model with high posterior probabilities. While for $p = n^{1/2}$ and $p = n^{3/4}$, Method 1 performs the worst, and Method 3 performs the best, among the methods.

			$n = 100$		$n = 200$		$n = 400$	
			mean	std	mean	std	mean	std
$p = n^{1/4}$	Case I	Method 1	0.94	0.05	0.96	0.04	0.92	0.10
		Method 2	0.98	0.02	0.99	0.02	0.97	0.05
		Method 3	0.99	0.01	0.99	0.01	0.99	0.02
		Method 4	0.96	0.04	0.96	0.05	0.95	0.05
	Case II	Method 1	0.86	0.14	0.91	0.08	0.86	0.10
		Method 2	0.94	0.10	0.97	0.04	0.95	0.05
		Method 3	0.98	0.05	0.99	0.02	0.98	0.02
		Method 4	0.92	0.05	0.94	0.05	0.89	0.08
$p = n^{1/2}$	Case I	Method 1	0.60	0.14	0.56	0.14	0.53	0.12
		Method 2	0.82	0.10	0.81	0.11	0.80	0.10
		Method 3	0.94	0.05	0.93	0.06	0.93	0.05
		Method 4	0.68	0.10	0.63	0.13	0.62	0.12
	Case II	Method 1	0.34	0.12	0.29	0.11	0.27	0.11
		Method 2	0.65	0.14	0.63	0.14	0.62	0.14
		Method 3	0.86	0.09	0.85	0.10	0.84	0.09
		Method 4	0.42	0.10	0.41	0.10	0.36	0.10
$p = n^{3/4}$	Case I	Method 1	0.14	0.07	0.07	0.04	0.04	0.03
		Method 2	0.47	0.13	0.38	0.12	0.33	0.10
		Method 3	0.77	0.10	0.71	0.13	0.68	0.11
		Method 4	0.21	0.08	0.16	0.05	0.16	0.06
	Case II	Method 1	0.02	0.01	0.00	0.00	0.00	0.00
		Method 2	0.20	0.10	0.13	0.06	0.08	0.05
		Method 3	0.55	0.15	0.48	0.12	0.41	0.12
		Method 4	0.04	0.02	0.03	0.02	0.04	0.02

Table 1: Means and standard deviations of the 100 $p(\gamma^0|Z_m)$'s. Methods 1 to 3 correspond to $\phi_n = 10, 100, 1000$ under the Bayesian model (2.3), and Method 4 uses hyper g -prior with tuning parameter 3.

2.5 Discussion

Previous work about posterior model consistency (PMC) includes Fernández *et al.* (2001) and Liang *et al.* (2008) when the number of parameters p is fixed. In this paper, I have studied PMC when the model dimension p grows with sample size n . Specifically, I have shown that, under a variation of the Bayesian model proposed by George and McCulloch (1993), the posterior probability of the true model converges to one, i.e., PMC holds. I have obtained this result in two situations: (i) design matrix X is general while p grows slower than n , e.g., $p \log n = o(n)$; (ii) $X^T X/n$ is the identity matrix and p may grow as fast as n , e.g., $p = n$. Furthermore, I have demonstrated under a special framework that the consistency results may fail if ϕ_n is too small or too large, where ϕ_n is the hyperparameter controlling the prior variance of the nonzero model coefficients. Precisely, when $\phi_n = o(n^{-1})$ (an example of small order) or when $n^{n \log n} = O(\phi_n)$ (an example of large order), both PMC and consistency of the posterior odds fail. Besides that, our results do not require that the candidate models are pairwise nested.

Berger *et al.* (2003), Moreno *et al.* (2010) and Girón *et al.* (2010) have proved the consistency of Bayes factor when p is growing with n . This form of consistency, under our framework, is equivalent to the consistency of the posterior odds if the prior odds are uniformly bounded from above and below, so it is of interest to illustrate the relationship between PMC and consistency of posterior odds. I have considered a special framework and shown that PMC implies consistency of the posterior odds but the reverse may not be true. This is different from the finding by Liang *et al.* (2008) who demonstrate the equivalence of PMC and consistency of the Bayes factor when p is fixed. When combined with dimension reduction procedures such as SIS (Fan and Lv, 2008), our results can be also extended to ultrahigh-dimensional situations. I have also generalized the consistency results to a g -prior

setting studied by Zellner (1986) and Liang *et al.* (2008).

An extension of the current work is noteworthy. Assumption 2.2.7 is a technical assumption used to facilitate the proof and may not be the weakest possible. I leave it to future work to determine whether this condition can be further weakened or even removed.

2.6 Appendix A: Proofs of the results in Section 2.2

In this section, I prove the main results in Section 2.2. I also prove some lemmas which are useful to establish the main results. Let $P(\cdot)$ denote the probability measure associated with the underlying probability space.

Proof of Proposition 2.2.1. It follows by assumption that $\frac{1}{n}X_{\bar{\gamma}}^T X_{\bar{\gamma}} \geq cI_{|\bar{\gamma}|}$. Letting $X_{\bar{\gamma}} = (X_{\gamma}, X_{\bar{\gamma} \setminus \gamma})$, I can write $\frac{1}{n}X_{\bar{\gamma}}^T X_{\bar{\gamma}} = \begin{pmatrix} A & B \\ B^T & C \end{pmatrix}$, where $A = X_{\gamma}^T X_{\gamma}/n$, $B = X_{\gamma}^T X_{\bar{\gamma} \setminus \gamma}/n$ and $C = X_{\bar{\gamma} \setminus \gamma}^T X_{\bar{\gamma} \setminus \gamma}/n$. By formula for the inverse of blocked matrix (Seber and Lee, 2003, page 466), the lower right corner of $(\frac{1}{n}X_{\bar{\gamma}}^T X_{\bar{\gamma}})^{-1}$ is B_{22}^{-1} with $B_{22} = C - B^T A^{-1} B = \frac{1}{n}X_{\bar{\gamma} \setminus \gamma}^T (I_n - P_{\gamma}) X_{\bar{\gamma} \setminus \gamma}$. Then $B_{22}^{-1} \leq c^{-1}I$, which implies $\lambda_{-}(B_{22}) \geq c$. \square

Lemma 2.6.1. Suppose $\epsilon \sim N(0, \sigma_0^2 I_n)$. Then:

- (a). Let $v_{\gamma} = (I_n - P_{\gamma})X_{\gamma^0 \setminus \gamma} \beta_{\gamma^0 \setminus \gamma}^0$. If S_2 is nonnull, then $\max_{\gamma \in S_2} |v_{\gamma}^T \epsilon| / \|v_{\gamma}\|_2 = O_p(\sqrt{p})$, where I adopt the convention that $|v_{\gamma}^T \epsilon| / \|v_{\gamma}\|_2 = 0$ when $v_{\gamma} = 0$.
- (b). If S_1 is nonnull, then for any $\alpha > 2$, with probability approaching one, $\max_{\gamma \in S_1} \epsilon^T (P_{\gamma} - P_{\gamma^0}) \epsilon / (|\gamma| - s_n) \leq \alpha \sigma_0^2 \log p$.
- (c). If S_2 is nonnull, and I adopt the convention that $\epsilon^T P_{\gamma} \epsilon / |\gamma| = 0$ when γ is null, then for any $\alpha > 2$, with probability approaching one, $\max_{\gamma \in S_2} \epsilon^T P_{\gamma} \epsilon / |\gamma| \leq \alpha \sigma_0^2 \log p$.

Proof of Lemma 2.6.1. I prove the result for the case where X is deterministic, and briefly talk about the proofs for the case where X is random and independent of ϵ .

(a) I first assume that X is deterministic. By inequality (9.3) in Durrett (2005), if $\xi \sim N(0, 1)$, then there exists a C_0 such that for any $t > 1$, $P(|\xi| \geq t) \leq C_0 \exp(-t^2/2)$. Note that $|v_\gamma^T \epsilon|/(\sigma_0 \|v_\gamma\|_2) \sim N(0, 1)$, and therefore, by Bonferroni's inequality,

$$P\left(\max_{\gamma \in S_2} \frac{|v_\gamma^T \epsilon|}{\|v_\gamma\|_2} \geq t\right) \leq \sum_{\gamma \in S_2} P\left(\frac{|v_\gamma^T \epsilon|}{\|v_\gamma\|_2} \geq t\right) \leq C_0 2^p \exp\left(-\frac{t^2}{2\sigma_0^2}\right).$$

Then the result holds by setting $t = C\sigma_0\sqrt{2p}$ with large C . When X is random but independent of ϵ , note that the conditional distribution of $|v_\gamma^T \epsilon|/(\sigma_0 \|v_\gamma\|_2)$ given X is $N(0, 1)$. Thus, the proof can be finished by the above arguments.

(b) Suppose X is deterministic. First, if $\xi = \chi_\mu^2$, then by Chebyshev's inequality, for any $2 < \alpha' < \alpha$,

$$\begin{aligned} & P(\xi \geq \alpha\mu \log p) \\ &= P(\exp(\xi/\alpha') \geq \exp((\alpha/\alpha')\mu \log p)) \\ &\leq \exp(-(\alpha/\alpha')\mu \log p) E\{\exp(\xi/\alpha')\} \\ &= (1 - 2/\alpha')^{-\mu/2} \exp(-(\alpha/\alpha')\mu \log p). \end{aligned}$$

Using this inequality, Bonferroni's inequality, and the fact that when $\gamma \in S_1$, $\epsilon^T(P_\gamma - P_{\gamma^0})\epsilon \sim$

$\sigma_0^2 \chi_{|\gamma|-s_n}^2$, I have

$$\begin{aligned}
& P \left(\max_{\gamma \in S_1} \frac{\epsilon^T (P_\gamma - P_{\gamma^0}) \epsilon}{|\gamma| - s_n} \geq \alpha \sigma_0^2 \log p \right) \\
& \leq \sum_{\gamma \in S_1} P \left(\epsilon^T (P_\gamma - P_{\gamma^0}) \epsilon \geq \alpha \sigma_0^2 (|\gamma| - s_n) \log p \right) \\
& \leq \sum_{\gamma \in S_1} (1 - 2/\alpha')^{-(|\gamma|-s_n)/2} \exp(-(\alpha/\alpha')(|\gamma| - s_n) \log p) \\
& = \sum_{r=1}^{p-s_n} \binom{p-s_n}{r} (1 - 2/\alpha')^{-r/2} \exp(-(\alpha/\alpha')r \log p) \\
& = \left(1 + (1 - 2/\alpha')^{-1/2} p^{-\alpha/\alpha'} \right)^{p-s_n} - 1 \rightarrow 0.
\end{aligned}$$

When X is random and independent of ϵ , then conditioning on X , $\epsilon^T (P_\gamma - P_{\gamma^0}) \epsilon \sim \sigma_0^2 \chi_{|\gamma|-s_n}^2$.

Thus, the conclusion follows from the above arguments.

(c) I let X be deterministic. The case where X is random can be handled similarly.

Assume that S_2 contains nonnull models, and note that when γ is nonnull, $\epsilon^T P_\gamma \epsilon \sim \sigma_0^2 \chi_{|\gamma|}^2$.

Fix arbitrarily α' such that $2 < \alpha' < \alpha$. Then by the proof of part (b) I have

$$\begin{aligned}
& P \left(\max_{\gamma \in S_2} \frac{\epsilon^T P_\gamma \epsilon}{|\gamma|} \geq \alpha \sigma_0^2 \log p \right) \\
& = P \left(\max_{\gamma \in S_2 \setminus \{\emptyset\}} \frac{\epsilon^T P_\gamma \epsilon}{|\gamma|} \geq \alpha \sigma_0^2 \log p \right) \\
& \leq \sum_{\gamma \in S_2 \setminus \{\emptyset\}} P \left(\epsilon^T P_\gamma \epsilon \geq \alpha \sigma_0^2 |\gamma| \log p \right) \\
& \leq \sum_{\gamma \in S_2 \setminus \{\emptyset\}} (1 - \alpha'/2)^{-|\gamma|/2} \exp(-(\alpha/\alpha')|\gamma| \log p) \\
& \leq \sum_{r=1}^p \binom{p}{r} (1 - 2/\alpha')^{-r/2} p^{-(\alpha/\alpha')r} \\
& = \left(1 + (1 - 2/\alpha')^{-1/2} p^{-\alpha/\alpha'} \right)^p - 1 \rightarrow 0. \quad \square
\end{aligned}$$

Proof of Theorem 2.2.2. I have

$$\begin{aligned}
-\log(p(\gamma|Z)/p(\gamma^0|Z)) &= -\log\left(\frac{p(\gamma)}{p(\gamma^0)}\right) + \frac{1}{2}\log\left(\frac{\det(W_\gamma)}{\det(W_{\gamma^0})}\right) \\
&\quad + \frac{n+\nu}{2}\log\left(\frac{1+\mathbf{y}^T(I_n - X_\gamma U_\gamma^{-1} X_\gamma^T)\mathbf{y}}{1+\mathbf{y}^T(I_n - X_{\gamma^0} U_{\gamma^0}^{-1} X_{\gamma^0}^T)\mathbf{y}}\right) \\
&= -\log\left(\frac{p(\gamma)}{p(\gamma^0)}\right) + \frac{1}{2}\log\left(\frac{\det(W_\gamma)}{\det(W_{\gamma^0})}\right) \\
&\quad + \frac{n+\nu}{2}\log\left(\frac{1+\mathbf{y}^T(I_n - X_\gamma U_\gamma^{-1} X_\gamma^T)\mathbf{y}}{1+\mathbf{y}^T(I_n - P_\gamma)\mathbf{y}}\right) \\
&\quad - \frac{n+\nu}{2}\log\left(\frac{1+\mathbf{y}^T(I_n - X_{\gamma^0} U_{\gamma^0}^{-1} X_{\gamma^0}^T)\mathbf{y}}{1+\mathbf{y}^T(I_n - P_{\gamma^0})\mathbf{y}}\right) \\
&\quad + \frac{n+\nu}{2}\log\left(\frac{1+\mathbf{y}^T(I_n - P_\gamma)\mathbf{y}}{1+\mathbf{y}^T(I_n - P_{\gamma^0})\mathbf{y}}\right). \tag{2.12}
\end{aligned}$$

Denote the above summands by T_1, T_2, T_3, T_4, T_5 . By Assumption 2.2.6, T_1 is bounded below.

Since $U_\gamma \geq X_\gamma^T X_\gamma$, I have $T_3 \geq 0$ for any n .

To approximate T_4 , let

$$\Delta = \mathbf{y}^T X_{\gamma^0} (X_{\gamma^0}^T X_{\gamma^0})^{-1} \left(\Sigma_{\gamma^0} + (X_{\gamma^0}^T X_{\gamma^0})^{-1} \right)^{-1} (X_{\gamma^0}^T X_{\gamma^0})^{-1} X_{\gamma^0}^T \mathbf{y}.$$

By the Sherman-Morrison-Woodbury matrix identity (Seber and Lee, 2003, page 467),

$$U_{\gamma^0}^{-1} - (X_{\gamma^0}^T X_{\gamma^0})^{-1} = - (X_{\gamma^0}^T X_{\gamma^0})^{-1} \left(\Sigma_{\gamma^0} + (X_{\gamma^0}^T X_{\gamma^0})^{-1} \right)^{-1} (X_{\gamma^0}^T X_{\gamma^0})^{-1}. \tag{2.13}$$

By (2.13) and the fact that $\left(\Sigma_{\gamma^0} + \left(X_{\gamma^0}^T X_{\gamma^0}\right)^{-1}\right)^{-1} \leq \Sigma_{\gamma^0}^{-1}$, I have

$$\begin{aligned}
& \frac{1 + \mathbf{y}^T (I_n - X_{\gamma^0} U_{\gamma^0}^{-1} X_{\gamma^0}^T) \mathbf{y}}{1 + \mathbf{y}^T (I_n - P_{\gamma^0}) \mathbf{y}} \\
&= 1 + \frac{\Delta}{1 + \mathbf{y}^T (I_n - P_{\gamma^0}) \mathbf{y}} \\
&\leq 1 + 2 \left(\frac{(\beta_{\gamma^0}^0)^T \Sigma_{\gamma^0}^{-1} \beta_{\gamma^0}^0 + \epsilon^T X_{\gamma^0} (X_{\gamma^0}^T X_{\gamma^0})^{-1} \Sigma_{\gamma^0}^{-1} (X_{\gamma^0}^T X_{\gamma^0})^{-1} X_{\gamma^0}^T \epsilon}{1 + \mathbf{y}^T (I_n - P_{\gamma^0}) \mathbf{y}} \right) \\
&\leq 1 + 2 \phi_n^{-1} \left(\frac{\|\beta_{\gamma^0}^0\|_2^2 + \epsilon^T X_{\gamma^0} (X_{\gamma^0}^T X_{\gamma^0})^{-2} X_{\gamma^0}^T \epsilon}{1 + \mathbf{y}^T (I_n - P_{\gamma^0}) \mathbf{y}} \right).
\end{aligned}$$

Since $\mathbf{y}^T (I_n - P_{\gamma^0}) \mathbf{y} / n = \epsilon^T (I_n - P_{\gamma^0}) \epsilon / n \rightarrow_p \sigma_0^2$, and $E\{\epsilon^T X_{\gamma^0} (X_{\gamma^0}^T X_{\gamma^0})^{-2} X_{\gamma^0}^T \epsilon\} \leq \sigma_0^2 s_n (n \varphi_{\min}(n))^{-1}$, I have $\epsilon^T X_{\gamma^0} (X_{\gamma^0}^T X_{\gamma^0})^{-2} X_{\gamma^0}^T \epsilon = O_p(s_n (n \varphi_{\min}(n))^{-1})$. Therefore, by Assumptions 2.2.2 and 2.2.3, and the fact that $k_n \geq s_n \psi_n^2$, I can show that

$$\frac{1 + \mathbf{y}^T (I_n - X_{\gamma^0} U_{\gamma^0}^{-1} X_{\gamma^0}^T) \mathbf{y}}{1 + \mathbf{y}^T (I_n - P_{\gamma^0}) \mathbf{y}} \leq 1 + \frac{2k_n}{n \phi_n \sigma_0^2} (1 + o_p(1)). \quad (2.14)$$

Consequently, $0 \leq -T_4 = O_p(1)$ follows from the condition that $k_n = O(\phi_n)$ (Assumption 2.2.7).

Next I approximate T_2 and T_5 in the following Lemmas 2.6.2 and 2.6.3.

Lemma 2.6.2. Under Assumption 2.2.8, if $\gamma \in S_1$, then uniformly for c_j 's $\in [\phi_n, \bar{\phi}_n]$, $T_2 \geq 2^{-1}(|\gamma| - s_n) \log(1 + C_3 n^{1-\delta} \phi_n)$. Under Assumption 2.2.2, if $\gamma \in S_2$, then uniformly for c_j 's $\in [\phi_n, \bar{\phi}_n]$, $T_2 \geq -2^{-1} s_n \log(1 + C_2 n \bar{\phi}_n)$, where C_2 and C_3 are constants given in Assumptions 2.2.2 and 2.2.8 respectively.

Proof of Lemma 2.6.2. If $\gamma \in S_1$, it follows from the determinant formula for block

matrices (Seber and Lee, 2003, page 468), and Assumption 2.2.8 that

$$\begin{aligned}
\det(U_\gamma) &= \det(U_{\gamma^0}) \det\left(\Sigma_{\gamma \setminus \gamma^0}^{-1} + X_{\gamma \setminus \gamma^0}^T (I_n - X_{\gamma^0} U_{\gamma^0}^{-1} X_{\gamma^0}^T) X_{\gamma \setminus \gamma^0}\right) \\
&\geq \det(U_{\gamma^0}) \det\left(\Sigma_{\gamma \setminus \gamma^0}^{-1} + X_{\gamma \setminus \gamma^0}^T (I_n - P_{\gamma^0}) X_{\gamma \setminus \gamma^0}\right) \\
&\geq \det(U_{\gamma^0}) \det\left(\Sigma_{\gamma \setminus \gamma^0}^{-1} + C_3 n^{1-\delta} I_{|\gamma \setminus \gamma^0|}\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\frac{\det(W_\gamma)}{\det(W_{\gamma^0})} &= \frac{\det(\Sigma_\gamma)}{\det(\Sigma_{\gamma^0})} \frac{\det(U_\gamma)}{\det(U_{\gamma^0})} \\
&\geq \det(\Sigma_{\gamma \setminus \gamma^0}) \det\left(\Sigma_{\gamma \setminus \gamma^0}^{-1} + C_3 n^{1-\delta} I_{|\gamma \setminus \gamma^0|}\right) \\
&= \det\left(I_{|\gamma \setminus \gamma^0|} + C_3 n^{1-\delta} \Sigma_{\gamma \setminus \gamma^0}\right) \\
&\geq \det\left((1 + C_3 n^{1-\delta} \underline{\phi}_n) I_{|\gamma \setminus \gamma^0|}\right) = (1 + C_3 n^{1-\delta} \underline{\phi}_n)^{|\gamma| - s_n}, \tag{2.15}
\end{aligned}$$

which shows that $T_2 \geq 2^{-1}(|\gamma| - s_n) \log(1 + C_3 n^{1-\delta} \underline{\phi}_n)$. If $\gamma \in S_2$, note that $\det(W_\gamma) \geq 1$, and by Assumption 2.2.2

$$T_2 \geq -\frac{1}{2} \log(\det(W_{\gamma^0})) \geq -\frac{1}{2} \log(\det(I_{s_n} + C_2 n \Sigma_{\gamma^0})) \geq -2^{-1} s_n \log(1 + C_2 n \bar{\phi}_n),$$

which completes the proof of Lemma 2.6.2. \square

Lemma 2.6.3. Let $\alpha_0 > 2$. If either Assumption 2.2.4 or 2.2.5 is satisfied, when n is large, with large probability and uniformly for $\gamma \in S_1$, $T_5 \geq -2^{-1}(|\gamma| - s_n) \alpha_0 \log p$. If both Assumptions 2.2.2 and 2.2.4 are satisfied, there exists a constant C' such that when n is large, with large probability and uniformly for $\gamma \in S_2$, $T_5 \geq 2^{-1}(n + \nu) \log(1 + C' \psi_n^2)$.

Proof of Lemma 2.6.3. I consider $\gamma \in S_1$ and S_2 separately. Notice that Assumption 2.2.4 implies that $p \log p = o(n \log(1 + \psi_n^2))$, and therefore implies that $p \log p = o(n \psi_n^2)$. Let

$v_\gamma = (I_n - P_\gamma)X_{\gamma^0 \setminus \gamma} \beta_{\gamma^0 \setminus \gamma}^0$. From Lemma 2.6.1 (a) and (c), there exists $C > 0$ such that when n is sufficiently large, with large probability, for any $\gamma \in S_2$,

$$\begin{aligned}
\mathbf{y}^T(I_n - P_\gamma)\mathbf{y} &= \|v_\gamma\|_2^2 + 2v_\gamma^T \epsilon + \epsilon^T(I_n - P_\gamma)\epsilon \\
&\geq \|v_\gamma\|_2^2 - 2C\sqrt{p}\|v_\gamma\|_2 + \epsilon^T \epsilon - C|\gamma| \log p \\
&\geq \|v_\gamma\|_2^2 \left(1 - 2C \frac{\sqrt{p}}{\|v_\gamma\|_2} - C \frac{p \log p}{\|v_\gamma\|_2^2}\right) + \epsilon^T \epsilon \\
&\geq \|v_\gamma\|_2^2 \left(1 - 2C \sqrt{\frac{p}{n\varphi_{\min}(n)\psi_n^2}} - C \frac{p \log p}{n\varphi_{\min}(n)\psi_n^2}\right) + \epsilon^T \epsilon \\
&= \|v_\gamma\|_2^2(1 + o(1)) + \epsilon^T \epsilon \\
&\geq n\varphi_{\min}(n)\|\beta_{\gamma^0 \setminus \gamma}^0\|_2^2(1 + o(1)) + \epsilon^T \epsilon \\
&\geq n\varphi_{\min}(n)\psi_n^2(1 + o(1)) + \epsilon^T \epsilon.
\end{aligned} \tag{2.16}$$

It is easy to see that Assumption 2.2.4 implies that $s_n = o(n)$, and therefore, $\epsilon^T(I_n - P_{\gamma^0})\epsilon = n\sigma_0^2(1 + o_p(1))$. Thus, by (2.16), there exists a C' such that for sufficiently large n , with large probability, uniformly for $\gamma \in S_2$,

$$T_5 \geq \frac{n + \nu}{2} \log \left(\frac{1 + n\varphi_{\min}(n)\psi_n^2(1 + o(1)) + \epsilon^T \epsilon}{1 + \epsilon^T(I_n - P_{\gamma^0})\epsilon} \right) \geq \frac{n + \nu}{2} \log (1 + C'\psi_n^2). \tag{2.17}$$

On the other hand, by properties of projection matrices and Lemma 2.6.1 (b), when n is

sufficiently large, with large probability, I have uniformly for $\gamma \in S_1$,

$$\begin{aligned}
& \frac{1 + \mathbf{y}^T(I_n - P_\gamma)\mathbf{y}}{1 + \mathbf{y}^T(I_n - P_{\gamma^0})\mathbf{y}} \\
&= 1 - \frac{\mathbf{y}^T(P_\gamma - P_{\gamma^0})\mathbf{y}}{1 + \mathbf{y}^T(I_n - P_{\gamma^0})\mathbf{y}} \\
&= 1 - \frac{(\beta_{\gamma^0}^0)^T X_{\gamma^0}^T (P_\gamma - P_{\gamma^0}) X_{\gamma^0} \beta_{\gamma^0} + 2(\beta_{\gamma^0}^0)^T X_{\gamma^0}^T (P_\gamma - P_{\gamma^0}) \epsilon + \epsilon^T (P_\gamma - P_{\gamma^0}) \epsilon}{1 + \mathbf{y}^T(I_n - P_{\gamma^0})\mathbf{y}} \\
&= 1 - \frac{\epsilon^T (P_\gamma - P_{\gamma^0}) \epsilon}{1 + \epsilon^T (I_n - P_{\gamma^0}) \epsilon} \geq 1 - \frac{\alpha(|\gamma| - s_n) \log p}{n},
\end{aligned}$$

where I have temporarily fixed an α such that $2 < \alpha < \sqrt{2\alpha_0}$. It follows by the inequality that $\log(1 - x) \geq -(\alpha/2)x$ when $x \in (0, 1 - 2/\alpha)$, and by Assumption 2.2.4 or 2.2.5 (which both imply that $(|\gamma| - s_n) \log p/n$ approaches zero uniformly for $\gamma \in S_1$) that for sufficiently large n , with large probability and uniformly for $\gamma \in S_1$,

$$T_5 \geq \frac{n + \nu}{2} \log \left(1 - \frac{\alpha(|\gamma| - s_n) \log p}{n} \right) \geq -2^{-1}(|\gamma| - s_n) \alpha_0 \log p, \quad (2.18)$$

which completes the proof of Lemma 2.6.3. \square

Now I am ready to finish the proof of Theorem 2.2.2. By (2.14), Lemma 2.6.2, Lemma 2.6.3, Assumption 2.2.4, and the fact that $p^{\alpha_0} = o(\rho_n)$ with $\rho_n \equiv n^{1-\delta} \underline{\phi}_n$, with large probability, uniformly for $\gamma \in S_1$ and $c_1, \dots, c_p \in [\underline{\phi}_n, \bar{\phi}_n]$,

$$\begin{aligned}
p(\gamma|Z)/p(\gamma^0|Z) &\leq \tilde{C} \exp \left(-2^{-1}(|\gamma| - s_n) \log((1 + C_3 \rho_n)/p^{\alpha_0}) \right) \\
&= \tilde{C} \left(\frac{1 + C_3 \rho_n}{p^{\alpha_0}} \right)^{-2^{-1}(|\gamma| - s_n)} \rightarrow 0.
\end{aligned} \quad (2.19)$$

By Assumptions 2.2.4 and 2.2.6, it can be verified that $s_n \log(1 + C_2 n \bar{\phi}_n) \ll \frac{n+\nu}{2} \log(1 + C' \psi_n^2)$.

So, with large probability, uniformly for $\gamma \in S_2$ and $c_1, \dots, c_p \in [\underline{\phi}_n, \bar{\phi}_n]$,

$$\begin{aligned} p(\gamma|Z)/p(\gamma^0|Z) &\leq \tilde{C} \exp \left(2^{-1} s_n \log(1 + C_2 n \bar{\phi}_n) - \frac{n + \nu}{2} \log(1 + C' \psi_n^2) \right) \\ &\leq \tilde{C} (1 + C' \psi_n^2)^{-\frac{n+\nu}{4}} \rightarrow 0, \end{aligned} \quad (2.20)$$

where \tilde{C} in (2.19) and (2.20) depends on the lower bounds of T_1 and T_4 . For the proof of PMC, I consider two cases. It is easy to see from (2.19) that

$$\begin{aligned} \sum_{\gamma \in S_1} p(\gamma|Z)/p(\gamma^0|Z) &\leq \tilde{C} \sum_{\gamma \in S_1} \left(\frac{1 + C_3 \rho_n}{p^{\alpha_0}} \right)^{-2^{-1}(|\gamma| - s_n)} \\ &= \tilde{C} \sum_{r=1}^{p-s_n} \binom{p-s_n}{r} \left(\frac{1 + C_3 \rho_n}{p^{\alpha_0}} \right)^{-\frac{r}{2}} \\ &= \tilde{C} \left[\left(1 + \left(\frac{1 + C_3 \rho_n}{p^{\alpha_0}} \right)^{-\frac{1}{2}} \right)^{p-s_n} - 1 \right] \rightarrow 0, \end{aligned}$$

where the last limit result follows from the assumption that $p^{\alpha_0+2} = o(\rho_n)$.

Similarly, by (2.20), and $p \log n = o(n \log(1 + \psi_n^2))$ (which follows from Assumption 2.2.4),

I can show that

$$\sum_{\gamma \in S_2} p(\gamma|Z)/p(\gamma^0|Z) \leq \tilde{C} 2^p (1 + C' \psi_n^2)^{-(n+\nu)/4} \rightarrow 0. \quad (2.21)$$

This completes the proof of Theorem 2.2.2. \square

Proof of Theorem 2.2.4. The assumption that γ^0 is null implies that the model class S_2 is empty. Similar to the proof of Theorem 2.2.2, I need to approximate T_1 to T_5 in (2.12). This is easier when the true model is null since $T_4 = 0$, and by Lemma 2.6.2, when γ is nonnull, $T_2 \geq 2^{-1} |\gamma| \log(1 + C_3 n^{1-\delta} \underline{\phi}_n)$. Since T_1 and T_3 are still bounded below, the proof is reduced to approximate T_5 . By Lemma 2.6.3, Assumption 2.2.5, and that $s_n = 0$, when n is large, with large probability and uniformly for $\gamma \in S_1$, $T_5 \geq -2^{-1} |\gamma| \alpha_0 \log p$. Therefore,

the remaining proofs can be finished by arguments similar to (2.19) and (2.21). \square

Proof Theorem 2.3.1. (i) Let $p(\gamma|Z, c)$ be the posterior probability of γ given Z and c , as specified by (2.4). Applying Theorem 2.2.2, I have that in probability

$$\sup_{c \in [\underline{\phi}, \bar{\phi}]} \max_{\gamma \neq \gamma^0} p(\gamma|Z, c)/p(\gamma^0|Z, c) \rightarrow 0.$$

Then the result follows from $p(\gamma|Z) = \int_{\underline{\phi}}^{\bar{\phi}} p(\gamma|Z, c)g(c)dc$, and

$$\frac{\int_{\underline{\phi}}^{\bar{\phi}} p(\gamma|Z, c)g(c)dc}{\int_{\underline{\phi}}^{\bar{\phi}} p(\gamma^0|Z, c)g(c)dc} \leq \sup_{c \in [\underline{\phi}, \bar{\phi}]} \max_{\gamma \neq \gamma^0} p(\gamma|Z, c)/p(\gamma^0|Z, c).$$

(ii) Let $0 < \underline{\phi} < \bar{\phi}$. By Theorem 2.2.2, $\inf_{c \in [\underline{\phi}, \bar{\phi}]} p(\gamma^0|Z, c) \rightarrow 1$ in probability. Since

$$p(\gamma^0|Z) = \int_{\underline{\phi}}^{\bar{\phi}} (p(\gamma^0|Z, c) - 1)g(c)dc + \int_{\underline{\phi}}^{\bar{\phi}} g(c)dc + \int_{[0, \infty) \setminus [\underline{\phi}, \bar{\phi}]} p(\gamma^0|Z, c)g(c)dc,$$

the result follows by fixing $\underline{\phi}$ and $\bar{\phi}$ so that $\int_{\underline{\phi}}^{\bar{\phi}} g(c)dc$ is close to 1, and letting n go to ∞ .

Proof Theorem 2.3.2. Proof is similar to those of Theorem 2.3.1.

Supplement Materials

Appendix B: Generalizations of Bayesian consistency to ultra-high dimensional settings.

Appendix C: Proof of Corollaries 2.2.5 and 2.2.6.

Appendix D: Almost Sure Consistency of $p(\gamma^0|Z)$.

Acknowledgement The authors wish to thank Professor Jun Shao for suggestions that helped to improve the present work.

2.7 Appendix B: Generalizations of Bayesian consistency to ultra-high dimensional settings

In this paper I have focused on the case $p \leq n$. Nowadays, data with ultrahigh dimensions, i.e., $p \gg n$, is of interest, and variable selection problem under this framework has received much attention. See for instance, Meinshausen and Bühlmann (2006); Meinshausen and Yu (2009); Zhang and Huang (2010); Bühlmann and Kalisch (2010). In this section, I will extend the results in Section 2 to the case that data is ultrahigh-dimensional. I assume that the original normal linear model contains p variables with $p \gg n$, while only s_n variables are included in the “true” model with $s_n \leq n$. To get consistency results, I proceed in two stages. First, I apply dimension reduction approaches, such as SIS proposed by Fan and Lv (2008), to reduce the size of the original model. A reduced model will be produced with size $\tilde{p} \leq n$. Second, I will apply the Bayesian hierarchical model (2.3) to the reduced model and apply the results in Section 2 to obtain consistency. Since all the work in the second stage has been completed by conditioning on the reduced model, the probability measure used to characterize the consistency should be the conditional probability measure given the reduced model. In what follows, I will introduce the ultrahigh-dimensional model framework and the corresponding asymptotic results in detail.

First of all, let us generalize the previous framework to ultrahigh-dimensional setting. I suppose that the model now becomes

$$\mathbf{y} = X\beta + \epsilon \tag{2.22}$$

where X is an $n \times p$ design matrix with $p \gg n$, β is a p -vector of model parameters and $\epsilon \sim N(\mathbf{0}, \sigma_0^2 I_n)$. Suppose that there is a p -dimensional true model parameter vector β^0 , i.e., data

were actually generated from $\mathbf{y} = X\beta^0 + \epsilon$ with $\epsilon \sim N(\mathbf{0}, \sigma_0^2 I_n)$, and denote the state vector (of dimension p) corresponding to β^0 to be γ^0 , and the size of the true model to be $s_n = |\gamma^0|$, which satisfies $s_n \leq n$. Without loss of generality, I suppose $(\gamma^0)_1 = \dots = (\gamma^0)_{s_n} = 1$, and $(\gamma^0)_j = 0$ for $s_n + 1 \leq j \leq p$.

To incorporate the results in Section 2 into the above setting, I let \tilde{p} be a non-stochastic positive integer satisfying $\tilde{p} \leq n$, and assume that model (2.22) can be reduced to the following submodel

$$\mathbf{y} = \tilde{X}\tilde{\beta} + \tilde{\epsilon}, \quad (2.23)$$

where \tilde{X} is an $n \times \tilde{p}$ submatrix of X , $\tilde{\beta}$ is a \tilde{p} -vector of parameters. Let $\hat{\gamma}^n$ be the p -dimensional state vector corresponding to model (2.23), i.e., $\hat{\gamma}_j^n = 1$ if variable X_j is involved in model (2.23), and 0 otherwise. Usually, $\hat{\gamma}^n$ might be random since it might come from a group of random samples. Note that there are exactly \tilde{p} ones and $(p - \tilde{p})$ zeros in $\hat{\gamma}^n$.

Throughout this section, I let (\mathbf{y}, X) be samples drawn from model (2.22) which are independent of $\hat{\gamma}^n$, and $Z = (\mathbf{y}, \tilde{X})$ be the “reduced ” data corresponding to model (2.23). Following Fan and Lv (2008), one way to achieve independence between $\hat{\gamma}^n$ and (\mathbf{y}, X) is to obtain $\hat{\gamma}^n$ through a new group of samples which are generated from model (2.22) and independent of (\mathbf{y}, X) .

Definition 2.7.1. $\hat{\gamma}^n$ is said to be *desirable* if $P(E_N) \rightarrow 1$ as N approaches ∞ , where $E_N := \bigcap_{n \geq N} \{\gamma^0 \subset \hat{\gamma}^n\}$, i.e., the event that γ^0 is nested in $\hat{\gamma}^n$ for all $n \geq N$.

Remark A.1. One important way to produce a desirable $\hat{\gamma}^n$ is through the SIS (sure independence screening) procedures introduced by Fan and Lv (2008) where the authors have shown that, under certain regularity conditions, $P(\gamma^0 \subset \hat{\gamma}^n) = 1 - O(\exp(-n^\tau))$, for some constant $\tau > 0$. Thus, it can be verified that $\hat{\gamma}^n$ produced by SIS is desirable. \square

In the next, I identify the posterior probability of the submodels of $\hat{\gamma}^n$. Let $\hat{\gamma}^n$ be desirable

with size \tilde{p} . Suppose $\hat{\gamma}^n$ contains γ^0 and let $\tilde{\gamma}^0$ be a \tilde{p} -dimensional state vector indicating the true model in (2.23), i.e., $(\tilde{\gamma}^0)_1 = \dots = (\tilde{\gamma}^0)_{s_n} = 1$ and $(\tilde{\gamma}^0)_{s_n+1} = \dots = (\tilde{\gamma}^0)_{\tilde{p}} = 0$. The \tilde{p} -dimensional “true” model parameter vector corresponding to model $\hat{\gamma}^n$ will be identified by $\tilde{\beta}_j^0 = \beta_j^0$, for $1 \leq j \leq s_n$, and $\tilde{\beta}_j^0 = 0$, for $s_n + 1 \leq j \leq \tilde{p}$. Now it is safe to fit our Bayesian hierarchical model (2.3) on $\hat{\gamma}^n$. For any \tilde{p} -dimensional state vector $\tilde{\gamma}$ indicating a submodel of $\hat{\gamma}^n$, I could identify the posterior probability $p(\tilde{\gamma}|Z)$ of $\tilde{\gamma}$ through (2.4). Since $\tilde{\gamma}^0$ is one of the submodels of $\hat{\gamma}^n$, $p(\tilde{\gamma}^0|Z)$ well defines the posterior probability of the true model.

Note that here I have only defined $p(\tilde{\gamma}^0|Z)$ based on those reduced models which contain the true model. By assuming that $\hat{\gamma}^n$ is desirable, it is not necessary to consider the case that $\hat{\gamma}^n$ does not include the true model, since this is a rare event.

To study the posterior probability of the true model $p(\tilde{\gamma}^0|Z)$, I need the following assumption which regularizes the design matrix X .

Assumption 2.7.1. There is a positive constant c'' such that with probability equal to one,

$$1/c'' \leq \lambda_- (X_\gamma^T X_\gamma/n) \leq \lambda_+ (X_\gamma^T X_\gamma/n) \leq c'',$$

for any p -dimensional state vector γ with $|\gamma| = \tilde{p}$ and $\gamma^0 \subset \gamma$.

Remark A.2. Assumption 2.7.1 is called the *sparse Riesz condition*, which is a reasonable assumption on design matrix X . This assumption has been used by several authors to establish asymptotic results under non-Bayesian frameworks when the model dimension is ultrahigh, see Meinshausen and Yu (2009) and Zhang and Huang (2010). It states that the sub-diagonal matrix of $X^T X/n$ “surrounding” the true model is nonsingular although $X^T X/n$ is singular. By Remark 2.2, Assumption 2.7.1 guarantees that the “reduced” design matrix \tilde{X} in (2.23) will satisfy Assumption 2.2.2 and the inequality (2.9) with $\delta = 0$, as long as $\hat{\gamma}^n$ contains γ^0 . \square

For PMC under the above setting, I have the following result.

Theorem 2.7.1. Suppose that $n, \tilde{p}, s_n, \bar{\phi}_n, \underline{\phi}_n, \tilde{\beta}^0$ satisfy Assumptions 2.2.3, 2.2.4, 2.2.6, 2.2.7 and, as $n \rightarrow \infty$, $n\underline{\phi}_n \rightarrow \infty$. Suppose that X satisfies Assumption 2.7.1 and the prior distribution $p(\tilde{\gamma})$ with $\tilde{\gamma} \in \{0, 1\}^{\tilde{p}}$ satisfies Assumption 2.2.1. If $\tilde{p}^{\alpha_0} = o(n\underline{\phi}_n)$, then

$$\sup_{c_1, \dots, c_p \in [\underline{\phi}_n, \bar{\phi}_n]} \max_{\tilde{\gamma} \neq \tilde{\gamma}^0} p(\tilde{\gamma}|Z)/p(\tilde{\gamma}^0|Z) \rightarrow_p 0. \text{ If } \tilde{p}^{\alpha_0+2} = o(n\underline{\phi}_n), \text{ then } \inf_{c_1, \dots, c_p \in [\underline{\phi}_n, \bar{\phi}_n]} p(\tilde{\gamma}^0|Z) \rightarrow_p 1.$$

Proof of Theorem 2.7.1. I only show the second part, i.e., $\inf_{c_1, \dots, c_p \in [\underline{\phi}_n, \bar{\phi}_n]} p(\tilde{\gamma}^0|Z) \rightarrow_p 1$ when $\tilde{p}^{\alpha_0+2} = o(n\underline{\phi}_n)$, since the proof of the first part is similar. It is enough to show that for any $\varepsilon > 0$, when $n \rightarrow \infty$

$$P \left(\inf_{c_1, \dots, c_p \in [\underline{\phi}_n, \bar{\phi}_n]} p(\tilde{\gamma}^0|Z) > 1 - \varepsilon \right) \rightarrow 1. \quad (2.24)$$

Let $E_N = \bigcap_{n \geq N} \{\gamma^0 \subset \hat{\gamma}^n\}$. The desirability of $\hat{\gamma}^n$ guarantees that $\lim_{N \rightarrow \infty} P(E_N) = 1$. I temporarily fix an N . On the event E_N , for any $n \geq N$, $\hat{\gamma}^n$ represents a well selected model that contains the true model $\tilde{\gamma}^0$. By the assumption that $\hat{\gamma}^n$ is independent of (\mathbf{y}, X) , and by following the arguments similar to the proofs of Theorem 2.2.2, it can be shown that $p(\tilde{\gamma}^0|Z)$ converges to 1 in terms of conditional probability induced by $\hat{\gamma}^n$, i.e., $P(\cdot|\hat{\gamma}^n)$. In other words, the following result holds when $n \rightarrow \infty$,

$$P \left(\inf_{c_1, \dots, c_p \in [\underline{\phi}_n, \bar{\phi}_n]} p(\tilde{\gamma}^0|Z) > 1 - \varepsilon \middle| \hat{\gamma}^n \right) \rightarrow 1, \text{ almost surely on } E_N. \quad (2.25)$$

So when $n \rightarrow \infty$, $P \left(\inf_{c_1, \dots, c_p \in [\underline{\phi}_n, \bar{\phi}_n]} p(\tilde{\gamma}^0|Z) > 1 - \varepsilon \middle| \hat{\gamma}^n \right) I(E_N) \rightarrow I(E_N)$, a.s. Meanwhile,

it is clear that the following decomposition holds,

$$\begin{aligned}
& P \left(\inf_{c_1, \dots, c_p \in [\underline{\phi}_n, \bar{\phi}_n]} p(\tilde{\gamma}^0 | Z) > 1 - \varepsilon \right) \\
&= E \left\{ P \left(\inf_{c_1, \dots, c_p \in [\underline{\phi}_n, \bar{\phi}_n]} p(\tilde{\gamma}^0 | Z) > 1 - \varepsilon \middle| \hat{\gamma}^n \right) I(E_N) \right\} \\
&+ E \left\{ P \left(\inf_{c_1, \dots, c_p \in [\underline{\phi}_n, \bar{\phi}_n]} p(\tilde{\gamma}^0 | Z) > 1 - \varepsilon \middle| \hat{\gamma}^n \right) I(E_N^c) \right\}.
\end{aligned}$$

I denote the above summands to be Q_1 and Q_2 respectively. Note that $Q_2 \leq P(E_N^c)$ can be arbitrarily small when N is large. Meanwhile, by bounded convergence theorem, letting n go to ∞ leads to $Q_1 \rightarrow P(E_N)$, which is close to one for large N . Therefore, (2.24) holds. \square

Remark A.3.

- (1). The key step in the proof of Theorem 2.7.1 is the limit (2.25), which has been proved in detail in Appendix B'.
- (2). Using a similar technique, Theorem 2.2.4 can be also generalized to the setting of this section, which deals with the case $\gamma^0 = \emptyset$.
- (3). Fan and Lv (2008) have shown the consistency of several non-Bayesian estimation approaches, such as SCAD, Adaptive-Lasso and Dantzig Selector, when combined with SIS. Theorem 2.7.1 demonstrates that, when combined with SIS, consistency for Bayesian variable selection procedures also holds. I should mention that the idea in the proof of Theorem 2.7.1 is rooted in Fan and Lv (2008). \square

Appendix B': Proof of (2.25).

I consider the settings in Section 7 and keep all the notation in the paper. And I also keep all the labels for the assumptions, theorems, lemmas and equations in the paper.

A key step in the proof of Theorem 2.7.1 is that when $n \rightarrow \infty$, (2.25) holds. In fact, (2.25) is a variation of Theorem 2.2.2 in the setting of conditional probability induced by $\hat{\gamma}^n$, i.e., $P(\cdot|\hat{\gamma}^n)$. In this short note, I briefly discuss the proof of (2.25). Note that $P\left(\inf_{c_1, \dots, c_p \in [\underline{\phi}_n, \bar{\phi}_n]} p(\tilde{\gamma}^0|Z) > 1 - \varepsilon \middle| \hat{\gamma}^n\right)(\omega) = P\left(\inf_{c_1, \dots, c_p \in [\underline{\phi}_n, \bar{\phi}_n]} p(\tilde{\gamma}^0|Z) > 1 - \varepsilon \middle| \hat{\gamma}^n = \gamma(n)\right)(\omega)$, a.s. $\omega \in \Omega$, where Ω denotes the underlying sample space. Therefore, to prove (2.25), I only need to prove that for any element $\omega \in E_N$,

$$P\left(\inf_{c_1, \dots, c_p \in [\underline{\phi}_n, \bar{\phi}_n]} p(\tilde{\gamma}^0|Z) > 1 - \varepsilon \middle| \hat{\gamma}^n = \gamma(n)\right) \rightarrow 1, \quad (2.26)$$

where $\gamma(n) := \hat{\gamma}^n(\omega)$. Since $\omega \in E_N$, for any $n \geq N$, $\tilde{\gamma}^0 \subset \gamma(n)$.

Next, I will prove (2.26). The proof can be finished by arguments similar to the proof of Theorem 2.2.2. I denote $S_1 = \{\tilde{\gamma} \subset \gamma(n) | \tilde{\gamma}^0 \subset \tilde{\gamma}, \tilde{\gamma} \neq \tilde{\gamma}^0\}$ and $S_2 = \{\tilde{\gamma} \subset \gamma(n) | \tilde{\gamma}^0 \text{ is not nested in } \tilde{\gamma}\}$. Thus, $S_1 \cup S_2 \cup \{\tilde{\gamma}^0\}$ is the collection of all submodels of $\gamma(n)$. Define $P_{\tilde{\gamma}} = \tilde{X}_{\tilde{\gamma}}(\tilde{X}_{\tilde{\gamma}}^T \tilde{X}_{\tilde{\gamma}})^{-1} \tilde{X}_{\tilde{\gamma}}^T$ for any $\tilde{\gamma} \subset \gamma(n)$.

I first need to establish a variation of Lemma 2.6.1 in terms of $P(\cdot|\hat{\gamma}^n)$, which has been stated as follows.

Lemma 2.7.2. Suppose $\epsilon \sim N(0, \sigma_0^2 I_n)$ is independent of $\hat{\gamma}^n$. Then:

- (a). Let $v_{\tilde{\gamma}} = (I_n - P_{\tilde{\gamma}}) \tilde{X}_{\tilde{\gamma}^0 \setminus \tilde{\gamma}} \tilde{\beta}_{\tilde{\gamma}^0 \setminus \tilde{\gamma}}^0$. Let S_2 be nonnull. Then for any $\varepsilon > 0$, there is a constant C_ε such that

$$P\left(\max_{\tilde{\gamma} \in S_2} |v_{\tilde{\gamma}}^T \epsilon| / \|v_{\tilde{\gamma}}\|_2 > C_\varepsilon \sqrt{\tilde{p}} \middle| \hat{\gamma}^n = \gamma(n)\right) < \varepsilon, \text{ for any } n \geq 1,$$

where I have adopted the convention that $|v_{\tilde{\gamma}}^T \epsilon| / \|v_{\tilde{\gamma}}\|_2 = 0$ when $v_{\tilde{\gamma}} = 0$.

(b). If S_1 is nonnull, then for any $\alpha > 2$,

$$P \left(\max_{\tilde{\gamma} \in S_1} \epsilon^T (P_{\tilde{\gamma}} - P_{\tilde{\gamma}^0}) \epsilon / (|\tilde{\gamma}| - s_n) \leq \alpha \sigma_0^2 \log \tilde{p} \middle| \hat{\gamma}^n = \gamma(n) \right) \rightarrow 1.$$

(c). If S_2 is nonnull, and I adopt the convention that $\epsilon^T P_{\tilde{\gamma}} \epsilon / |\tilde{\gamma}| = 0$ when $\tilde{\gamma}$ is null, then for any $\alpha > 2$,

$$P \left(\max_{\tilde{\gamma} \in S_2} \epsilon^T P_{\tilde{\gamma}} \epsilon / |\tilde{\gamma}| \leq \alpha \sigma_0^2 \log \tilde{p} \middle| \hat{\gamma}^n = \gamma(n) \right) \rightarrow 1.$$

Proof of Lemma 2.7.2.

(a). Since $\hat{\gamma}^n$ is independent of ϵ and X ,

$$\begin{aligned} & P \left(\max_{\tilde{\gamma} \in S_2} |v_{\tilde{\gamma}}^T \epsilon| / \|v_{\tilde{\gamma}}\|_2 > C_\varepsilon \sqrt{\tilde{p}} \middle| \hat{\gamma}^n = \gamma(n) \right) \\ & \leq \sum_{\tilde{\gamma} \in S_2} P \left(|v_{\tilde{\gamma}}^T \epsilon| / \|v_{\tilde{\gamma}}\|_2 > C_\varepsilon \sqrt{\tilde{p}} \middle| \hat{\gamma}^n = \gamma(n) \right) \\ & = \sum_{\tilde{\gamma} \in S_2} P \left(|v_{\tilde{\gamma}}^T \epsilon| / \|v_{\tilde{\gamma}}\|_2 > C_\varepsilon \sqrt{\tilde{p}} \right) \\ & \leq C_0 2^{\tilde{p}} \exp \left(-\frac{C_\varepsilon^2 \tilde{p}}{2\sigma_0^2} \right) \rightarrow 0, \text{ when } C_\varepsilon \text{ is large and } \tilde{p} = \tilde{p}_n \rightarrow \infty \text{ as } n \rightarrow \infty. \end{aligned}$$

Here, C_0 is the constant identified as in Lemma 2.6.1 (a). Thus, the conclusion holds.

(b). Note that

$$\begin{aligned} & P \left(\max_{\tilde{\gamma} \in S_1} \epsilon^T (P_{\tilde{\gamma}} - P_{\tilde{\gamma}^0}) \epsilon / (|\tilde{\gamma}| - s_n) \leq \alpha \sigma_0^2 \log \tilde{p} \middle| \hat{\gamma}^n = \gamma(n) \right) \\ & \leq \sum_{\tilde{\gamma} \in S_1} P \left(\epsilon^T (P_{\tilde{\gamma}} - P_{\tilde{\gamma}^0}) \epsilon / (|\tilde{\gamma}| - s_n) \leq \alpha \sigma_0^2 \log \tilde{p} \middle| \hat{\gamma}^n = \gamma(n) \right). \end{aligned}$$

By independence between $\hat{\gamma}^n$ and (ϵ, X) , conditioning on the event $\{\hat{\gamma}^n = \gamma(n)\}$, $\epsilon^T (P_{\tilde{\gamma}} - P_{\tilde{\gamma}^0}) \epsilon$ is distributed as $\sigma_0^2 \chi_{|\tilde{\gamma}| - |\tilde{\gamma}^0|}^2$. Then the conclusion follows by arguments similar to

the proof of Lemma 2.6.1 (b).

(c). The proof follows by arguments similar to (b). \square

Let T_1, T_2, T_3, T_4, T_5 be defined as in (2.12). Note that the term T_1 is bounded below by Assumption 2.2.1, the term T_2 can be approximated similarly to Lemma 2.6.2, and the term T_3 is always non-negative.

By independence between (\mathbf{y}, X) and $\hat{\gamma}^n$, and the arguments similar to (2.14), it can be shown that for any $\eta > 0$, there exists $C_\eta > 0$ such that for any n

$$P \left(\inf_{c_1, \dots, c_p \in [\underline{\phi}_n, \bar{\phi}_n]} (-T_4) \geq C_\eta \middle| \hat{\gamma}^n = \gamma(n) \right) \leq \eta,$$

i.e., $-T_4$ is stochastically bounded in terms of $P(\cdot | \hat{\gamma}^n = \gamma(n))$.

Next, I approximate T_5 under $P(\cdot | \hat{\gamma}^n = \gamma(n))$. By Lemma 2.6.1, independence between (\mathbf{y}, X) and $\hat{\gamma}^n$, and by the proof of Lemma 2.6.3, one can show that for some constant C'

$$P \left(\text{uniformly for } \tilde{\gamma} \in S_2, T_5 \geq \frac{n + \nu}{2} \log (1 + C' \psi_n^2) \middle| \hat{\gamma}^n = \gamma(n) \right)$$

is large for sufficiently large n , and for sufficiently large n ,

$$P \left(\text{uniformly for } \tilde{\gamma} \in S_1, T_5 \geq -2^{-1}(|\tilde{\gamma} - s_n|)\alpha_0 \log p \middle| \hat{\gamma}^n = \gamma(n) \right)$$

is large. Then by arguments similar to (2.19), (2.20), (2.21) and (2.21) in the proof Theorem 2.2.2, one can show that (2.26) holds. \square

2.8 Appendix C: Proof of Corollaries 2.2.5 and 2.2.6

Proof of Corollary 2.2.5. It is easy to see that Assumption 2.2.2 holds with $C_1 = C_2 = 1$ and Assumption 2.2.8 holds with $\delta = 0$ and $C_3 = 1$. I still use the method of proof of Theorem 2.2.2. Note that the following equation holds,

$$\begin{aligned} -\log(p(\gamma|Z)/p(\gamma^0|Z)) &= -\log\left(\frac{p(\gamma)}{p(\gamma^0)}\right) + \frac{1}{2}\log\left(\frac{\det(W_\gamma)}{\det(W_{\gamma^0})}\right) \\ &\quad + \frac{n+\nu}{2}\log\left(\frac{1+\mathbf{y}^T(I_n - X_\gamma U_\gamma^{-1} X_\gamma^T)\mathbf{y}}{1+\mathbf{y}^T(I_n - X_{\gamma^0} U_{\gamma^0}^{-1} X_{\gamma^0}^T)\mathbf{y}}\right). \end{aligned} \quad (2.27)$$

Denote the above summands to be J_1, J_2, J_3 . Following Assumption 2.2.1, J_1 is bounded below. By Lemma 2.6.2, if $\gamma \in S_1$, then $J_2 \geq 2^{-1}(|\gamma| - s_n)\log(1 + n\phi_n)$; if $\gamma \in S_2$, then $J_2 \geq -2^{-1}s_n\log(1 + n\phi_n)$.

Now I approximate J_3 . Note that

$$E((\beta_{\gamma^0}^0)^T X_{\gamma^0}^T \epsilon \epsilon^T X_{\gamma^0} \beta_{\gamma^0}^0) = \sigma_0^2 n k_n,$$

and hence $|(\beta_{\gamma^0}^0)^T X_{\gamma^0}^T \epsilon|/(n k_n) = O_p(\sigma_0(n k_n)^{-1/2}) = O_p(1/(\sqrt{n}\psi_n)) = o_p(1)$. By a direct calculation and $s_n = o(n)$ (Assumption 2.2.9(i)), and the fact that $U_\gamma = (n + \phi_n^{-1})I_{|\gamma|}$ (by orthogonality of X), I can show that

$$\begin{aligned} &\mathbf{y}^T(I_n - X_{\gamma^0} U_{\gamma^0}^{-1} X_{\gamma^0}^T)\mathbf{y} = \mathbf{y}^T(I_n - X_{\gamma^0} X_{\gamma^0}^T/(n + \phi_n^{-1}))\mathbf{y} \\ &= \frac{n\phi_n^{-1}}{n + \phi_n^{-1}}\|\beta_{\gamma^0}^0\|_2^2 + \frac{2\phi_n^{-1}}{n + \phi_n^{-1}}(\beta_{\gamma^0}^0)^T X_{\gamma^0}^T \epsilon + \epsilon^T(I_n - X_{\gamma^0} X_{\gamma^0}^T/(n + \phi_n^{-1}))\epsilon \\ &= \frac{n\phi_n^{-1}}{n + \phi_n^{-1}}k_n \left(1 + \frac{2(\beta_{\gamma^0}^0)^T X_{\gamma^0}^T \epsilon}{n k_n}\right) + n\sigma_0^2(1 + o_p(1)) \\ &= \left(\frac{n\phi_n^{-1}}{n + \phi_n^{-1}}k_n + n\sigma_0^2\right)(1 + o_p(1)) = \sigma_0^2 a_n(1 + o_p(1)). \end{aligned} \quad (2.28)$$

Let $\nu_\gamma = (I_n - P_\gamma)X_{\gamma^0 \setminus \gamma} \beta_{\gamma^0 \setminus \gamma}^0$. By the orthogonality of X , $\nu_\gamma = X_{\gamma^0 \setminus \gamma} \beta_{\gamma^0 \setminus \gamma}^0$. Then

$$\max_{\gamma \in S_2} \frac{|\nu_\gamma^T \epsilon|}{\|\nu_\gamma\|_2} = \max_{\gamma \in S_2} \frac{|(\beta_{\gamma^0 \setminus \gamma}^0)^T X_{\gamma^0 \setminus \gamma}^T \epsilon|}{\|X_{\gamma^0 \setminus \gamma} \beta_{\gamma^0 \setminus \gamma}^0\|_2} \leq \max_{\gamma \subset \gamma^0} \frac{|(\beta_\gamma^0)^T X_\gamma^T \epsilon|}{\|X_\gamma \beta_\gamma^0\|_2}.$$

Therefore, by Bonferroni's inequality, I get that

$$\begin{aligned} P \left(\max_{\gamma \in S_2} \frac{|\nu_\gamma^T \epsilon|}{\|\nu_\gamma\|_2} > t \right) &\leq P \left(\max_{\gamma \subset \gamma^0} \frac{|(\beta_\gamma^0)^T X_\gamma^T \epsilon|}{\|X_\gamma \beta_\gamma^0\|_2} > t \right) \\ &\leq \sum_{\gamma \subset \gamma^0} P \left(\frac{|(\beta_\gamma^0)^T X_\gamma^T \epsilon|}{\|X_\gamma \beta_\gamma^0\|_2} > t \right) \\ &\leq C_0 2^{s_n} \exp \left(-\frac{t^2}{2\sigma_0^2} \right). \end{aligned}$$

Letting $t = C\sigma_0\sqrt{2s_n}$ for a sufficiently large C , I can show that $\max_{\gamma \in S_2} |\nu_\gamma^T \epsilon|/\|\nu_\gamma\|_2 = O_p(\sqrt{s_n})$.

By Assumption 2.2.9 (i), $s_n = o(n\psi_n^2)$. Therefore, by a similar proof to (2.16), uniformly for $\gamma \in S_2$,

$$\mathbf{y}^T (I_n - X_\gamma U_\gamma^{-1} X_\gamma^T) \mathbf{y} \geq \mathbf{y}^T (I_n - P_\gamma) \mathbf{y} \geq n\psi_n^2(1 + o_p(1)). \quad (2.29)$$

Consequently, by (2.28) and (2.29), I can see that with large probability and uniformly for $\gamma \in S_2$,

$$J_3 \geq \frac{n + \nu}{2} \log \left(\frac{n\psi_n^2}{\sigma_0^2 \zeta a_n} \right).$$

When $\gamma \in S_1$,

$$\begin{aligned} \frac{1 + \mathbf{y}^T (I_n - X_\gamma U_\gamma^{-1} X_\gamma^T) \mathbf{y}}{1 + \mathbf{y}^T (I_n - X_{\gamma^0} U_{\gamma^0}^{-1} X_{\gamma^0}^T) \mathbf{y}} &= 1 - \frac{\mathbf{y}^T (X_\gamma X_\gamma^T - X_{\gamma^0} X_{\gamma^0}^T) \mathbf{y} / (n + \phi_n^{-1})}{1 + \mathbf{y}^T (I_n - X_{\gamma^0} U_{\gamma^0}^{-1} X_{\gamma^0}^T) \mathbf{y}} \\ &= 1 - \frac{\epsilon^T (X_\gamma X_\gamma^T - X_{\gamma^0} X_{\gamma^0}^T) \epsilon / (n + \phi_n^{-1})}{1 + \sigma_0^2 a_n (1 + o_p(1))}. \end{aligned}$$

Following the proofs of Lemma 3 by Meinshausen and Yu (2009), with large probability,

uniformly for $\gamma \in S_1$, $\epsilon^T(X_\gamma X_\gamma^T - X_{\gamma^0} X_{\gamma^0}^T)\epsilon \leq 2n\sigma_0^2(|\gamma| - s_n) \log p$. Note that $p \log p = o(a_n)$ (Assumption 2.2.9(ii)), and the inequality $\log(1 - x) \geq -(\alpha/2)x$ holds when $x \in (0, 1 - 2/\alpha)$ for any $2 < \alpha < \alpha_0$. Thus, when n is sufficiently large, with large probability, for any $\gamma \in S_1$,

$$\begin{aligned} J_3 &= \frac{n + \nu}{2} \log \left(1 - \frac{1}{n + \phi_n^{-1}} \frac{\epsilon^T(X_\gamma X_\gamma^T - X_{\gamma^0} X_{\gamma^0}^T)\epsilon}{1 + \sigma_0^2 a_n (1 + o_p(1))} \right) \\ &\geq \frac{n + \nu}{2} \log \left(1 - \frac{2(|\gamma| - s_n) \log p}{a_n} (1 + o_p(1)) \right) \\ &\geq -2^{-1}(|\gamma| - s_n) \frac{n + \nu}{a_n} \alpha_0 \log p. \end{aligned}$$

Then, $\max_{\gamma \neq \gamma^0} p(\gamma|Z)/p(\gamma^0|Z) \rightarrow 0$ in probability follows by $s_n = o\left(\frac{(n+\nu)\log(n\psi_n^2/(\sigma_0^2\zeta a_n))}{\log(1+n\phi_n)}\right)$ (Assumption 2.2.9 (i)) and assumption $p^{\alpha_0(n+\nu)/a_n} = o(n\phi_n)$, and by applying arguments similar to (2.19) and (2.20). By $p = o\left((n + \nu) \log\left(\frac{n\psi_n^2}{\sigma_0^2\zeta a_n}\right)\right)$, $p^{2+\alpha_0(n+\nu)/a_n} = o(n\phi_n)$, and arguments similar to (2.21) and (2.21), it can be shown that $\sum_{\gamma \neq \gamma^0} p(\gamma|Z)/p(\gamma^0|Z) \rightarrow_p 0$, and therefore $p(\gamma^0|Z) \rightarrow_p 1$. \square

Proof of Corollary 2.2.6. (a) Assumptions 2.2.1 and 2.2.6 are easy to verify. Define $\alpha_n \sim \beta_n$ to mean that there exist positive constants d_1 and d_2 such that $d_1 < \alpha_n/\beta_n < d_2$ when n goes to ∞ . It is easy to see that $a_n \sim n^{1+\delta_1-\delta_2}(\log n)^2$ and $n\psi_n^2 \sim n^{2+\delta_1}(\log n)^2$, therefore $n\psi_n^2 \gg a_n$, so Assumption 2.2.9 can be verified by a direct calculation. Notice that $(n + \nu)/a_n \rightarrow 0$, $n^{\alpha_0(n+\nu)/a_n} = o(n\phi_n)$ holds for any α_0 . By Corollary 2.2.5, $\max_{\gamma \neq \gamma^0} p(\gamma|Z)/p(\gamma^0|Z) \rightarrow_p 0$. When $1 < \delta_2 \leq \delta_1$, it is easy to see that $n = o\left((n + \nu) \log\left(\frac{n\psi_n^2}{a_n}\right)\right)$ and $n^{2+\alpha_0(n+\nu)/a_n} = o(n\phi_n)$, so by Corollary 2.2.5, $p(\gamma^0|Z) \rightarrow_p 1$.

Next, I assume $-1 < \delta_2 \leq 1$ and identify the limit of $p(\gamma^0|Z)$. Without loss of generality, let $\beta_j^0 \neq 0$, $1 \leq j \leq s$, and $\beta_{s+1}^0 = \dots = \beta_n^0 = 0$. For $s + 1 \leq j \leq n$, define $\gamma(j)$ to be an n -vector with 1s in the first s positions and j th position, and zero in others, i.e., $\gamma(j)_1 = \dots = \gamma(j)_s = \gamma(j)_j = 1$, $\gamma(j)_i = 0$ when $i \neq 1, \dots, s, j$. Clearly, each $\gamma(j)$

corresponds to a model in S_1 .

Denote $\gamma = \gamma(j)$ for some $s + 1 \leq j \leq n$, then $\det(W_\gamma) = (1 + n\phi_n)^{s+1}$ and $\det(W_{\gamma^0}) = (1 + n\phi_n)^s$. Let J_1 , J_2 and J_3 be defined as in (2.27). Consequently, $J_2 = \frac{1}{2} \log(1 + n\phi_n)$. Using the representation (2.27) and the fact that $J_1 = 0$ and J_3 is almost surely non-positive when $\gamma \in S_1$, I have

$$-\log(p(\gamma|Z)/p(\gamma^0|Z)) \leq \frac{1}{2} \log(1 + n\phi_n), \text{ a.s.}, \quad (2.30)$$

and so $p(\gamma|Z)/p(\gamma^0|Z) \geq (1 + n\phi_n)^{-1/2}$, a.s., which leads to

$$\sum_{\gamma \neq \gamma^0} p(\gamma|Z)/p(\gamma^0|Z) \geq \sum_{j=s+1}^n p(\gamma(j)|Z)/p(\gamma^0|Z) \geq \frac{n-s}{(1+n\phi_n)^{1/2}}, \text{ a.s.} \quad (2.31)$$

When $-1 < \delta_2 < 1$, it follows from (2.31) that $\sum_{\gamma \neq \gamma^0} p(\gamma|Z)/p(\gamma^0|Z) \rightarrow \infty$, a.s. Therefore, $p(\gamma^0|Z) \rightarrow 0$, a.s. follows from relationship (2.8). When $\delta_2 = 1$, $\frac{n-s}{(1+n\phi_n)^{1/2}}$ converges to some positive constant c . It thus follows from (2.31) that $\limsup_n p(\gamma^0|Z) \leq c_0 := 1/(1+c)$, a.s.

(b) Using the relationship (2.8), it is sufficient to prove $p(\emptyset|Z)/p(\gamma^0|Z) \rightarrow \infty$. Let J_1 , J_2 , J_3 be defined as in the representation (2.27). Then $J_1 = 0$ follows from assumption. It is easy to see that $J_2 = -\frac{s}{2} \log(1 + n\phi_n)$. Next, I approximate J_3 . Since $n^{n \log n} = O(\phi_n)$, $a_n \sim n$. Thus, by the proof of (2.28), $\mathbf{y}^T \left(I_n - X_{\gamma^0} U_{\gamma^0}^{-1} X_{\gamma^0}^T \right) \mathbf{y} = \sigma_0^2 a_n (1 + o_p(1)) \sim \sigma_0^2 n (1 + o_p(1))$. Since $\epsilon^T X_{\gamma^0} \beta_{\gamma^0}^0 = O_p((nk_n)^{1/2})$, $\epsilon^T X_{\gamma^0} X_{\gamma^0}^T \epsilon = O_p(n)$ and $X_{\gamma^0}^T X_{\gamma^0} = nI_s$, and by the fact

that $nk_n \geq sn\psi_n^2 \rightarrow \infty$, I have

$$\begin{aligned}
\mathbf{y}^T X_{\gamma^0} X_{\gamma^0}^T \mathbf{y} &= (X_{\gamma^0} \beta_{\gamma^0}^0 + \epsilon)^T X_{\gamma^0} X_{\gamma^0}^T (X_{\gamma^0} \beta_{\gamma^0}^0 + \epsilon) \\
&= n^2 \|\beta_{\gamma^0}^0\|_2^2 + 2n \epsilon^T X_{\gamma^0} \beta_{\gamma^0}^0 + \epsilon^T X_{\gamma^0} X_{\gamma^0}^T \epsilon \\
&= n^2 k_n + O_p(n(nk_n)^{1/2}) + O_p(n) \\
&= n^2 k_n (1 + o_p(1)),
\end{aligned}$$

and so it is easy to see that

$$\begin{aligned}
J_3 &= \frac{n + \nu}{2} \log \left(1 + \frac{1}{n + \phi_n^{-1}} \frac{\mathbf{y}^T X_{\gamma^0} X_{\gamma^0}^T \mathbf{y}}{1 + \mathbf{y}^T (I_n - X_{\gamma^0} U_{\gamma^0}^{-1} X_{\gamma^0}^T) \mathbf{y}} \right) \\
&= O_p \left(\frac{n + \nu}{2} \log(1 + k_n) \right).
\end{aligned}$$

By a direct calculation, it is not hard to verify that $(n + \nu) \log(1 + k_n) \ll \log(1 + n\phi_n)$, so

$$\begin{aligned}
-\log(p(\emptyset|Z)/p(\gamma^0|Z)) &= -\frac{s}{2} \log(1 + n\phi_n) + O_p \left(\frac{n + \nu}{2} \log(1 + k_n) \right) \\
&= -\frac{s}{2} \log(1 + n\phi_n) (1 + o_p(1)),
\end{aligned}$$

which leads to $p(\emptyset|Z)/p(\gamma^0|Z) \rightarrow \infty$ in probability. Thus, $p(\gamma^0|Z) \rightarrow 0$ in probability follows immediately from (2.8).

(c) The proof can be finished by constructing a sequence of models containing the true model using the approach in part (a) and by arguments (2.30) and (2.31). \square

2.9 Appendix D: Almost Sure Consistency of $p(\gamma^0|Z)$

The consistency results in Theorems 2.2.2–2.2.4 hold in probability. I can also establish almost sure consistency. Before stating the exact results, I need to introduce the following technical condition which plays a role similar to that of Assumption 2.2.3, but imposes a greater restriction on ψ_n .

Assumption 2.9.1. There exists a positive sequence ψ_n such that $\min_{j \in \gamma^0} |\beta_j^0| \geq \psi_n$ and, as $n \rightarrow \infty$, $\psi_n \sqrt{n/\log n} \rightarrow \infty$.

With this assumption, I can derive the following results on almost sure convergence when the true model is either nonnull or null.

Theorem 2.9.1. Suppose γ^0 is nonnull, and Assumptions 2.2.1, 2.2.2, 2.2.4, 2.2.6–2.2.8 and 2.9.1 are satisfied. Let $\delta \geq 0$ satisfy Assumption 2.2.8. If $n^{1-\delta-\alpha_0} \underline{\phi}_n \rightarrow \infty$ for some $\alpha_0 > 4$, then $\sup_{c_1, \dots, c_p \in [\underline{\phi}_n, \bar{\phi}_n]} \max_{\gamma \neq \gamma^0} p(\gamma|Z)/p(\gamma^0|Z) \rightarrow 0$, a.s. If $p^2 = o(n^{1-\delta-\alpha_0} \underline{\phi}_n)$ for some $\alpha_0 > 4$, then $\sup_{c_1, \dots, c_p \in [\underline{\phi}_n, \bar{\phi}_n]} \sum_{\gamma \neq \gamma^0} p(\gamma|Z) \rightarrow 0$, a.s., and consequently, $\inf_{c_1, \dots, c_p \in [\underline{\phi}_n, \bar{\phi}_n]} p(\gamma^0|Z) \rightarrow 1$, a.s.

Theorem 2.9.2. Suppose γ^0 is null. Suppose that Assumptions 2.2.1, 2.2.5 and 2.2.8 are satisfied. If $n^{1-\delta-\alpha_0} \underline{\phi}_n \rightarrow \infty$ for some $\alpha_0 > 4$, then $\sup_{c_1, \dots, c_p \in [\underline{\phi}_n, \bar{\phi}_n]} \max_{\gamma \neq \gamma^0} p(\gamma|Z)/p(\gamma^0|Z) \rightarrow 0$, a.s. If $p^2 = o(n^{1-\delta-\alpha_0} \underline{\phi}_n)$ for some $\alpha_0 > 4$, then $\sup_{c_1, \dots, c_p \in [\underline{\phi}_n, \bar{\phi}_n]} \sum_{\gamma \neq \gamma^0} p(\gamma|Z) \rightarrow 0$, a.s., and consequently, $\inf_{c_1, \dots, c_p \in [\underline{\phi}_n, \bar{\phi}_n]} p(\gamma^0|Z) \rightarrow 1$, a.s.

The following lemma is useful to establish the almost sure convergence of $p(\gamma^0|Z)$.

Lemma 2.9.3. Let $\epsilon \sim N(0, \sigma_0^2 I_p)$.

(a). Let $v_\gamma = (I_n - P_\gamma)X_{\gamma^0 \setminus \gamma} \beta_{\gamma^0 \setminus \gamma}^0$. Then $\limsup_{n \rightarrow \infty} \max_{\gamma \in S_2} |v'_\gamma \epsilon| / (\|v_\gamma\|_2 \sqrt{p \log n}) < 2\sigma_0$, a.s.

(b). If S_1 is nonnull, then $\limsup_{n \rightarrow \infty} \max_{\gamma \in S_1} \epsilon^T (P_\gamma - P_{\gamma^0}) \epsilon / ((|\gamma| - s_n) \log n) \leq 4\sigma_0^2$, a.s.

(c). If S_2 is nonnull, then $\limsup_{n \rightarrow \infty} \max_{\gamma \in S_2} \epsilon^T P_\gamma \epsilon / (|\gamma| \log n) \leq 4\sigma_0^2$, a.s.

(d). $\limsup_n \epsilon^T P_{\gamma^0} \epsilon / (s_n \log n) \leq 4\sigma_0^2$, a.s.

Proof of Lemma 2.9.3. The proofs can be completed by using the Borel-Cantelli lemma and the techniques in the proof Lemma 2.6.1.

(a) Similar to the proof of part (a) in Lemma 2.6.1, it can be shown that

$$P \left(\max_{\gamma \in S_2} \frac{|v'_\gamma \epsilon|}{\|v_\gamma\|_2} \geq 2\sigma_0 \sqrt{p \log n} \right) \leq C_0 \left(\frac{2}{n^2} \right)^p,$$

therefore, $\sum_n P \left(\max_{\gamma \in S_2} |v'_\gamma \epsilon| / \|v_\gamma\|_2 \geq 2\sigma_0 \sqrt{2p \log n} \right) < \infty$. By the Borel-Cantelli lemma, the desired result holds.

(b) For any $\alpha > 4$, I temporarily fix $\alpha' > 2$ such that $2\alpha' < \alpha$. Following the proof of part (b) in Lemma 2.6.1, for large n ,

$$\begin{aligned} P \left(\max_{\gamma \in S_1} \frac{\epsilon^T (P_\gamma - P_{\gamma^0}) \epsilon}{(|\gamma| - s_n) \log n} \geq \alpha \sigma_0^2 \right) &\leq \left(1 + (1 - 2/\alpha')^{-1/2} n^{-\alpha/\alpha'} \right)^{p-s_n} - 1 \\ &\leq \exp \left((1 - 2/\alpha')^{-1/2} n^{-\alpha/\alpha'} p \right) - 1 \\ &\leq 2(1 - 2/\alpha')^{-1/2} n^{1-\alpha/\alpha'}. \end{aligned}$$

Therefore, by the Borel-Cantelli lemma, $\limsup_n \max_{\gamma \in S_1} \epsilon^T (P_\gamma - P_{\gamma^0}) \epsilon / ((|\gamma| - s_n) \log n) \leq \alpha \sigma_0^2$, a.s. Then the desired result holds by selecting a sequence of α s approaching 4.

(c) & (d) These proofs can be accomplished by arguments similar to part (b). \square

Proof of Theorem 2.9.1. I start with (2.12) to approximate the ratio $p(\gamma|Z)/p(\gamma^0|Z)$. The terms T_1 and T_3 in (2.12) are bounded below. The approximation of T_2 is given by Lemma 2.6.2. By Lemma 2.9.3 (d), Assumptions 2.2.2, 2.2.7 and 2.9.1, and a similar argument to (2.14), it can be shown that $0 \leq -T_4 = O(1)$, a.s.

To approximate T_5 , by Assumption 2.2.4, Lemma 2.9.3 and a careful revision of (2.16), (2.17) and (2.18), one can show that with probability equal to one, for some constant $C''' > 0$, there exists a large N such that if $n \geq N$, uniformly for $\gamma \in S_2$, $T_5 \geq 2^{-1}(n+\nu) \log(1+C''\psi_n^2)$; uniformly for $\gamma \in S_1$, $T_5 \geq -2^{-1}(|\gamma| - s_n)\alpha_0 \log n$.

Thus, by an argument similar to (2.20) and (2.21), it can be shown that

$$\sup_{c_1, \dots, c_p \in [\underline{\phi}_n, \bar{\phi}_n]} \sum_{\gamma \in S_2} p(\gamma|Z)/p(\gamma^0|Z) \rightarrow 0, \text{ a.s.},$$

which implies that

$$\sup_{c_1, \dots, c_p \in [\underline{\phi}_n, \bar{\phi}_n]} \max_{\gamma \in S_2} p(\gamma|Z)/p(\gamma^0|Z) \rightarrow 0, \text{ a.s.}$$

On the other hand, by the above approximations of T_1 to T_5 , with probability equal to one, there exists a constant $C''' > 0$ such that when n is sufficiently large, uniformly for $\gamma \in S_1$,

$$\begin{aligned} & \sup_{c_1, \dots, c_p \in [\underline{\phi}_n, \bar{\phi}_n]} p(\gamma|Z)/p(\gamma^0|Z) \\ & \leq C''' \exp \left(-2^{-1}(|\gamma| - s_n) \log(1 + C_3 n^{1-\delta} \underline{\phi}_n) + 2^{-1}(|\gamma| - s_n) \log n^{\alpha_0} \right) \\ & = C''' \left(\frac{1 + C_3 n^{1-\delta} \underline{\phi}_n}{n^{\alpha_0}} \right)^{-2^{-1}(|\gamma| - s_n)}. \end{aligned} \tag{2.32}$$

Thus, if $n^{1-\delta-\alpha_0} \underline{\phi}_n \rightarrow \infty$, then $\sup_{c_1, \dots, c_p \in [\underline{\phi}_n, \bar{\phi}_n]} \max_{\gamma \in S_1} p(\gamma|Z)/p(\gamma^0|Z) \rightarrow 0, \text{ a.s.}$

By (2.32), with probability equal to one,

$$\begin{aligned} \sup_{c_1, \dots, c_p \in [\underline{\phi}_n, \bar{\phi}_n]} \sum_{\gamma \in S_1} p(\gamma|Z)/p(\gamma^0|Z) &\leq C''' \sum_{\gamma \in S_1} \left(\frac{1 + C_3 n^{1-\delta} \underline{\phi}_n}{n^{\alpha_0}} \right)^{-2^{-1}(|\gamma| - s_n)} \\ &= C''' \left[\left(1 + \left(\frac{1 + C_3 n^{1-\delta} \underline{\phi}_n}{n^{\alpha_0}} \right)^{-1/2} \right)^{p - s_n} - 1 \right]. \end{aligned}$$

Thus, if $p^2 = o(n^{1-\delta-\alpha_0} \underline{\phi}_n)$, then $\sup_{c_1, \dots, c_p \in [\underline{\phi}_n, \bar{\phi}_n]} \sum_{\gamma \in S_1} p(\gamma|Z)/p(\gamma^0|Z) \rightarrow 0$, a.s., which, by (2.8), implies $\inf_{c_1, \dots, c_p \in [\underline{\phi}_n, \bar{\phi}_n]} p(\gamma^0|Z) \rightarrow 1$, a.s. \square

Proof of Theorem 2.9.2. Proof is completed by similar arguments to the proofs of Theorems 2.2.4 and 2.9.1. \square

Chapter 3

An Application of Bayesian Variable Selection to Spatial Concurrent Linear Models

3.1 Models and Algorithms

In this chapter, I develop our specific modeling approach to handle the spatial concurrent model (1.3). To simplify the details, I only consider $K = 1$ in model (1.3), i.e., only one slope surface is involved, although generalization to multiple slope surfaces is not difficult. Thus, model (1.3) becomes the following model with a single covariate surface x

$$y(\mathbf{s}_i) = A(\mathbf{s}_i) + x(\mathbf{s}_i)B(\mathbf{s}_i) + \epsilon(\mathbf{s}_i), \quad i = 1, \dots, n, \quad (3.1)$$

where $n = 4^{J+2}$, $\{\mathbf{s}_i\}_{i=1}^n = \{(2^{-J-2}k_1, 2^{-J-2}k_2) | k_1, k_2 = 0, 1, \dots, 2^{J+2} - 1\}$ is the set of locations evenly spaced over $[0, 1) \times [0, 1)$, and the $\epsilon(\mathbf{s}_i)$ s $\stackrel{iid.}{\sim} N(0, \sigma^2)$. By performing a two-dimensional Haar DWT with maximal level of decomposition J on A and B , model (3.1) can be written as a linear model $\mathbf{y} = X\beta + \epsilon$, which is a special case of (1.5) when $K = 1$. Here, X is the $n \times m$ design matrix induced by Haar DWT with $m = 2(4^{J+1})$, $\epsilon \sim N(\mathbf{0}, \sigma^2 I_n)$ is an n -vector of errors, and $\beta = [\mathbf{a}', \mathbf{b}']'$ with \mathbf{a} and \mathbf{b} being the $(m/2)$ -vectors of wavelet

coefficients corresponding to surfaces A and B .

Instead of imposing stationary prior distributions in the spatial domain of A and B , I assign mixture priors in the wavelet domain β corresponding to the resolution levels, which may produce nonstationary priors for A and B and accommodate more complex structures in spatial domain. Even if the components of β are assumed to be *a priori* independent, when $\mathbf{s} \neq \tilde{\mathbf{s}}$, $A(\mathbf{s})$ and $A(\tilde{\mathbf{s}})$, $B(\mathbf{s})$ and $B(\tilde{\mathbf{s}})$ may still be spatially correlated. In fact, as \mathbf{s} and $\tilde{\mathbf{s}}$ become closer in space, $A(\mathbf{s})$ and $A(\tilde{\mathbf{s}})$, $B(\mathbf{s})$ and $B(\tilde{\mathbf{s}})$ will share more common wavelet coefficients in their wavelet expansions, which makes their spatial correlations stronger.

I will consider two different Bayesian models and provide corresponding MCMC algorithms. In both models, I assume

$$\mathbf{y}|X, \beta, \sigma^2 \sim N(X\beta, \sigma^2 I_n), \quad 1/\sigma^2 \sim \chi_\nu^2,$$

where ν is a fixed hyperparameter. Let $\gamma = (\gamma_1, \dots, \gamma_m)$ with γ_j s being the 0-1 Bernoulli variables indicating the exclusion and inclusion of β_j s. In both models I place Bernoulli priors on γ , i.e., $p(\gamma_1, \dots, \gamma_m) = \prod_{j=1}^m \theta_j^{\gamma_j} (1 - \theta_j)^{1-\gamma_j}$, where $\theta_j = p(\gamma_j = 1)$ is the inclusion probability. However, I consider different priors for β .

Our first Bayesian model requires all the nonzero components of β to possess a common prior variance τ^2 . Given γ and τ^2 , the β_j s are independent with mixture priors.

$$\textbf{Model I:} \quad \beta_j | \gamma_j, \tau^2 \sim (1 - \gamma_j) \delta_0 + \gamma_j N(0, \tau^2), \quad 1/\tau^2 \sim \chi_\mu^2,$$

where μ is fixed. Based on Model I, the posterior distribution of $(\beta, \gamma, \sigma^2, \tau^2)$ is

$$\begin{aligned} & p(\beta, \gamma, \sigma^2, \tau^2 | \mathbf{y}, X) \\ & \propto \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp(-\|\mathbf{y} - X\beta\|^2 / (2\sigma^2)) \prod_{\gamma_j=1} \left(\frac{1}{\sqrt{2\pi}\tau} \right) \exp(-\beta_j^2 / (2\tau^2)) \\ & \cdot \prod_{\gamma_j=0} \delta_0(\beta_j) \frac{2^{-\nu/2}}{\Gamma(\nu/2)} \sigma^{-\nu-2} \exp(-1/(2\sigma^2)) \frac{2^{-\mu/2}}{\Gamma(\mu/2)} \tau^{-\mu-2} \exp(-1/(2\tau^2)) p(\gamma). \quad (3.2) \end{aligned}$$

If $\tau = \sigma$, then Model I is similar to one proposed by Clyde *et al.* (1998) and Li and Zhang (2010). Here I do not assume that the variances of the coefficients are related to σ , which makes our model flexible. A blockwise Gibbs sampler introduced by Godsill and Rayner (1998) and Wolfe *et al.* (2004) will be used to draw samples from the posterior distribution, as I now describe.

Algorithm I. Given a current state $(\beta^{(t)}, \gamma^{(t)}, \sigma^{(t)}, \tau^{(t)})$.

(A) Update (γ, β) :

$$p(\gamma_j^{(t+1)} = 1 | \beta_{-j}, \gamma_{-j}, \sigma^{(t)}, \tau^{(t)}, \mathbf{y}, X) = \frac{1}{1 + \rho_j},$$

$$p(\beta_j^{(t+1)} = 0 | \gamma_j^{(t+1)} = 0, \beta_{-j}, \gamma_{-j}, \sigma^{(t)}, \tau^{(t)}, \mathbf{y}, X) = 1,$$

$$\beta_j^{(t+1)} | \gamma_j^{(t+1)} = 1, \beta_{-j}, \gamma_{-j}, \sigma^{(t)}, \tau^{(t)}, \mathbf{y}, X \sim N\left(\frac{u_j}{v_j^2}, \frac{(\sigma^{(t)})^2}{v_j^2}\right),$$

where $\gamma_{-j} = (\gamma_1^{(t+1)}, \dots, \gamma_{j-1}^{(t+1)}, \gamma_{j+1}^{(t)}, \dots, \gamma_m^{(t)})'$, $\beta_{-j} = (\beta_1^{(t+1)}, \dots, \beta_{j-1}^{(t+1)}, \beta_{j+1}^{(t)}, \dots, \beta_m^{(t)})'$,

$$u_j = (\mathbf{y} - X_{-j}\beta_{-j})'X_j, \quad v_j = \left(X_j'X_j + \frac{(\sigma^{(t)})^2}{(\tau^{(t)})^2}\right)^{1/2}$$

with X_j being the j -th column of X and $X_{-j} = (X_1, \dots, X_{j-1}, X_{j+1}, \dots, X_m)$, and

$$\rho_j = \frac{p(\gamma_j = 0 | \gamma_{-j})}{p(\gamma_j = 1 | \gamma_{-j})} \frac{\tau^{(t)} v_j}{\sigma^{(t)}} \exp \left(-\frac{u_j^2}{2(\sigma^{(t)})^2 v_j^2} \right).$$

(B) Update (σ, τ) :

$$(\sigma^{(t+1)})^2 | \gamma^{(t+1)}, \beta^{(t+1)}, \tau^{(t)}, \mathbf{y}, X \sim IG \left(\frac{n + \nu}{2}, \frac{1 + \|\mathbf{y} - X_{\gamma^{(t+1)}} \beta_{\gamma^{(t+1)}}^{(t+1)}\|_2^2}{2} \right),$$

$$(\tau^{(t+1)})^2 | \gamma^{(t+1)}, \beta^{(t+1)}, \sigma^{(t+1)}, \mathbf{y}, X \sim IG \left(\frac{|\gamma^{(t+1)}| + \mu}{2}, \frac{1 + \|\beta^{(t+1)}\|_2^2}{2} \right),$$

where $IG(a, b)$ denotes the inverse gamma distribution with density $g(x) \propto x^{-a-1} \exp(-b/x)$ for $x > 0$.

The derivation of Algorithm I can be found in Appendix. Unlike the usual non-blockwise Gibbs sampler, Algorithm I involves no matrix inversion, and hence, is computationally efficient when m is moderate. However, when m is large, a direct application of Algorithm I will still be time-consuming because evaluating the quantity u_j in step (A) involves intensive matrix multiplication. To address this problem, I notice that $V_j = \mathbf{y} - X_{-j} \beta_{-j}$ and $V_{j-1} = \mathbf{y} - X_{-(j-1)} \beta_{-(j-1)}$ satisfy

$$V_j = V_{j-1} + \beta_j^{(t)} X_j - \beta_{j-1}^{(t+1)} X_{j-1}. \quad (3.3)$$

By (3.3), V_j can be obtained directly through V_{j-1} , which is available from the last updating. This effectively avoids unnecessary matrix multiplications in each iteration. A technique similar in spirit to (3.3) to reduce the computational burden was employed by Li and Zhang (2010), who proposed a non-blockwise Gibbs sampler for high-dimensional structured models.

In Model I, the prior variances of the nonzero β_j s have been set to be a common hyperparameter τ^2 , which seems restrictive. Our second Bayesian model overcomes this restriction by introducing different prior variances τ_j^2 s for β_j s. Given γ and τ_j^2 s, I assume the β_j s are independent with mixture priors as follows:

$$\textbf{Model II:} \quad \beta_j | \gamma_j, \tau_j^2 \sim (1 - \gamma_j) \delta_0 + \gamma_j N(0, \tau_j^2), \quad 1/\tau_1^2, \dots, 1/\tau_m^2 \stackrel{iid.}{\sim} \chi_\mu^2,$$

where μ is fixed. Based on Model II, the posterior distribution of $(\beta, \gamma, \sigma^2, \tau_1^2, \dots, \tau_m^2)$ is

$$\begin{aligned} & p(\beta, \gamma, \sigma^2, \tau_1^2, \dots, \tau_m^2 | \mathbf{y}, X) \\ & \propto \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp(-\|\mathbf{y} - X\beta\|^2 / (2\sigma^2)) \prod_{\gamma_j=1} \left(\frac{1}{\sqrt{2\pi}\tau_j} \right) \exp(-\beta_j^2 / (2\tau_j^2)) \\ & \quad \cdot \prod_{\gamma_j=0} \delta_0(\beta_j) \frac{2^{-\nu/2}}{\Gamma(\nu/2)} \sigma^{-\nu-2} \exp(-1/(2\sigma^2)) \\ & \quad \cdot \prod_{j=1}^m \frac{2^{-\mu/2}}{\Gamma(\mu/2)} \tau_j^{-\mu-2} \exp(-1/(2\tau_j^2)) p(\gamma). \end{aligned} \quad (3.4)$$

Using the blockwise technique, one can draw posterior samples from $p(\beta, \gamma, \sigma^2, \tau_1^2, \dots, \tau_m^2 | \mathbf{y}, X)$ with the following algorithm:

Algorithm II.

Given a current state $(\beta^{(t)}, \gamma^{(t)}, \sigma^{(t)}, \tau_1^{(t)}, \dots, \tau_m^{(t)})$.

(A) Update (γ, β) :

$$\begin{aligned} p(\gamma_j^{(t+1)} = 1 | \beta_{-j}, \gamma_{-j}, \tau_1^{(t)}, \dots, \tau_m^{(t)}, \sigma^{(t)}, \mathbf{y}, X) &= \frac{1}{1 + \rho_j}, \\ p(\beta_j^{(t+1)} = 0 | \gamma_j^{(t+1)} = 0, \beta_{-j}, \gamma_{-j}, \tau_1^{(t)}, \dots, \tau_m^{(t)}, \sigma^{(t)}, \mathbf{y}, X) &= 1, \end{aligned}$$

$$\beta_j^{(t+1)} | \gamma_j^{(t+1)} = 1, \beta_{-j}, \gamma_{-j}, \tau_1^{(t)}, \dots, \tau_m^{(t)}, \sigma^{(t)}, \mathbf{y}, X \sim N \left(\frac{u_j}{v_j^2}, \frac{(\sigma^{(t)})^2}{v_j^2} \right),$$

$$\text{where } \gamma_{-j} = \left(\gamma_1^{(t+1)}, \dots, \gamma_{j-1}^{(t+1)}, \gamma_{j+1}^{(t)}, \dots, \gamma_m^{(t)} \right)', \beta_{-j} = \left(\beta_1^{(t+1)}, \dots, \beta_{j-1}^{(t+1)}, \beta_{j+1}^{(t)}, \dots, \beta_m^{(t)} \right)',$$

$$u_j = (\mathbf{y} - X_{-j} \beta_{-j})' X_j, \quad v_j = \left(X_j' X_j + \frac{(\sigma^{(t)})^2}{(\tau_j^{(t)})^2} \right)^{1/2}$$

with X_j being the j -th column of X and $X_{-j} = (X_1, \dots, X_{j-1}, X_{j+1}, \dots, X_m)$, and

$$\rho_j = \frac{p(\gamma_j = 0 | \gamma_{-j})}{p(\gamma_j = 1 | \gamma_{-j})} \frac{\tau_j^{(t)} v_j}{\sigma^{(t)}} \exp \left(-\frac{u_j^2}{2(\sigma^{(t)})^2 v_j^2} \right).$$

(B) Update τ_j :

$$(\tau_j^{(t+1)})^2 | \beta_j^{(t+1)}, \gamma_j^{(t+1)} = 0, \mathbf{y}, X \sim 1/\chi_\mu^2,$$

$$(\tau_j^{(t+1)})^2 | \beta_j^{(t+1)}, \gamma_j^{(t+1)} = 1, \mathbf{y}, X \sim IG \left(\frac{1 + \mu}{2}, \frac{1 + (\beta_j^{(t+1)})^2}{2} \right), j = 1, \dots, m.$$

(C) Update σ :

$$(\sigma^{(t+1)})^2 | \gamma^{(t+1)}, \beta^{(t+1)}, \mathbf{y}, X \sim IG \left(\frac{n + \nu}{2}, \frac{1 + \|\mathbf{y} - X_{\gamma^{(t+1)}} \beta_{\gamma^{(t+1)}}^{(t+1)}\|_2^2}{2} \right).$$

The derivation of Algorithm II is similar to that of Algorithm I. Since $2m + 2$ parameters have been involved in Model I, while $3m + 1$ parameters have been involved in Model II, it takes more time to use Algorithm II than Algorithm I for MCMC sampling. However, Bayesian estimates resulting from Model II may sometimes have better performance than those resulting from Model I, which will be seen in next section. To reduce computational cost, a technique similar to (3.3) will also be applied to Algorithm II.

3.2 Numerical Results

In this section, I apply the Bayesian methods developed in Section 2 to the concurrent linear model (3.1) and illustrate these methods with simulated and real datasets. In Section 3.1, I consider the problem of reconstructing both intercept and slope surfaces, and use them to obtain the fitted response surface. I assess the performance of Models I and II through four criteria: squared bias, variance, mean square error for the estimate of the coefficient surface, and mean square error for the response. Comparison with the LASSO approach proposed by Zhang *et al.* (2011) will also be demonstrated. In Section 3.2, I try to find the locations where the relationship between the response and the covariate is strong. In Section 3.3, I apply our methods to gypsy moth defoliation data.

Let $\{\mathbf{s}_i\}_{i=1}^n$ be the lattice set of locations specified in Section 2. Denote $\mathbf{A} = (A(\mathbf{s}_1), \dots, A(\mathbf{s}_n))'$ and $\mathbf{B} = (B(\mathbf{s}_1), \dots, B(\mathbf{s}_n))'$. After obtaining the estimates $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$ of \mathbf{a} and \mathbf{b} , I perform an inverse DWT to obtain the estimates of \mathbf{A} and \mathbf{B} through $\hat{\mathbf{A}} = W\hat{\mathbf{a}}$ and $\hat{\mathbf{B}} = W\hat{\mathbf{b}}$, where $W \in \mathbb{R}^{n \times \frac{m}{2}}$ corresponds to the two-dimensional Haar DWT and satisfies $W'W = I_{m/2}$.

The Markov chains simulated from posterior likelihoods (3.2) and (3.4) will converge quickly if the initial points of these chains are carefully selected. Here, I adopt an empirical procedure for this purpose. I first let $\hat{\beta} = (X'X)^{-1}X'\mathbf{y}$ be the least squares estimate, then I choose the initial point $\beta^{(0)}$ for the Markov chains as a draw from $N(\hat{\beta}, \tilde{\sigma}^2 I_m)$ with $\tilde{\sigma}^2$ predetermined to be the variance of $\beta^{(0)}$.

3.2.1 Assessing the Performance of Models I and II

I assessed the performance of Models I and II through the numerical results by Algorithms I and II. I chose the true intercept surface to be

$$A(s_1, s_2) = \begin{cases} 1, & 0 \leq s_1 < 0.5, 0 \leq s_2 < 0.5 \\ 4, & 0.5 \leq s_1 < 1, 0 \leq s_2 < 0.5 \\ 7, & 0 \leq s_1 < 0.5, 0.5 \leq s_2 < 1 \\ 10, & 0.5 \leq s_1 < 1, 0.5 \leq s_2 < 1, \end{cases}$$

and considered two different slope surfaces: (Case I)

$$B(s_1, s_2) = \begin{cases} 1, & 0 \leq s_1 < 0.47, 0 \leq s_2 < 0.5 \\ 3, & 0.47 \leq s_1 < 1, 0 \leq s_2 < 0.5 \\ 5, & 0 \leq s_1 < 0.5, 0.5 \leq s_2 < 1 \\ 7, & 0.5 \leq s_1 < 1, 0.5 \leq s_2 < 1, \end{cases}$$

and (Case II) $B(s_1, s_2) = 4 \sin(2\pi s_1) \cos(2\pi s_2)$, for $0 \leq s_1, s_2 < 1$.

To further explore the role played by the covariate surface, three covariate surfaces with different types of oscillation were considered:

$$x_a(s_1, s_2) = 4 \sin(4\pi(s_1 + s_2)), \quad (3.5)$$

$$x_b(s_1, s_2) = 4 \sin(10\pi(s_1 + s_2)), \quad (3.6)$$

$$x_c(s_1, s_2) = 4 \sin(15\pi(s_1 + s_2)), \quad 0 \leq s_1, s_2 \leq 1. \quad (3.7)$$

I chose $J = 3$ and generated data from model (3.1) with $\sigma = 1$. Therefore, $n = 1024$

and $m = 512$. There are 3 nonzero wavelet coefficients for A . In Case I, B is locally flat corresponding to 3 nonzero wavelet coefficients. (Recall that I are using Haar wavelets.) However, in Case II, B has little local flatness and all 256 wavelet coefficients of B are nonzero. I fixed $\mu = \nu = 6$. Let $\{a_0, a_{jk}^r | r = 1, 2, 3, j = 0, 1, \dots, J, k \in \Lambda_j\}$ and $\{b_0, b_{jk}^r | r = 1, 2, 3, j = 0, 1, \dots, J, k \in \Lambda_j\}$ be the components of \mathbf{a} and \mathbf{b} , and $\gamma_0^a = I(a_0 \neq 0)$, $\gamma_0^b = I(b_0 \neq 0)$, $\gamma_{jkr}^a = I(a_{jk}^r \neq 0)$, $\gamma_{jkr}^b = I(b_{jk}^r \neq 0)$, where j denotes the resolution level of the wavelet coefficients and Λ_j denotes the collection of the indexes of the wavelet coefficients at the j -th resolution level. I considered the following three different Bernoulli priors for γ .

Prior (1):

$$p(\gamma_0^a = 1) = p(\gamma_0^b = 1) = 0.5, \quad p(\gamma_{jkr}^a = 1) = p(\gamma_{jkr}^b = 1) = 0.5\phi^j, \quad r = 1, 2, 3, j = 0, \dots, J, \quad k \in \Lambda_j.$$

Prior (2):

$$p(\gamma_0^a = 1) = p(\gamma_0^b = 1) = 0.5, \quad p(\gamma_{jkr}^a = 1) = 0.5\phi^j, \quad p(\gamma_{jkr}^b = 1) = 0.5, \quad r = 1, 2, 3, j = 0, \dots, J, \quad k \in \Lambda_j.$$

Prior (3):

$$p(\gamma_0^a = 1) = p(\gamma_0^b = 1) = 0.5, \quad p(\gamma_{jkr}^a = 1) = 0.5\phi^{8j}, \quad p(\gamma_{jkr}^b = 1) = 0.5, \quad r = 1, 2, 3, j = 0, \dots, J, \quad k \in \Lambda_j.$$

I considered $\phi = 1, 0.9, 0.8, 0.7$. Note that when $\phi = 1$, Priors (1)–(3) all become indifference priors. I applied Prior (1) to Case I, and applied Priors (2) and (3) to Case II. Prior (1) puts smaller weights on the higher level wavelet coefficients of both surfaces A and B so that they have larger prior probability to be zero, while Priors (2) and (3) only do this for surface A but assign neutral probabilities to the wavelet coefficients of surface B .

For each of the covariate surfaces (3.5)–(3.7) and for both Cases I and II, I repeated the

simulations $L = 50$ times. For the l -th replication with $l = 1, \dots, L$, Markov chains with length 5000 were generated from the posterior distribution (3.2), and the first 2500 served as burn-ins. Gelman-Rubin's factors (see Gelman *et al.*, 2003) for all chains were below 1.1, suggesting that all chains converged well. The estimates \hat{A}^l and \hat{B}^l of A and B based on the l -th replication were obtained through averaging the last 2500 posterior samples.

To assess performance, I borrowed an idea from Fan *et al.* (2010) to calculate the squared bias, variance and mean square errors of the estimates. To state our method, I let \hat{A}_i^l , \hat{B}_i^l , A_i and B_i be the values of \hat{A}^l , \hat{B}^l , A and B at pixel $\tilde{\mathbf{s}}_i$ with $\{\tilde{\mathbf{s}}_i\} = \{(s_1/100, s_2/100) | s_1, s_2 = 0, 1, \dots, 99\}$ being the 100×100 uniform grid of pixels over $[0, 1) \times [0, 1)$. Thus, there are $N = 10^4$ pixels being evaluated. Note that $\{\tilde{\mathbf{s}}_i\}$ have been chosen to be different from the locations where data were drawn for the purposes of assessing the performance of the estimates at new locations. I define the average squared bias to be

$$Bias_A^2 = \frac{1}{N} \sum_{i=1}^N \left(\sum_{l=1}^L \frac{\hat{A}_i^l - A_i}{L} \right)^2,$$

$$Bias_B^2 = \frac{1}{N} \sum_{i=1}^N \left(\sum_{l=1}^L \frac{\hat{B}_i^l - B_i}{L} \right)^2,$$

and define the average variance to be

$$Var_A = \frac{1}{N} \sum_{i=1}^N \sum_{l=1}^L \left(\hat{A}_i^l - \frac{1}{L} \sum_{l=1}^L \hat{A}_i^l \right)^2 / L,$$

$$Var_B = \frac{1}{N} \sum_{i=1}^N \sum_{l=1}^L \left(\hat{B}_i^l - \frac{1}{L} \sum_{l=1}^L \hat{B}_i^l \right)^2 / L.$$

The average mean square errors for A , B are then defined to be $MSE_A = Bias_A^2 + Var_A$ and $MSE_B = Bias_B^2 + Var_B$. The average mean square error for the response is defined to

be $MSE_y = \sum_{i=1}^N \sum_{l=1}^L \left(\hat{A}_i^l + x_i \hat{B}_i^l - (A_i + x_i B_i) \right)^2 / (NL)$, where $x_i = x(\tilde{\mathbf{s}}_i)$.

I first assessed the performance of Model I with Algorithm I. Tables 2 and 3 summarize the average squared bias, variance and mean square error of the estimates by using both Algorithm I and LASSO. Since Priors (2) and (3) coincide with each other when $\phi = 1$, I only recorded the results corresponding to Prior (2) when $\phi = 1$. Several findings result from these tables. First, for Case I where both A and B are piecewise constant, the Bayesian estimates corresponding to all the covariate surfaces x_a , x_b and x_c have similar performance in terms of MSE_A , MSE_B and MSE_y . For estimating A , the Bayesian method results in smaller mean square errors than LASSO, while for estimating B , the Bayesian and LASSO methods result in comparable mean square errors. Second, for Case II where A is piecewise constant but B is smooth, the Bayesian estimates corresponding to x_c are slightly better than those corresponding to x_a and x_b in terms of MSE_A and MSE_B . Zhang *et al.* (2011) observed similar effects of the covariate surfaces on the LASSO estimates. I can also see that, for $\phi = 0.9, 0.8, 0.7$, Prior (3) results in smaller MSE_A than Prior (2). Compared with LASSO, the Bayesian approach corresponding to Prior (3) produces smaller MSE_A , but produces slightly larger MSE_B . Third, for both Priors (2) and (3), when ϕ decreases, the average variances of the posterior estimates of both A and B decrease.

To examine Model II with Algorithm II, I repeated the simulations 50 times and each time generated 5000 MCMC samples based on the posterior distribution (3.4). I then treated the first half as burn-ins. Convergence was monitored through Gelman-Rubin's factors. Tables 4 and 5 summarize the results of using Algorithm II. Comparing Tables 2 and 4, and Tables 3 and 5, two observations can be made: (1) for Case I in which B is piecewise constant, Model I and Model II result in comparable MSE_A and MSE_B , while Model I corresponds to slightly smaller MSE_y ; (2) for Case II in which B is smooth, Model II outperforms Model

Surface	Method	$Bias_A^2$	$Bias_B^2$	Var_A	Var_B	MSE_A	MSE_B	MSE_y
x_a	$\phi = 1$	0.0004	0.0600	0.0146	0.0010	0.0149	0.0610	0.0223
	$= 0.9$	0.0002	0.0600	0.0091	0.0007	0.0093	0.0607	0.0150
	$= 0.8$	0.0002	0.0601	0.0072	0.0006	0.0074	0.0606	0.0115
	$= 0.7$	0.0002	0.0601	0.0054	0.0005	0.0056	0.0605	0.0094
	LASSO	0.0389	0.0599	0.0038	0.0088	0.0427	0.0687	0.0864
x_b	$\phi = 1$	0.0002	0.0601	0.0132	0.0008	0.0134	0.0609	0.0209
	$= 0.9$	0.0001	0.0601	0.0086	0.0006	0.0087	0.0607	0.0140
	$= 0.8$	0.0001	0.0601	0.0064	0.0005	0.0065	0.0606	0.0108
	$= 0.7$	0.0001	0.0601	0.0051	0.0004	0.0052	0.0605	0.0089
	LASSO	0.0312	0.0595	0.0034	0.0061	0.0346	0.0656	0.0754
x_c	$\phi = 1$	0.0002	0.0603	0.0131	0.0010	0.0133	0.0613	0.0216
	$= 0.9$	0.0001	0.0602	0.0082	0.0007	0.0083	0.0610	0.0145
	$= 0.8$	0.0001	0.0602	0.0060	0.0006	0.0061	0.0608	0.0111
	$= 0.7$	0.0001	0.0602	0.0046	0.0005	0.0047	0.0607	0.0090
	LASSO	0.0341	0.0602	0.0038	0.0044	0.0379	0.0646	0.0731

Table 2: Average squared bias, variance and mean square error related to Case I when Bayesian and LASSO approaches have been applied. For the Bayesian approach, Model I with Algorithm I has been implemented and Prior (1) has been imposed on the vector of Bernoulli variables γ .

I in terms of MSE_A , MSE_B and MSE_y .

3.2.2 Detecting Where the Slopes Are Nonzero

Our modeling approach allows for nonstationarity in the B surface, and in particular, it is possible that the relationship between the y and x surfaces vary over space. Therefore, it is of interest to detect the regions where the response has a strong relationship with the covariate. This is equivalent to detecting the locations or pixels on the image where the slopes deviate from zero. To accomplish this, I construct a $100(1 - \alpha)\%$ credible interval for $B(\mathbf{s})$ at each pixel \mathbf{s} . If the credible interval at \mathbf{s} excludes zero, then that gives evidence that $B(\mathbf{s})$ deviates from zero. Note that the upper and lower bounds of all the credible intervals form two-dimensional surfaces which together I call an uncertainty band. Unlike one-dimensional wavelet regression problem where the graphical demonstration of uncertainty bands is feasible (see, e.g., Chipman *et al.* 1997), it is difficult to effectively plot the two-dimensional

Surface	Method		$Bias_A^2$	$Bias_B^2$	Var_A	Var_B	MSE_A	MSE_B	MSE_y
x_a	Prior (2)	$\phi = 1$	0.8216	0.3064	0.6070	0.0768	1.4286	0.3832	0.9386
		$= 0.9$	0.3351	0.2641	0.3963	0.0591	0.7314	0.3232	0.9582
		$= 0.8$	0.1473	0.2479	0.1629	0.0400	0.3103	0.2879	0.9732
		$= 0.7$	0.0828	0.2429	0.1196	0.0370	0.2023	0.2799	0.9872
	Prior (3)	$\phi = 0.9$	0.0237	0.2403	0.0336	0.0298	0.0573	0.2702	1.0124
		$= 0.8$	0.0144	0.2405	0.0139	0.0285	0.0283	0.2691	1.0281
		$= 0.7$	0.0137	0.2411	0.0121	0.0284	0.0259	0.2695	1.0305
	LASSO		0.1298	0.1987	0.0222	0.0355	0.1520	0.2342	0.8599
x_b	Prior (2)	$\phi = 1$	0.0952	0.2069	0.1149	0.0318	0.2101	0.2387	0.9641
		$= 0.9$	0.0691	0.2037	0.0817	0.0304	0.1507	0.2341	0.9696
		$= 0.8$	0.0535	0.2034	0.0596	0.0294	0.1131	0.2327	0.9789
		$= 0.7$	0.0440	0.2033	0.0440	0.0286	0.0880	0.2319	0.9879
	Prior (3)	$\phi = 0.9$	0.0313	0.2060	0.0166	0.0269	0.0479	0.2329	1.0134
		$= 0.8$	0.0288	0.2065	0.0076	0.0267	0.0364	0.2332	1.0270
		$= 0.7$	0.0285	0.2068	0.0061	0.0268	0.0346	0.2336	1.0302
	LASSO		0.1212	0.1885	0.0067	0.0237	0.1279	0.2122	0.9374
x_c	Prior (2)	$\phi = 1$	0.0427	0.1994	0.0707	0.0286	0.1134	0.2280	1.0131
		$= 0.9$	0.0248	0.1986	0.0486	0.0271	0.0734	0.2257	1.0271
		$= 0.8$	0.0149	0.1990	0.0345	0.0265	0.0494	0.2255	1.0428
		$= 0.7$	0.0090	0.1993	0.0243	0.0261	0.0333	0.2253	1.0569
	Prior (3)	$\phi = 0.9$	0.0025	0.1997	0.0088	0.0263	0.0113	0.2260	1.0893
		$= 0.8$	0.0015	0.1999	0.0032	0.0260	0.0047	0.2259	1.1052
		$= 0.7$	0.0014	0.1999	0.0029	0.0261	0.0043	0.2259	1.1075
	LASSO		0.0549	0.1941	0.0039	0.0199	0.0588	0.2139	1.2012

Table 3: Average squared bias, variance and mean square error related to Case II when Bayesian and LASSO approaches have been applied. For the Bayesian approach, Model I with Algorithm I has been implemented and Priors (2) and (3) have been imposed on the vector of Bernoulli variables γ .

Surface	Method	$Bias_A^2$	$Bias_B^2$	Var_A	Var_B	MSE_A	MSE_B	MSE_y
x_a	$\phi = 1$	0.0003	0.0604	0.0132	0.0067	0.0135	0.0670	0.0863
	$= 0.9$	0.0002	0.0602	0.0103	0.0046	0.0105	0.0648	0.0639
	$= 0.8$	0.0002	0.0601	0.0079	0.0033	0.0080	0.0633	0.0466
	$= 0.7$	0.0001	0.0600	0.0061	0.0023	0.0062	0.0623	0.0333
x_b	$\phi = 1$	0.0006	0.0601	0.0188	0.0075	0.0194	0.0676	0.0909
	$= 0.9$	0.0004	0.0601	0.0137	0.0054	0.0141	0.0655	0.0669
	$= 0.8$	0.0003	0.0601	0.0100	0.0039	0.0103	0.0640	0.0486
	$= 0.7$	0.0002	0.0601	0.0073	0.0028	0.0075	0.0628	0.0346
x_c	$\phi = 1$	0.0004	0.0598	0.0217	0.0082	0.0221	0.0680	0.0933
	$= 0.9$	0.0003	0.0598	0.0152	0.0059	0.0155	0.0657	0.0680
	$= 0.8$	0.0002	0.0598	0.0106	0.0042	0.0108	0.0640	0.0488
	$= 0.7$	0.0001	0.0598	0.0075	0.0030	0.0076	0.0628	0.0345

Table 4: Average squared bias, variance and mean square error related to Case I when a Bayesian approach has been applied. Model II with Algorithm II has been implemented and Prior (1) has been imposed on the vector of Bernoulli variables γ .

Surface	Method		$Bias_A^2$	$Bias_B^2$	Var_A	Var_B	MSE_A	MSE_B	MSE_y
x_a	Prior (2)	$\phi = 1$	0.0510	0.2212	0.0222	0.0232	0.0733	0.2444	0.9304
		$= 0.9$	0.0451	0.2187	0.0206	0.0234	0.0657	0.2421	0.9356
		$= 0.8$	0.0382	0.2168	0.0187	0.0236	0.0569	0.2404	0.9398
		$= 0.7$	0.0342	0.2151	0.0181	0.0238	0.0523	0.2389	0.9436
	Prior (3)	$\phi = 0.9$	0.0221	0.2109	0.0168	0.0251	0.0389	0.2360	0.9500
		$= 0.8$	0.0199	0.2102	0.0161	0.0253	0.0360	0.2355	0.9526
		$= 0.7$	0.0203	0.2103	0.0163	0.0253	0.0366	0.2356	0.9531
x_b	Prior (2)	$\phi = 1$	0.0200	0.1846	0.0112	0.0218	0.0312	0.2064	0.9020
		$= 0.9$	0.0162	0.1852	0.0088	0.0213	0.0250	0.2065	0.9035
		$= 0.8$	0.0143	0.1848	0.0072	0.0214	0.0215	0.2062	0.9080
		$= 0.7$	0.0134	0.1845	0.0060	0.0215	0.0194	0.2060	0.9115
	Prior (3)	$\phi = 0.9$	0.0133	0.1828	0.0048	0.0222	0.0181	0.2050	0.9232
		$= 0.8$	0.0134	0.1828	0.0046	0.0223	0.0180	0.2051	0.9270
		$= 0.7$	0.0133	0.1829	0.0046	0.0223	0.0179	0.2052	0.9280
x_c	Prior (2)	$\phi = 1$	0.0093	0.1811	0.0114	0.0203	0.0207	0.2014	0.9373
		$= 0.9$	0.0052	0.1808	0.0085	0.0204	0.0138	0.2013	0.9551
		$= 0.8$	0.0029	0.1808	0.0065	0.0204	0.0095	0.2012	0.9663
		$= 0.7$	0.0015	0.1807	0.0052	0.0203	0.0067	0.2010	0.9760
	Prior (3)	$\phi = 0.9$	0.0002	0.1810	0.0033	0.0203	0.0035	0.2013	0.9920
		$= 0.8$	0.0001	0.1810	0.0031	0.0203	0.0032	0.2013	1.0004
		$= 0.7$	0.0001	0.1810	0.0031	0.0203	0.0032	0.2013	1.0010

Table 5: Average squared bias, variance and mean square error related to Case II when a Bayesian approach has been applied. Model II with Algorithm II has been implemented and Priors (2) and (3) have been imposed on the vector of Bernoulli variables γ .

uncertainty bands. In this section, I use an alternative method to address this difficulty. Before proceeding further, I perform some useful calculations.

I denote $B_i = B(\mathbf{s}_i)$, and let $\mathbf{b}^{(1)}, \dots, \mathbf{b}^{(T)}$ be T posterior samples of \mathbf{b} , where \mathbf{b} denotes the vector of wavelet coefficients of the surface B . Let $B_i^{(t)} = W(\mathbf{s}_i)\mathbf{b}^{(t)}$ for $t = 1, \dots, T$. The Bayesian estimate of B_i is

$$\hat{B}_i = \sum_{t=1}^T B_i^{(t)} / T = W(\mathbf{s}_i)\hat{\mathbf{b}},$$

where $\hat{\mathbf{b}} = \sum_{t=1}^T \mathbf{b}^{(t)} / T$. The posterior variance of $B_i^{(t)}$, $t = 1, \dots, T$, is

$$\hat{\sigma}_i^2 = W(\mathbf{s}_i)\hat{\Sigma}W(\mathbf{s}_i)',$$

where $\hat{\Sigma} = \sum_{t=1}^T (\mathbf{b}^{(t)} - \hat{\mathbf{b}}) (\mathbf{b}^{(t)} - \hat{\mathbf{b}})' / (T - 1)$ is an $m \times m$ matrix. I call $\hat{\sigma}_i$ the posterior standard deviation (PSD) of B at pixel \mathbf{s}_i .

I find the pixels at which the slopes deviate from zero, and also classify the pixels according to the magnitudes and signs of the slopes. For this purpose, I construct a choropleth map to indicate $\hat{B}_i \geq \Delta$, $0 \leq \hat{B}_i < \Delta$, $-\Delta < \hat{B}_i < 0$ and $\hat{B}_i \leq -\Delta$, with $\Delta > 0$ a suitably selected threshold.

In the simulated and real data examples discussed later, a majority of the posterior distributions $p(B_i|\mathbf{y}, X)$ of B_i are unimodal and roughly symmetric. Therefore, it is convenient to approximate $p(B_i|\mathbf{y}, X)$ by a normal distribution with center and scale being \hat{B}_i and $\hat{\sigma}_i$. Using an analogy to the concept of frequentist p -value, if $|\hat{B}_i/\hat{\sigma}_i| > 1.96$, then I believe with strong evidence that $B_i \neq 0$ and represent this situation by $p < 0.05$; if $1.64 \leq |\hat{B}_i/\hat{\sigma}_i| \leq 1.96$, then I believe with moderate evidence that $B_i \neq 0$ and represent this situation by $0.05 \leq p \leq 0.1$; otherwise, I believe that B_i might be close to zero and represent this situation by $p > 0.1$. Note that this is analogous to the interpretation of a frequentist p -value. In a choropleth map, I designate the various possibilities for p by different using different line-patterns.

In a simulation study, the A surface was defined as in Section 4.2.1 and the B surface was defined by Case II in Section 3.1, i.e., $B(s_1, s_2) = 4 \sin(2\pi s_1) \cos(2\pi s_2)$, $0 \leq s_1, s_2 \leq 1$. Note that B is smooth with zero values at some pixels. Algorithm I under Model I was implemented, and I set $\Delta = 2$ which is half of the maximum value of $|B|$.

In addition, I chose $J = 4$ and generated data from model (3.1) with $\sigma = 1$. Thus, $n = 4096$ and $m = 2048$. I chose the hyperparameters $\mu = \nu = 6$ and prior (3) defined in Section 4.2.1 was used for the Bernoulli variable γ for each of the cases $\phi = 1, 0.9, 0.8, 0.7$. Markov chains of length 5000 were simulated with the first half burn-ins, and I used the second half for calculations. Convergence was assessed through Gelman-Rubin's factors.

Figure 4 displays the images of \hat{B} and the PSD of B corresponding to $\phi = 1, 0.9, 0.8, 0.7$ when using x_c defined in Section 4.2.1 as the covariate surface. I observe that all the \hat{B} images graphically resemble the true B , and the PSD of B for $\phi = 1$ appear to be greater than those for $\phi = 0.9, 0.8, 0.7$. I also observe that when ϕ decreases, the \hat{B} images become slightly sparser in the sense that larger square regions appear on the images. This is because when the Bernoulli probabilities associated with higher level wavelet coefficients become smaller, the finer details will be dropped and the basis supports with smaller sizes will merge into larger square regions.

As displayed in Figure 4, there are three peaks (indicated by red) and three valleys (indicated by blue) regularly arranged on the true B image, and the values of the true B at the pixels around the peaks and valleys deviate from zero, while they are close to zero elsewhere. Figure 5 displays the choropleth map for \hat{B} corresponding to various ϕ values. I observe that the locations where the B values deviate from zero are correctly detected and changing ϕ makes little change in the detection results.

3.2.3 Applications to Gypsy Moth Defoliation Data

I next use the proposed Bayesian approach to analyze the gypsy moth defoliation data introduced in Section 1. Recall that the defoliation data contains images of defoliation rates (response) and elevations (covariate). After processing (see Zhang *et al.*, 2011), the images consist of 64×64 evenly spaced pixels \mathbf{s}_i s, and therefore, $n = 4096$. The response $y(\mathbf{s})$ and the covariate $x_1(\mathbf{s})$ represent the centered-and-scaled defoliation rate and scaled elevation measured at pixel \mathbf{s} respectively (as displayed in Figure 1). I used the centered-and-scaled x_1 as the covariate surface x , i.e., $x(\mathbf{s}) = (x_1(\mathbf{s}) - \overline{vec(x_1)})/\text{std}(vec(x_1))$, where $vec(x_1)$ denotes the vector of x_1 values at the 4096 pixels, and $\overline{vec(x_1)}$ and $\text{std}(vec(x_1))$ are the

sample mean and standard deviation of $vec(x_1)$. $J = 4$ was used, and thus, $m = 2048$ wavelet coefficients are involved in our model.

I fixed $\mu = \nu = 6$ and fit Model I. Prior (1) was placed on γ with the Bernoulli probabilities corresponding to resolution levels 0 to 4 being $0.5, 0.5\phi, 0.5\phi^2, 0.5\phi^3$ and $0.5\phi^4$ respectively. I somewhat arbitrarily chose $\phi = 0.9$ to produce some degree of flatness in the estimates. A Markov chain of length 20,000 was simulated from the posterior distribution $p(\beta, \gamma, \sigma, \tau | \mathbf{y}, X)$ specified by (3.2) using Algorithm I, and the first half was treated as burn-ins. The initial point $\beta^{(0)}$ for the β chain was generated from $N(\hat{\beta}, 10^{-4}I_m)$, where $\hat{\beta}$ was chosen as the least squares estimate of β . It took about 2.25 hours to draw 10,000 posterior samples. Convergence was assessed by applying Gelman-Rubin factors to 5 parallel Markov chains. I also applied the method introduced in Section 4.2.2 to classify the pixels.

Figure 6 displays the estimated intercept \hat{A} , the estimated slope \hat{B} , the fitted defoliation rate \hat{y} and the PSD of the slope B . In particular, the images of \hat{A} , \hat{B} and the PSD were constructed over a 100×100 lattice set of locations in $[0, 1) \times [0, 1)$ to display the posterior samples at new locations; while the \hat{y} image was constructed over the 64×64 lattice set of locations in $[0, 1) \times [0, 1)$ where the data were drawn allowing us to compare \hat{y} with y at the observed locations. I observe that \hat{B} is positive at most of the pixels, which shows an overall positive relationship between the defoliation rate and elevation. Furthermore, \hat{B} is slightly smaller at the locations where the elevation is small. I also observe that in the regions where the elevation changes quickly, the PSD of the slope deviates considerably from zero. Finally, the image \hat{y} appears to resemble the observed defoliation rate image y . Our findings on \hat{B} and \hat{y} are similar to those made by Zhang *et al.* (2011) who used LASSO algorithm to perform the computations, but again, I am also able to characterize the uncertainty in the relationships.

Figure 7 displays the choropleth map of the slope in which I chose $\Delta = 0.8$ (about $1/3$

the maximum of $|\hat{B}|$). I observe that in the upper-left region, the relationship between defoliation rate and elevation is strong and positive, while in the nearly central region, the relationship between defoliation rate and elevation is not strong. I also observe that, at a small number of locations, $p < 0.05$ and $B \leq -0.8$ which shows that the relationship there is strong and negative.

3.3 Discussion

Gelfand *et al.* (2003) proposed a Bayesian framework for spatial concurrent linear models with model coefficients being spatially varying processes. They imposed Gaussian process priors on the coefficient processes which essentially assume stationarity. To remove the assumption of stationarity from the model, Zhang *et al.* (2011) applied a wavelet approach to transform the spatial concurrent linear model into a linear model with design matrix induced by a wavelet structure, and they implemented LASSO to handle the estimation problem. However, it is difficult to conduct inferences using their method. To address this, I have developed a Bayesian variable selection approach based on the model proposed by Zhang *et al.* (2011). Specifically, I applied a Bayesian model similar to one proposed by George and McCulloch (1993), in which I introduced a vector γ of Bernoulli variables for the model coefficients so that the selection and estimation of the nonzero coefficients can be simultaneously achieved. The proposed approach is highly flexible and computationally efficient, and should be useful in many practical situations where the data display complex nonstationary patterns. In addition, I use a Gibbs sampler for posterior sampling that involves no complicated matrix computation. Hence, this is efficient for handling relatively large datasets. Furthermore, as demonstrated in simulated and real data analysis, our approach is effective in detecting the spatial locations where the response has a relationship

with a covariate, and provides statistical evidence for such detections.

I have placed Bernoulli priors on γ . Other priors such as Markov chain priors can also be applied by invoking a tree structure (see Romberg *et al.*, 2001). The support of any Haar wavelet basis function, which I call a parent, is divided into four equal adjacent pieces at the same level, which I call children, with each piece being the support of a Haar wavelet basis function. Since any basis support corresponds to a 0-1 variable γ_j , I also call $\gamma_{j'}$ the parent of γ_j if their corresponding basis supports have such parent-children relationship. Following Romberg *et al.* (2001), a Markov chain prior is defined to be

$$p(\gamma_j|\gamma_{-j}) = p(\gamma_j|\gamma_{j'}), \quad (3.8)$$

where $\gamma_{-j} = \{\gamma_i | i \neq j\}$, and $\gamma_{j'}$ is the parent of γ_j . The equation (3.8) means that the distributional properties of a child only depends on its parent. Let the transition probability be $p(\gamma_j|\gamma_{j'}) = p_{\gamma_{j'}, \gamma_j}$. I have numerically examined Markov chain priors with $p_{0,0} = 0.9$, $p_{0,1} = 0.1$, $p_{1,0} = 0.1$, $p_{1,1} = 0.9$, and found that they did not perform as well as Bernoulli priors and LASSO when estimating a piecewise constant surface. The reason might be that a piecewise constant surface has too much local flatness, and hence, even if a parent corresponds to a nonzero wavelet coefficient, its four children may still correspond to zero wavelet coefficients, which makes the connection between the parent and children weak. Under such circumstances, Bernoulli priors which assume independence among the basis functions may be better choices.

Two future extensions of the current work might be also worth mentioning. First, Dunson (2009) proposed a nonparametric Bayesian approach to model the basis coefficients in a longitudinal model. In his method, the prior distribution of the basis coefficients is nonparametric; in particular, they used a Dirichlet process prior, which provides a great deal

of flexibility. Dunson (2009) found that the nonparametric prior distribution works well for modeling the model coefficients, and it seems reasonable to extend that work to our model.

Second, in our model, the coefficients are sparse, and so even if the the number of parameters is large, the estimation results are still satisfactory. Although a sparse coefficient vector is common in the regression models associated with wavelets, it is still interesting to fit a model with non-sparse coefficients and examine the results. One article about the identification of the sparseness pattern of the model coefficients is given by Meinshausen and Yu (2009) who examined the impact of sparseness on LASSO estimates. There seems to be little literature handling this problem under a Bayesian framework, and so I intend to explore this further in the future.

3.4 Appendix: Sampler Derivations for Algorithm I.

Following Godsill and Rayner (1998), I define the blocks $z_j = (\gamma_j, \beta_j)$ for $j = 1, \dots, m$. Thus, each z_j can be viewed as a two-dimensional parameter. The idea is to update $(z_1, \dots, z_m, \sigma^2, \tau^2)$ using a standard Gibbs sampler. For any $j = 1, \dots, m$, I let $\tilde{\gamma}_{-j} = (\gamma_1, \dots, \gamma_{j-1}, \gamma_{j+1}, \dots, \gamma_m)$ and $\tilde{\beta}_{-j} = (\beta_1, \dots, \beta_{j-1}, \beta_{j+1}, \dots, \beta_m)'$.

Sampling γ and β . It follows from (3.2) that

$$\begin{aligned}
& p(\beta_j, \gamma_j = 1 | z_1, \dots, z_{j-1}, z_{j+1}, \dots, z_m, \sigma, \tau, \mathbf{y}, X) \\
& \propto \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp \left(-\frac{\|\mathbf{y} - X\beta\|^2}{2\sigma^2} \right) \left(\frac{1}{\sqrt{2\pi}\tau} \right) \exp \left(-\frac{\beta_j^2}{2\tau^2} \right) P(\gamma_j = 1 | \tilde{\gamma}_{-j}) \\
& = \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp \left(-\frac{\|\mathbf{y} - X_j\beta_j - X_{-j}\tilde{\beta}_{-j}\|^2}{2\sigma^2} \right) \left(\frac{1}{\sqrt{2\pi}\tau} \right) \exp \left(-\frac{\beta_j^2}{2\tau^2} \right) P(\gamma_j = 1 | \tilde{\gamma}_{-j}) \\
& = \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp \left(-\frac{\beta_j^2 v_j^2 - 2(\mathbf{y} - X_{-j}\tilde{\beta}_{-j})' X_j \beta_j}{2\sigma^2} \right) p(\gamma_j = 1 | \tilde{\gamma}_{-j}) \\
& \quad \cdot \left(\frac{1}{\sqrt{2\pi}\tau} \right) \exp \left(-\frac{\|\mathbf{y} - X_{-j}\tilde{\beta}_{-j}\|^2}{2\sigma^2} \right), \tag{3.9}
\end{aligned}$$

where $v_j^2 = X_j' X_j + \sigma^2/\tau^2$. Similarly, I have

$$\begin{aligned}
& p(\beta_j, \gamma_j = 0 | z_1, \dots, z_{j-1}, z_{j+1}, \dots, z_m, \sigma, \tau, \mathbf{y}, X) \\
& \propto \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \delta_0(\beta_j) p(\gamma_j = 0 | \tilde{\beta}_{-j}) \exp \left(-\frac{\|\mathbf{y} - X_{-j}\tilde{\beta}_{-j}\|^2}{2\sigma^2} \right). \tag{3.10}
\end{aligned}$$

Integrating out β_j in (3.9) and (3.10), I have

$$\begin{aligned}
& p(\gamma_j = 1 | z_1, \dots, z_{j-1}, z_{j+1}, \dots, z_m, \sigma, \tau, \mathbf{y}, X) \\
& \propto \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp \left(\frac{\tilde{u}_j^2}{2\sigma^2 v_j^2} \right) \left(\frac{\sigma}{\tau v_j} \right) p(\gamma_j = 1 | \tilde{\gamma}_{-j}) \exp \left(-\frac{\|\mathbf{y} - X_{-j}\tilde{\beta}_{-j}\|^2}{2\sigma^2} \right),
\end{aligned}$$

and

$$\begin{aligned}
& p(\gamma_j = 0 | z_1, \dots, z_{j-1}, z_{j+1}, \dots, z_m, \sigma, \tau, \mathbf{y}, X) \\
& \propto \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n p(\gamma_j = 0 | \tilde{\gamma}_{-j}) \exp \left(-\frac{\|\mathbf{y} - X_{-j}\tilde{\beta}_{-j}\|^2}{2\sigma^2} \right),
\end{aligned}$$

where $\tilde{u}_j = (\mathbf{y} - X_{-j}\tilde{\beta}_{-j})'X_j$. Therefore,

$$p(\gamma_j = 1 | z_1, \dots, z_{j-1}, z_{j+1}, \dots, z_m, \sigma, \tau, \mathbf{y}, X) = 1/(1 + \tilde{\rho}_j),$$

where $\tilde{\rho}_j = \frac{p(\gamma_j=0|\tilde{\gamma}_{-j})}{p(\gamma_j=1|\tilde{\gamma}_{-j})} \cdot \frac{\tau v_j}{\sigma} \exp\left(-\frac{\tilde{u}_j^2}{2\sigma^2 v_j^2}\right)$. It follows from (3.9) and (3.10) that

$$p(\beta_j = 0 | \gamma_j = 0, z_1, \dots, z_{j-1}, z_{j+1}, \dots, z_m, \sigma, \tau, \mathbf{y}, X) = 1$$

and

$$\beta_j | \gamma_j = 1, z_1, \dots, z_{j-1}, z_{j+1}, \dots, z_m, \sigma, \tau, \mathbf{y}, X \sim N\left(\frac{\tilde{u}_j}{v_j^2}, \frac{\sigma^2}{v_j^2}\right).$$

Using the above procedure, all the blocks z_j s can be updated.

Sampling σ and τ . Step (B) in Algorithm I follows directly from (3.2).

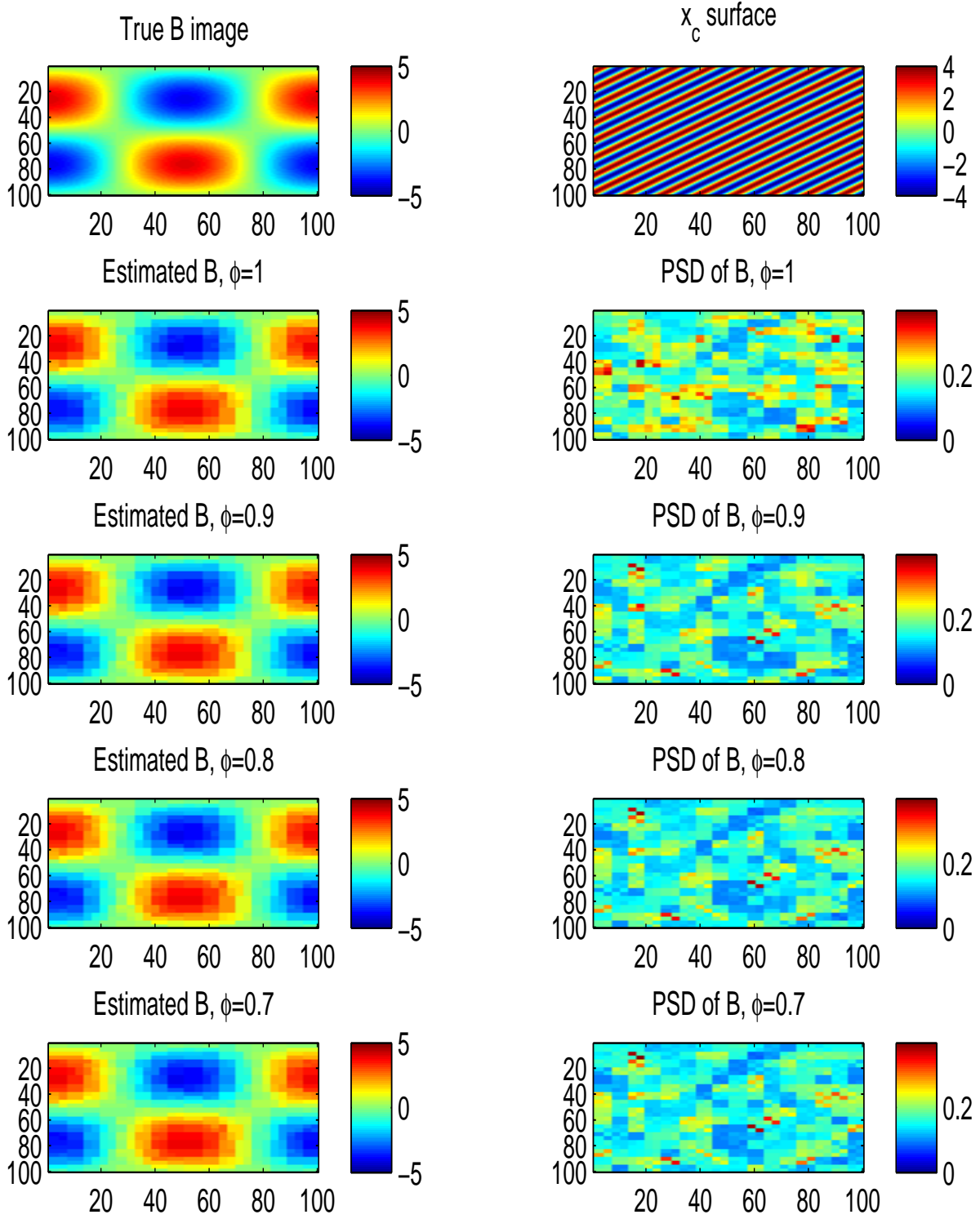


Figure 4: Section 4.2.2. Images of estimated B and the PSD of B for $\phi = 1, 0.9, 0.8, 0.7$. Covariate surface x_c and Prior (3) were used. Images of the true B and x_c are also shown.

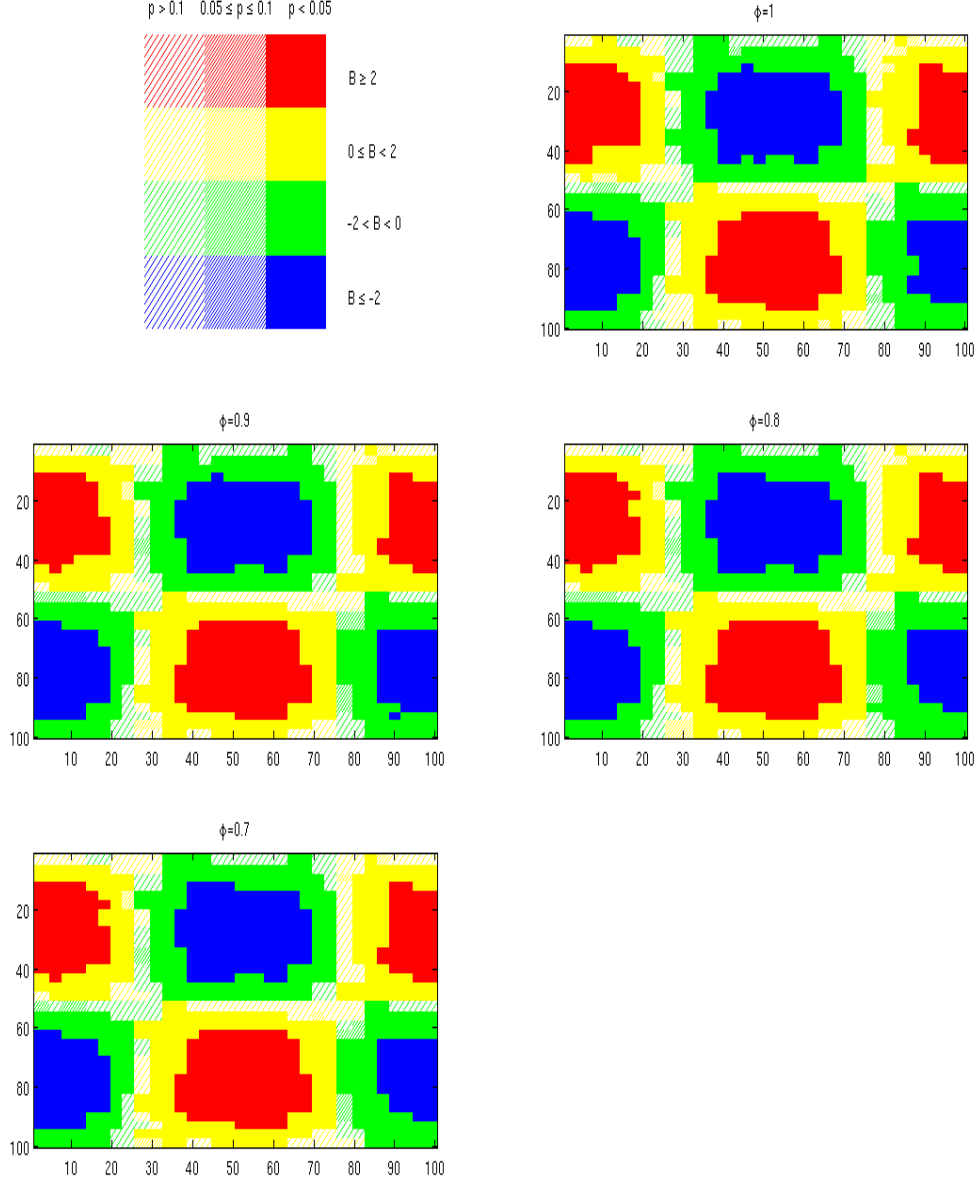


Figure 5: Section 4.2.2. Choropleth map of \hat{B} for $\phi = 1, 0.9, 0.8, 0.7$. Covariate surface x_c and Prior (3) were used. Red indicates $\hat{B}_i \geq 2$; Yellow: $0 \leq \hat{B}_i < 2$; Green: $-2 < \hat{B}_i < 0$; Blue: $\hat{B}_i \leq -2$. Filled boxes: $p < 0.05$; boxes with dense lines: $0.05 \leq p \leq 0.1$; boxes with thin lines: $p > 0.1$.

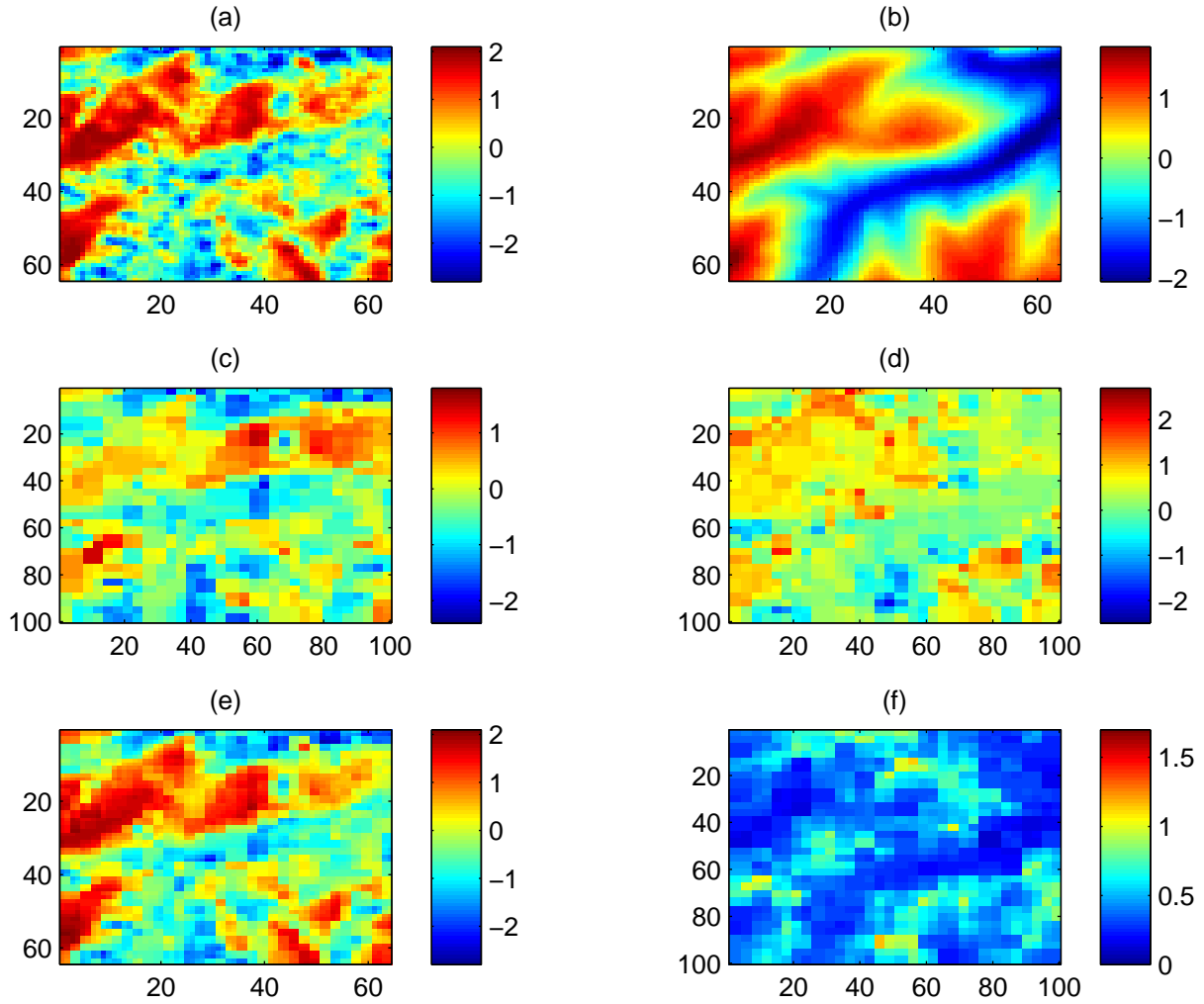


Figure 6: (a): Centered-and-scaled defoliation rate y ; (b): Centered-and-scaled elevation x ; (c): Estimated intercept \hat{A} ; (d): Estimated slope \hat{B} ; (e): Fitted value \hat{y} ; (f): PSD of the slope B . Prior (1) with $\phi = 0.9$ was used.

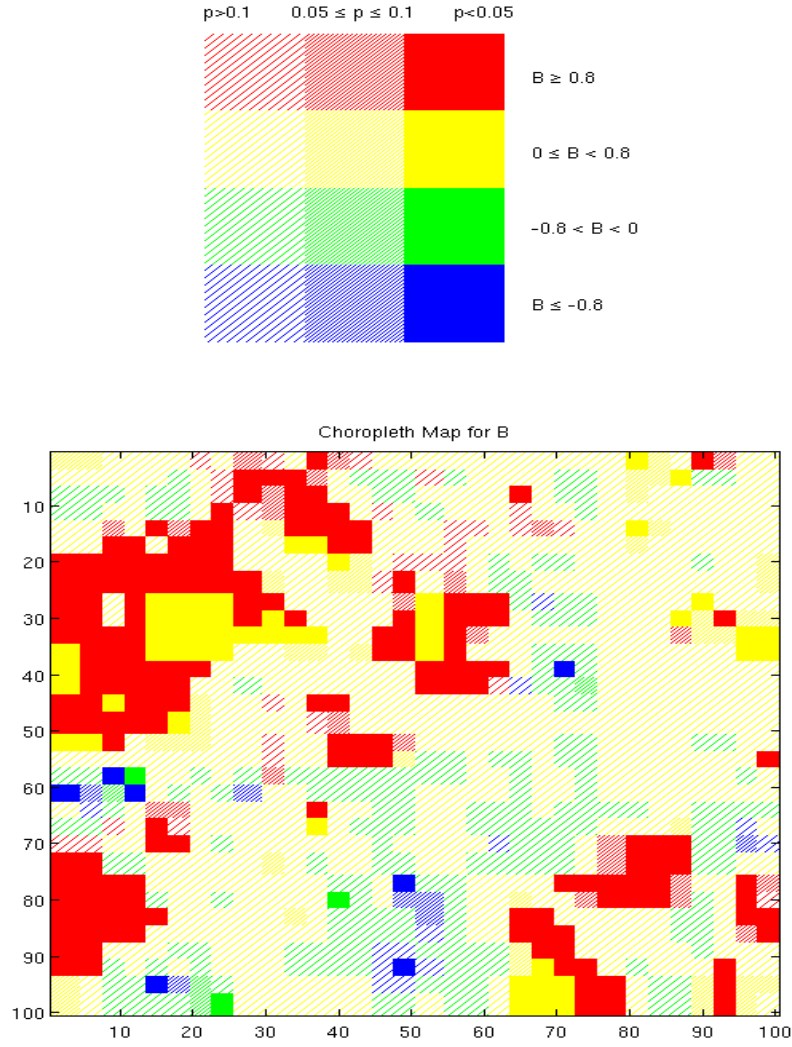


Figure 7: *Choropleth map for the slope surface B. Prior (1) with $\phi = 0.9$ was used. Red indicates $\hat{B}_i \geq 0.8$; Yellow: $0 \leq \hat{B}_i < 0.8$; Green: $-0.8 < \hat{B}_i < 0$; Blue: $\hat{B}_i \leq -0.8$. Filled boxes: $p < 0.05$; boxes with dense lines: $0.05 \leq p \leq 0.1$; boxes with thin lines: $p > 0.1$.*

Chapter 4

Conclusions And Future Work

The first part of my dissertation contains several theoretical results related to posterior linear model consistency when the model dimension p increases with sample size n . My results provide sufficient conditions under which the posterior probability of the true model converges in probability to one when p grows at certain rates with n . I also generalized these results to ultrahigh-dimensional situations and g -prior frameworks. In a g -prior framework, I proved that if the g -prior is predetermined and does not depend on any sample information, then when p grows slower than $n^{1/2}$, $p(\gamma^0|\text{data})$ dominates $p(\gamma|\text{data})$ for any $\gamma \neq \gamma^0$ (in probability); when p grows slower than $n^{1/4}$, $p(\gamma^0|\text{data}) \rightarrow 1$ (in probability). However, if p grows faster than the above two rates, to satisfy PMC, the g -prior needs to depend on the sample size n .

In the second part of my dissertation, I have applied a Bayesian approach to study several problems in spatial statistics. Specifically, based on Zhang's two-dimensional wavelet approach to transform spatial surfaces to sparse vectors of coefficients, I used a Bayesian variable selection approach to select and estimate the nonzero wavelet coefficients. Then I used the estimated coefficients to recover the spatial surfaces. Compared with the LASSO approach proposed by Zhang *et al.* (2011), our approach is more efficient and accurate in handling locally flat surfaces while performing as well as LASSO in handling smooth surfaces. In addition, our approach allows us to make inferences about the estimated coefficient surfaces. In particular, I applied the proposed Bayesian approach to analyze the defoliation

data of Dr. Townsend and detect the spatial locations, with given statistical evidence, where the relationship between the defoliation rates are elevations are strong and weak.

The proposed methodology, which is built upon a wavelet approach, is useful when the number of covariate surfaces is relatively small. When there are many covariate images, the proposed Bayesian framework will be computationally slow, an issue that I plan to explore in the future. Instead of using a wavelet approach, which produces a lot of parameters that need to be estimated, I will use a Bayesian regression spline method which requires a smaller number of parameters in the model (Smith and Kohn 1997). This framework can allow for more covariate images in the model and less computational burden, although it does require a stronger smoothness assumption on the unknown coefficient surfaces. For any covariate surface, the problem is to determine the spatial locations where there is a relationship between that covariate and the response. In the following text, I briefly illustrate this idea.

I still restrict attention to the unit square region $[0, 1) \times [0, 1)$. Denote $B_{j,l}(s, t) = B_j^1(s)B_l^2(t)$, $s, t \in [0, 1)$ and $j = 1, \dots, L_1$, $l = 1, \dots, L_2$ to be the tensor product spline basis functions. More precisely, both $\{B_j^1(s)\}$ and $\{B_l^2(t)\}$ represent one-dimensional spline bases, and the two-dimensional spline basis is formed by the tensor product of $\{B_j^1(s)\}$ and $\{B_l^2(t)\}$ (see He 1996). Any smooth function f on $[0, 1) \times [0, 1)$ will admit the following approximate expansion

$$f(s, t) \approx \sum_{j=1}^{L_1} \sum_{k=1}^{L_2} f_{j,k} B_{j,k}(s, t). \quad (4.1)$$

We still consider model (1.3), but use a slightly different notation for convenience. Suppose that the dataset contains response and covariate surfaces $y(s, t)$ and $x_k(s, t)$, $k = 1, \dots, K$,

which satisfy the following concurrent linear model

$$y(s, t) = \sum_{k=1}^K A_k(s, t)x_k(s, t) + \epsilon(s, t), \quad s, t \in [0, 1].$$

Write

$$A_k(s, t) \approx \sum_{j=1}^{L_1} \sum_{l=1}^{L_2} \theta_{k,j,l} B_{j,l}(s, t), \quad k = 1, \dots, K,$$

and let $\mathbf{B}(s, t) = (B_{1,1}(s, t), \dots, B_{L_1, L_2}(s, t))$ and $\theta_k = (\theta_{k,1,1}, \dots, \theta_{k, L_1, L_2})$. Therefore, both $\mathbf{B}(s, t)$ and θ_k are row vectors of length $L = L_1 L_2$. So our model takes the approximate form

$$y(s, t) \approx W(s, t)\theta + \epsilon(s, t), \quad s, t \in [0, 1],$$

where $W(s, t) = (x_1(s, t)B(s, t), \dots, x_K(s, t)B(s, t))$ and $\theta = (\theta_1, \dots, \theta_K)'$. Consider the following Bayesian model

$$y(s, t) | \theta, \sigma^2 \sim N(W(s, t)\theta, \sigma^2), \quad s, t \in [0, 1],$$

$$\theta_k \sim N(\mathbf{0}, \tau_k^2 M_k),$$

$$\sigma^2 \sim \text{Inverse Gamma}(a, b/2),$$

$$\tau_k^2 \sim \text{Inverse Gamma}(a_k, b_k/2), \quad k = 1, \dots, K.$$

Here, the selection of the matrix M_k is flexible. An idea is $M_k = P_1' P_1 + P_2' P_2$ for all k s, where P_1 and P_2 correspond to row and column penalties (Marx and Eilers, 2005). The

posterior distribution of model (4.2) is

$$\begin{aligned}
 p(\theta, \sigma^2, \tau_1^2, \dots, \tau_K^2 | \mathbf{y}, W) &\propto \sigma^{-N} \exp\left(-\frac{\|\mathbf{y} - W\theta\|^2}{2\sigma^2}\right) \\
 &\cdot \prod_{k=1}^K \left(\frac{1}{\sqrt{2\pi\tau_k}}\right)^L \det(M_k)^{-1/2} \exp\left(-\frac{\theta'_k M_k \theta_k}{2\tau_k^2}\right) \\
 &\cdot p(\sigma^2) \cdot \prod_{k=1}^K p(\tau_k^2). \tag{4.2}
 \end{aligned}$$

The MCMC algorithm based on posterior distribution (4.2) is derived as follows. Let W be the matrix with rows $W(s, t)$.

(i). $\theta | \sigma^2, \tau_1^2, \dots, \tau_K^2 \sim N((W'W + M)^{-1}W'\mathbf{y}, \sigma^2(W'W + M)^{-1})$, where

$$M = \begin{pmatrix} M_1\sigma^2/\tau_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & M_K\sigma^2/\tau_K^2 \end{pmatrix}.$$

(ii). $\sigma^2 | \theta, \tau_1^2, \dots, \tau_K^2 \sim \text{Inverse Gamma}\left(a + N/2, \frac{b + \|\mathbf{y} - W\theta\|^2}{2}\right)$.

(iii). $\tau_k^2 | \sigma^2, \tau_\kappa^2, \kappa \neq k \sim \text{Inverse Gamma}\left(a_k + L/2, \frac{b_k + \theta'_k M_k \theta_k}{2}\right)$.

Using the posterior samples drawn from the above algorithm (i)–(iii), we can get posterior samples for $B_k(s, t)$, namely, $B_k^{(1)}(s, t), \dots, B_k^{(T)}(s, t)$, for any $s, t \in [0, 1)$.

To detect the spatial locations (s, t) where $B_k(s, t) \neq 0$, I will apply two methods. In the first one, I will apply the approach introduced in Section 4.2.2 to construct a credible interval for $B_k(s, t)$, using a given level, and infer whether $B_k(s, t) \neq 0$. In the second, I will find the quantities

$$p_k(s, t) = \min \{p(B_k(s, t) > 0 | \mathbf{y}, W), p(B_k(s, t) < 0 | \mathbf{y}, W)\},$$

In practice, we use the following way to approximate $p(B_k(s, t) > 0|\mathbf{y}, W)$ and $p(B_k(s, t) < 0|\mathbf{y}, W)$.

$$p(B_k(s, t) > 0|\mathbf{y}, W) \approx \sum_{t=1}^T I(B_k^{(t)}(s, t) > 0)/T,$$

$$p(B_k(s, t) < 0|\mathbf{y}, W) \approx \sum_{t=1}^T I(B_k^{(t)}(s, t) < 0)/T.$$

Then I will use an idea analogous to the local false discovery rate (Lfdr), see Efron (2007), by treating $p_k(s, t)$ as analogous to a frequentist p -value. I also intend to find an effective way to control the false nondiscovery rate (FNR). Comparison between these two methods will be conducted by examining the powers to detect the locations where a true relationship between the response and the corresponding covariate exists.

Several additional extensions of my current work are worth mentioning. In the following text, I will introduce these extensions in details.

First, I plan to generalize my both theoretical and applied results to generalized linear models. Let data (y, x) be generated from the following exponential family

$$f(\mathbf{y}, X; \beta) = \exp \left(\mathbf{y}^T X \beta - \sum_{i=1}^n b(X_i \beta) + c(\mathbf{y}) \right), \quad (4.3)$$

where X represents an $n \times p$ design matrix and X_i denotes the i th row of X , \mathbf{y} denotes the vector of response values, and b and c are known functions. In particular, when $b(t) = t^2/2$, (4.3) reduces to a normal distribution which corresponds to the linear regression model $\mathbf{y} = X\beta + \epsilon$ with $\epsilon \sim N(0, I_n)$. Note that in model (4.3), for simplicity, there is no precision parameter although this will be formally considered in my future work. Let $\gamma_j = I(\beta_j \neq 0)$

for $j = 1, \dots, p$. Consider the following hierarchical model

$$\begin{aligned}\beta_j | \gamma_j, \lambda &\sim (1 - \gamma_j)\delta_0 + \gamma_j N(0, \lambda) \\ \lambda &\sim p(\lambda) \\ \gamma &\sim p(\lambda).\end{aligned}$$

By a standard calculation, the posterior probability of γ is

$$p(\gamma | \mathbf{y}, X) \propto p(\gamma) \int_0^\infty p(\lambda) \left(\frac{1}{2\pi\lambda} \right)^{|\gamma|/2} \left(\int_{\mathbb{R}^{|\gamma|}} \exp(-nl_n(\beta_\gamma)) d\beta_\gamma \right) d\lambda,$$

where

$$l_n(\beta_\gamma) = \left(-\mathbf{y}^T X_\gamma \beta_\gamma + \sum_{i=1}^n b(X_{i\gamma} \beta_\gamma) - \beta_\gamma^T \beta_\gamma / (2\lambda) \right) / n.$$

The integral $\int_{\mathbb{R}^{|\gamma|}} \exp(-nl_n(\beta_\gamma)) d\beta_\gamma$ generally does not have a closed form. To address this difficulty, one has to find an effective way to approximate this integral. Wang and George (2007) applied a Laplace approximation to this integral. However, it has been pointed out by Shun and McCullagh (1995) that the Laplace approximation is ineffective when the dimension of the integral is large, which occurs in our problem. To better approximate the integral with a large dimension, Shun and McCullagh (1995) proposed a modified Laplace approximation. In my future study, I plan to apply this modified Laplace approximation to approximate $p(\gamma | \mathbf{y}, X)$ and obtain some theoretical and computational results.

The second issue was motivated from compressed sensing and may have some potential applications in spatial data analysis. Usually, in order to recover an image, one needs to use the measurements from all the pixels on the image, which is expensive. In compressed sensing, a central problem is how to recover (as accurately as possible) an image using only a small number of samples. Usually, the number of samples used in compressed sensing is far

less than the number of pixels on the image, which saves a lot of sampling costs. In my future work, I intend to apply this idea to spatial image data (involving multiple images). This is a generalization of typical compressed sensing problems which only involve one image. This generalization, in addition to saving sampling costs, can also allow us to study the relationship among the spatial images.

To better formulate the problem, let $\mathbf{s}_i, i = 1, \dots, n$, be a set of (evenly spaced) locations in $[0, 1) \times [0, 1)$ as defined in Section 3.1. We still use y and x to denote the covariate and response images over the square region $[0, 1) \times [0, 1)$. In our real data, y denotes the defoliation rate and x denotes the elevation. Suppose that we observe the x values at all the n locations while we only observe $u_q = w_{q1}y(\mathbf{s}_1) + \dots + w_{qn}y(\mathbf{s}_n)$, $q = 1, \dots, Q$, where w_{li} s are either deterministic or random numbers. Each u_q is called a projection of the $y(\mathbf{s}_i)$ s. Note that it is the u_q s that are observed, instead of $y(\mathbf{s}_i)$ s. When the w_{qi} s are random, u_q is a random projection. The goal is to fit a regression model that links y and x and estimate the model parameters, for which we consider the following regression model

$$u_q = \mathbf{w}_q(\mathbf{A} + \mathbf{x} \circ \mathbf{B}) + \epsilon_q, q = 1, \dots, Q, \quad (4.4)$$

where $\mathbf{A} = (A(\mathbf{s}_1), \dots, A(\mathbf{s}_n))$ and $\mathbf{x} \circ \mathbf{B} = (x(\mathbf{s}_1)B(\mathbf{s}_1), \dots, x(\mathbf{s}_n)B(\mathbf{s}_n))$. Applying a wavelet transformation to the surfaces A and B , one can rewrite model (4.4) as

$$\mathbf{U} = \tilde{X}\beta + \epsilon,$$

where \tilde{X} is a design matrix induced by the w_{qi} s, the wavelet transformation and the x values at the entire spatial locations $\mathbf{s}_i, i = 1, \dots, n$, $\mathbf{U} = (u_1, \dots, u_Q)'$. (Note that if only part of the x values are available, one cannot construct \tilde{X} , and therefore, this approach does not

work.) In future work, I will use LASSO to estimate β and then reconstruct the surfaces A and B .

Bibliography

- [1] Agarwal, D. K., Gelfand, A. E., Sirmans, C. F., and Thibadeau, T. G. (2003). Non-stationary Spatial House Price Model. Unpublished paper.
- [2] Brown, P., Fearn, T. and Vannucci, M. (2001). Bayesian Wavelet Regression on Curves With Application to a Spectroscopic Calibration Problem. *Journal of the American Statistical Association*, **96**, 398–408.
- [3] Berger, J. O., Ghosh, J. K. and Mukhopadhyay, N. (2003). Approximations and consistency of Bayes factors as model dimension grows. *J. Statist. Planning. Inference*. **112**, 241–258.
- [4] Bühlmann, P., and Kalisch, M. and Maathuis, M. H. (2010). Variable Selection in High-Dimensional Linear Models: Partially Faithful Distributions and the PC-simple Algorithm. *Biometrika* **97**, 261–278.
- [5] Berger, J. O. and Pericchi, L. (1996). The Intrinsic Bayes Factor for Model Selection and Prediction. *J. Amer. Statist. Assoc.* **91**, 109–122.
- [6] Brown, P., Vannucci, M. and Fearn, T. (2002). Bayes Model Averaging with Selection of Regressors. *Journal of the Royal Statistical Society, Series B*, **64**, 519–536.
- [7] Casella, G., Girón, F. J., Martínez, M. L. and Moreno, E. (2009). Consistency of Bayesian Procedures for Variable Selection. *The Annals of Statistics*, **37**, 1207–1228.
- [8] Chipman, H., Kolaczyk, E. and McCulloch, R. (1997). Adaptive Bayesian Wavelet Shrinkage. *Journal of the American Statistical Association*, **92**, 1413–1421.

- [9] Clyde, M. and George, E. (2000). Flexible Empirical Bayes Estimation for Wavelets. *Journal of the Royal Statistical Society, Series B*, **62**, 681–698.
- [10] Clyde, M., Parmigiani, G. and Vidakovic, B. (1998). Multiple Shrinkage and Subset Selection in Wavelets. *Biometrika*, **85**, 391–401.
- [11] Crouse, M. S., Nowak, R. D. and Baraniuk, R. G. (1998). Wavelet-Based Statistical Signal Processing Using Hidden Markov Models. *IEEE Transactions on Signal Processing*, **46**, 886–902.
- [12] Daubechies, I. (1992). *Ten Lectures on Wavelets*. CBMS-NSF Regional Conference Series in Applied Mathematics 61.
- [13] Dunson, D. (2009). Nonparametric Bayes Local Partition Models for Random Effects. *Biometrika*, **96** 249–262.
- [14] Durrett, R. (2005). *Probability: Theorey and Examples*. 3rd Ed. Wadsworth-Brooks/Cole, Pacific Grove.
- [15] Size, power and false discovery rates. *The Annals of Statistics*, **35**, 1351–1377.
- [16] Efron, B., Johnstone, I., Hastie, T. and Tibshirani, R. (2002). Least Angle Regression. *The Annals of Statistics*, **32**, 407–451.
- [17] Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, **96**, 1348–1360.
- [18] Fan, J. and Lv, J. (2008). Sure Independence Screening for Ultrahigh Dimensional Feature Space. *J. R. Stat. Soc. Ser. B.* **70**, 849–911.

- [19] Fan, J. and Peng, H. (2004). Nonconcave Penalized Likelihood with a Diverging Number of Parameters. *The Annals of Statistics*, **32**, 928–961.
- [20] Fan J., Wu, Y. and Feng, Y. (2010). Local Quasi-likelihood With a Parametric Guide. *The Annals of Statistics*, **37**, 4153–4183.
- [21] Fernández, C., Ley, E., and Steel, M. F. (2001). Benchmark Priors for Bayesian Model Averaging. *Journal of Econometrics*, **100**, 381–427.
- [22] Fu, W. J. (1998). Penalized regressions: the bridge vs the lasso. *Journal of Computational and Graphical Statistics*, **7**, 396–416.
- [23] Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2003). *Bayesian Data Analysis* (2nd ed). Chapman & Hall/CRC.
- [24] Gelfand, A., Kim, H., Sirmans, C. and Banerjee, S.(2003). Spatial Modeling With Spatially Varying Coefficient Processes. *Journal of the American Statistical Association*, **98**, 387–396.
- [25] George, E. and McCulloch, R. (1993). Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association*, **88**, 881–889.
- [26] Girón, F. J., Moreno, E., Casella, G. and Martínez, M. L. (2010). Consistency of objective Bayes factors for nonnested linear models and increasing model dimension. *Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales. Serie A. Matemáticas* **104**, 57–67.
- [27] Godsill, J. S. and Rayner, P. J. W. (1998). Robust Reconstruction and Analysis of Autoregressive Signals in Impulsive Noise Using the Gibbs Sampler. *IEEE Transactions on Speech and Audio Processing*, **6**, 352–372.

- [28] He, X. (1996). Bivariate Tensor-Product B-Splines in a Partly Linear Model. *Journal of Multivariate Analysis*, **58**, 162–181.
- [29] Jeffreys, H. (1967). *Theory of Probability*. 4th Ed. Oxford Univ. Press, Oxford.
- [30] Jiang, W. (2007). Bayesian Variable Selection for High Dimensional Generalized Linear Models: Convergence Rates of the Fitted Densities. *The Annals of Statistics*, **35**, 1487–1511.
- [31] Jiang, W. and Tanner, M. (2008). Gibbs Posterior for Variable Selection in High-Dimensional Classification and Data Mining. *The Annals of Statistics*, **36**, 2207–2231.
- [32] Kim, S.-J., Koh, K., Lustig, M., Boyd, S. P., and Gorinevsky, D. (2007). An Interior-Point Method for Large- Scale l_1 -Regularized Least Squares. *IEEE Journal on Selected Topics in Signal Processing*, **1**, 606–617.
- [33] Kleiner, K. and Montgomery, M. (1994). Forest Stand Susceptibility to the Gypsy-Moth (lepidoptera, lymantriidae)–Species and Site Effects on Foliage Quality to Larvae. *Environmental Entomology*, **23**, 699–711.
- [34] Liang, F., Paulo, R., Molina, G., Clyde, M. and Berger, J. (2008). Mixtures of g -Priors for Bayesian Variable Selection. *Journal of the American Statistical Association*, **103**, 410–423.
- [35] Li, F. and Zhang, N. R. (2010). Bayesian Variable Selection in Structured High-Dimensional Covariate Spaces with Applications in Genomics. *Journal of the American Statistical Association*, **105**, 1202–1214.
- [36] Meinshausen, N. and Bühlmann, P. (2006). High-Dimensional Graphs and Variable Selection with the Lasso. *The Annals of Statistics* **34**, 1436–1462.

- [37] Meinshausen, N. and Yu, B. (2009). LASSO-type Recovery of Sparse Representations for High-Dimensional Data. *The Annals of Statistics*, **37**, 246-270.
- [38] Moreno, E., Bertolino, F. and Racugno, W. (1998). An Intrinsic Limiting Procedure for Model Selection and Hypotheses Testing. *J. Amer. Statist. Assoc.* **93**, 1451–1460.
- [39] Moreno, E. and Girón, F. J. (2005). Consistency of Bayes Factors for Intrinsic Priors in Normal Linear Models. *C. R. Math. Acad. Sci. Paris* **340**, 911–914.
- [40] Moreno, E., Girón, F. J. and Casella, G. (2010). Consistency of Objective Bayes Factors as the Model Dimension Grows. *The Annals of Statistics*, **38**, 1937–1952.
- [41] Marx, B. and Eilers, P. (2005). Multidimensional penalized signal regression. *Technometrics*, **47**, 13–22.
- [42] Nott, D. and Green, P. (2004). Bayesian Variable Selection and Swendsen-Wang Algorithm. *Journal of Computational and Graphical Statistics*, **13**, 141–157.
- [43] Osborne, M., Presnell, B. and Turlach, B. (2000). On the LASSO and its dual. *Journal of Computational and Graphical Statistics*, **9** 319–337.
- [44] Romberg, J., Choi, H. and Baraniul, R. (2001). Bayesian Tree-Structured Image Modeling Using Wavelet-Domain Hidden Markov Models. *IEEE Transactions on Image Processing*, **10**, 1056–1068.
- [45] Shao, J. (2003). *Mathematical Statistics*, 2nd Ed. Springer Texts in Statistics. Springer, New York.
- [46] Shang, Z. and Clayton, M. K. (2011). Consistency of Bayesian Model Selection for Linear Models With A Growing Number of Parameters. *Journal of Statistical Planning and Inference*, in press.

- [47] Shun, Z. and McCullagh, P. (1995). Laplace Approximation of High Dimensional Integrals. *Journal of the Royal Statistical Society, Series B*, **57**, 749–760.
- [48] Smith, M. and Fahrmeir, L. (2007). Spatial Bayesian Variable Selection With Application to Functional Magnetic Resonance Imaging. *Journal of the American Statistical Association*, **102**, 417–431.
- [49] Smith, M. S. and Kohn, R. (1996). Nonparametric Regression Using Bayesian Variable Selection. *Journal of Econometrics* **75**, 317–344.
- [50] Seber, G. A. F. and Lee, A. J. (2003). *Linear Regression Analysis*, 2nd Ed. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ.
- [51] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, **58** 267–288.
- [52] Townsend, P. A., Eshleman, K. N. and Welcker, C. (2004). Remote Sensing of Gypsy Moth Defoliation to Assess Variations in Stream Nitrogen Concentrations. *Ecological Applications*, **14**, 504–516.
- [53] Wang, X. and George, E. (2007). Adaptive Bayesian Criteria in Variable Selection for Generalized Linear Models. *Statistica Sinica*, **17**, 667–690.
- [54] Wolfe, P., Godsill, S. and Ng, W. (2004). Bayesian Variable Selection and Regularization for Time-Frequency Surface Estimation. *Journal of the Royal Statistical Society, Series B*, **66**, 575–589.
- [55] Yuan, M. and Lin, Y. (2005). Efficient Empirical Bayes Variable Selection and Estimation in Linear Models. *Journal of the American Statistical Association*, **100**, 1215–1225.

- [56] Yuan, M. and Lin, Y. (2006). Model Selection and Estimation in Regression with Grouped Variables. *Journal of the Royal Statistical Society, Series B*, **68**, 49–67.
- [57] Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. Wiley, New York.
- [58] Zellner, A. (1978). Jeffreys-Bayes Posterior Odds Ratio and the Akaike Information Criterion for Discriminating Between Models. *Econom. Lett.* **1**, 337–342.
- [59] Zhang, J., Clayton, M. K. and Townsend, P. A. (2011). Functional Concurrent Linear Regression Model for Spatial Images. *Journal of Agricultural, Biological and Environmental Statistics*, **16**, 105–130.
- [60] Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *The Annals of Statistics* **36**, 1567–1594.
- [61] Zou, H. (2006). The adaptive LASSO and its oracle property. *Journal of the American Statistical Association*, **101**, 1418–1428.
- [62] Zou, H. and Trevor, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B*, **67**, 301–320.