VARIABLE SELECTION FOR ULTRA HIGH DIMENSIONAL DATA

A Dissertation

by

QIFAN SONG

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Faming Liang
Committee Members,	Raymond Carroll
	Valen Johnson
	Soumendra Lahiri
	Jianxin Zhou
Head of Department,	Simon Sheather

August 2014

Major Subject: Statistics

ABSTRACT

Variable selection plays an important role for the high dimensional data analysis. In this work, we first propose a Bayesian variable selection approach for ultra-high dimensional linear regression based on the strategy of split-and-merge. The proposed approach consists of two stages: (i) split the ultra-high dimensional data set into a number of lower dimensional subsets and select relevant variables from each of the subsets, and (ii) aggregate the variables selected from each subset and then select relevant variables from the aggregated data set. Since the proposed approach has an embarrassingly parallel structure, it can be easily implemented in a parallel architecture and applied to big data problems with millions or more of explanatory variables. Under mild conditions, we show that the proposed approach is consistent. That is, asymptotically, the true explanatory variables will be correctly identified by the proposed approach as the sample size becomes large. Extensive comparisons of the proposed approach have been made with the penalized likelihood approaches, such as Lasso, elastic net, SIS and ISIS. The numerical results show that the proposed approach generally outperforms the penalized likelihood approaches. The models selected by the proposed approach tend to be more sparse and closer to the true model.

In the frequentist realm, penalized likelihood methods have been widely used in variable selection problems, where the penalty functions are typically symmetric about 0, continuous and nondecreasing in $(0, \infty)$. The second contribution of this work is that, we propose a new penalized likelihood method, reciprocal Lasso (or in short, rLasso), based on a new class of penalty functions which are decreasing in $(0, \infty)$, discontinuous at 0, and converge to infinity when the coefficients approach

zero. The new penalty functions give nearly zero coefficients infinity penalties; in contrast, the conventional penalty functions give nearly zero coefficients nearly zero penalties (e.g., Lasso and SCAD) or constant penalties (e.g., L_0 penalty). This distinguishing feature makes rLasso very attractive for variable selection: It can effectively avoid selecting overly dense models. We establish the consistency of the rLasso for variable selection and coefficient estimation under both the low and high dimensional settings. Since the rLasso penalty functions induce an objective function with multiple local minima, we also propose an efficient Monte Carlo optimization algorithm to solve the minimization problem. Our simulation results show that the rLasso outperforms other popular penalized likelihood methods, such as Lasso, SCAD, MCP, SIS, ISIS and EBIC. It can produce sparser and more accurate coefficient estimates, and have a higher probability to catch true models.

ACKNOWLEDGEMENTS

I would have never been able to finish my dissertation without the guidance of my committee members, help from friends, and support from my family.

Foremost, I would like to express my sincere gratitude to my Ph.D. advisor Prof. Faming Liang for his constant encouragement and support of my Ph.D study and research, for his patience, motivation and passion. His guidance and caring helped me in all the time of my research and writing of this dissertation. I could not have imagined having a better advisor and mentor.

Besides, I would like to thank the rest of my thesis committee: Prof Raymond Carroll, Prof. Valen Johnson, Prof. Soumendra Lahiri and Prof. Jianxin Zhou, for their insightful comments and suggestions. And my special thank goes to Dr. Ellen Toby, for offering my the opportunities to work for her as a teaching assistant.

I would also like to thank many of my fellow postgraduate students in our department: Dr. Yichen Cheng, Dr. Rubin Wei, Dr. Kun Xu, Ranye Sun, Jinsu Kim, Abhra Sarkar, Karl Gregory, Xiaoqing Wu, Nan Zhang, Dr. Yanqing Wang, for their kindness and friendship.

The Department of Statistics and Institute of Applied Mathematics and Computational Science have provided the support and equipment I have needed to produce and complete my thesis.

Last but not the least, I thank my parents and my wife for their unconditional love and support.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION: VARIABLE SELECTION AND HIGH DIMENSIONALITY

1.1 History of High Dimensional Variable Selection

Variable selection is fundamental to statistical modeling of high dimensional problems which nowadays appear in many areas of scientific discoveries. Consider the linear regression model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1.1}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ is the response variable, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$ is the vector of regression coefficients, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) \in \mathbb{R}^{n \times p}$ is the design matrix, $\boldsymbol{x}_i \in \mathbb{R}^n$ is the ith predictor, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T \in \mathbb{R}^n$, and ϵ_i's are independently identically distributed (i.i.d.) random variables with mean 0 and variance σ^2. Under the high dimensional setting, one often assumes that $p \gg n$ and p can increase with n, whilst the true model is sparse, i.e. there are only few predictors whose regression coefficients are nonzero. Such a sparsity assumption is introduced by either a mathematical requirement in order to derive a solution or an expert's opinion that only few key predictors can causally influence the outcome. Identification of the causal predictors (also known as true predictors) is of particular interest, as which can avoid overfitting in model estimation and yield interpretable systems for future studies.

When analyzing data in the high dimensional spaces, a phenomenon, that we always encounter, is so called the curse of dimensionality. In the problem of model selection, the total number of candidate models is 2^p, which increases exponentially as p increases. There are two main impacts due to the high dimensionality. The first

is the noise accumulation. If we have to estimate all p parameters, the accumulated noise by the parameter estimations leads to poor prediction. The second is the spurious correlation. As pointed out by [21], even independent variables will demonstrate very high sample correlation under the high dimensional situation. Figure 1.1 gives the estimated distributions of maximum absolute sample correlation and distribution of the maximum absolute sample correlation provided that the variables are generated by independent normal distribution, with $n = 50$ and $p = 1000$ or 10000. As a result, any variable, especially the true important one, can be approximated by other spurious variables, and it may lead to a total wrong conclusion.

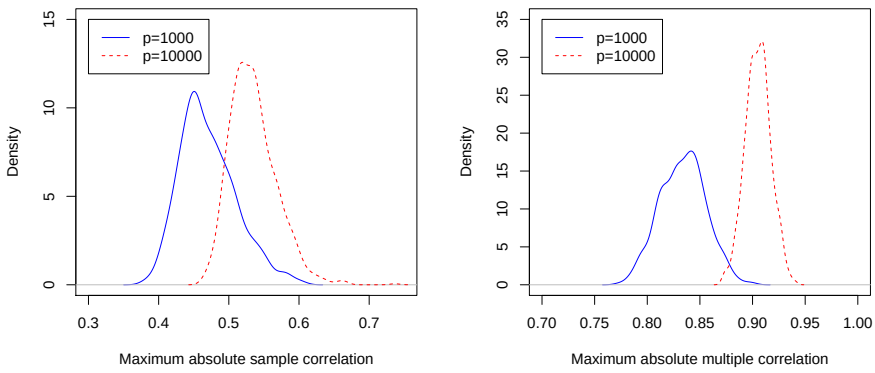

Figure 1.1: Distribution of spurious correlation due to dimensionality. Left: distribution of maximum absolute sample correlation with n=50; Right: distribution of the maximum absolute multiple correlation, with $n = 50$.

In the literature, the problem of variable selection is often treated with penalized likelihood methods. For linear regression, as the dispersion parameter σ can be estimated separately from $\boldsymbol{\beta}$, the variable selection can be done by minimizing the

penalized residual sum of squares

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + P_\lambda(\boldsymbol{\beta})\}, \tag{1.2}$$

where $P_\lambda(\cdot)$ is the so-called penalty function and λ is a tuning parameter which can be determined via cross-validation. The penalty function serves to control the complexity of the model, and its choice determines the behavior of the method. [65] proposed the Lasso method which employs a L_1-penalty of the form

$$P_\lambda(\boldsymbol{\beta}) = \lambda \sum_{i=1}^{p} |\beta_j|.$$

Since Lasso gives small penalties to small coefficients , the resulting model tends to be dense, which may include many spurious predictors with very small coefficients. As shown in [71] and [72], Lasso might not be consistent for variable selection unless a strong representable condition holds. To remedy this flaw, various methods have been proposed, such as adaptive Lasso [72], SCAD [18], and MCP [69]. Adaptive Lasso assigns different weights for penalizing different coefficients in the L_1 penalty. SCAD and MCP employ some penalty functions that are concave on $(0, \infty)$ and converge to constants as $|\beta|$ becomes large, and thus reduce the estimation bias when the true coefficients are large. Although these methods have shown some improvements over Lasso, they still tend to produce dense models especially when the ratio $\log(p)/n$ is large. This is because, as illustrated in Section 3, these methods share the same feature with the Lasso that nearly zero coefficients are given nearly zero penalties.

[11] showed that subject to the L_1 penalty, the estimator of $\boldsymbol{\beta}$ can only achieve a mean squared error up to a logarithmic factor $\log(p)$. Motivated by this result, [19] and [20] proposed the sure independence screening (SIS) method and its iterative

version, iterative SIS (ISIS). The SIS is to first reduce the dimension p by screening out the predictors whose marginal utility is low, and then apply a penalized likelihood method, such as Lasso or SCAD, to select appropriate predictors. The marginal utility measures the usefulness of a single predictor for predicting the response, and it can be chosen as the marginal likelihood or simply the marginal correlation for linear regression. ISIS iteratively selects predictors from remaining unselected predictors and thus reduces the risk of missing true predictors. Since many false predictors can be removed in the screening stage, SIS and ISIS can generally improve the performance of Lasso and SCAD in variable selection.

Back to 1970's, [1, 2] proposed the AIC which is to select predictors by minimizing the Kullback-Leibler divergence between the fitted and true models. Later, [60] proposed the Bayesian information criterion (BIC). Both AIC and BIC employ the L_0 penalty, which is given by

$$P_\lambda(\boldsymbol{\beta}) = \sum_{i=1}^{p} \lambda I(\beta_i \neq 0).$$

Recently, BIC has been extended to high dimensional problems by [12, 13], which establish the consistency of the extended BIC (EBIC) method for variable selection under the conditions $p = O(n^\kappa)$ and $\lambda = \log(n)/2 + \gamma \log(p)$, where $\gamma > 1 - 1/2\kappa$ is a user-specified parameter. The L_0-regularization method has a natural interpretation of best subset selection, where each coefficient is penalized by a constant factor. It is easy to figure out that this constant factor is of order $O(\log(p))$. Compared to the Lasso-type penalties, such as those used in Lasso, adaptive Lasso, SCAD and MCP, the L_0-penalty overcomes the drawback of small coefficients small penalties. However, for the problems for which the ratio $\log(p)/n$ is large, a penalty of order $O(\log(p))$ seems too big. It is known that the models produced by EBIC under

this scenario tend to be overly sparse. Other major concern with EBIC is the lack of an efficient optimization algorithm to search over the model space. Because the model space is discrete, some computationally intensive algorithms, such as Markov chain Monte Carlo (MCMC), have to be used. We note that the SCAD and MCP penalties can also be viewed as an approximation to the L_0 penalty, as they converge to constants when the coefficients become large.

Parallel to the penalized likelihood approaches, various Bayesian approaches have also been developed to tackle this problem. These approaches advance either in a sophisticated model search algorithm, such as evolutionary stochastic search [8], or in a prior specification that is particularly suitable for high dimensional problems, such as the Lasso prior [56], Ising prior [47], proportion of variance explanation (PVE) prior [29], and nonlocal prior ([37], [36]). Although various numerical examples have demonstrated the successes of Bayesian approaches on this problem, there still lacks a rigorous theory to show that the Bayesian approaches can lead to a consistent selection of true predictors when p is greater than n. On the other hand, the high computational demand of Bayesian approaches for sampling from the posterior distribution also hinders their popularity. Since the volume of the model space increases geometrically with the dimension p, the CPU time for a Bayesian approach should increase accordingly or even faster.

It is interesting to point out that many Bayesian variable selection methods are closely related to the L_0 penalty, see e.g., g-prior ([68], [51]), evolutionary stochastic search [8], and Bayesian subset selection (BSR) [52]. In these methods, a hierarchical prior is assumed for the model size and the regression coefficients associated with the model. For example, a commonly used prior for the model size is

$$P(\boldsymbol{\xi}) = \mu^{|\boldsymbol{\xi}|}(1-\mu)^{p-|\boldsymbol{\xi}|},$$

where $\boldsymbol{\xi}$ denotes the selected model with size $|\boldsymbol{\xi}|$, and μ denotes the prior probability that each predictor can be included in the model. The hyperparameter μ can be either determined using the empirical Bayes method [25] or treated as a beta random variable in a fully Bayesian method. It follows from [61] that both choices of μ will result in a prior of model size that is approximately equivalent to the L_0 penalty. Under the high dimensional setting, μ can also be chosen as a decreasing function of p as suggested by [52]. Like the empirical and fully Bayesian methods, this choice of μ also induces an automatic multiplicity adjustment for variable selection. It is worth pointing out that under the prior specification of [52], the maximum *a posteriori* (MAP) model is reduced to the EBIC model.

1.2 Two Proposed Approaches

In this work, two new approaches are proposed to solve the high dimensional variable selection problems in both Bayesian and Frequentist frameworks. The proposed Bayesian variable selection approach based on the strategy of split-and-merge. The proposed approach consists of two stages:

(i) Split the high dimensional data into a number of lower dimensional subsets and perform Bayesian variable selection for each subset.

(ii) Aggregate the variables selected from each subset and perform Bayesian variable selection for the aggregated data set.

Under mild conditions, we show that the proposed approach is consistent. That is, the true model will be identified in probability 1 as the sample size becomes large. Henceforth, we will call the proposed approach the split-and-merge (SaM) approach for the purpose of description simplicity.

Our contribution in this Bayesian variable selection method is two-fold. First, we propose a computationally feasible Bayesian approach for ultra-high dimensional

regression. Since SaM has an embarrassingly parallel structure, it can be easily implemented in a parallel architecture and applied to the big data problems with millions or more of predictors. This has been beyond the ability of conventional Bayesian approaches, as they directly work on the full data set. Second, under mild conditions we show that the Bayesian approach can share the same asymptotics, such as sure screening and the model selection consistency, as the SIS approach. In spirit, SaM is similar to SIS. The first stage serves the purpose of dimension reduction, which screens out uncorrelated predictors; and the second stage refines the selection of predictors. However, compared to SIS and ISIS, SaM can often lead to more accurate selection of true predictors. This is because SaM screens uncorrelated predictors based on the marginal inclusion probability, which has incorporated the joint information of all predictors contained in a subset. While, as previously mentioned, SIS makes use of only the marginal information of each predictor. ISIS tries to incorporate information from other predictors, but in an indirect way.

Finally, we note that the strategy of split-and-merge, or otherwise known as divide-and-conquer, has been often used in big data analysis, see e.g., [55], [41], [54] and [67]. Our use of this strategy is different from others. In SaM, the data split is done in the dimension of predictors, while this is done in the dimension of observations in other work. This difference necessitates the development of new theory for variable selection under the situation that the model is misspecified due to the missing of some true predictors.

In the second approach, we propose a new class of penalty functions for penalized likelihood variable selection, which are decreasing in $(0, \infty)$, discontinuous at 0, and converge to infinity when the coefficients approach zero. The new penalty function

has a typical form of

$$P_\lambda(\boldsymbol{\beta}) = \sum_{i=1}^{p} \frac{\lambda}{|\beta_i|} I(\beta_i \neq 0),$$

which gives nearly zero coefficients infinity penalties. In contrast, the conventional penalty functions give nearly zero coefficients nearly zero penalties (e.g., the Lasso-type penalties) or constant penalties (e.g., the L_0 penalty). Although unusual, the new penalty function possesses a very intuitive and appealing interpretation: The smaller coefficient a predictor has, the more likely the predictor is a false predictor and thus the higher penalty it should have. If a predictor has a nearly zero coefficient, then the effect of the predictor on model estimation and future prediction should be very limited and, hence, it had better be excluded from the model for the sake of model simplicity. Our numerical results show that the new penalty can outperform both the Lasso-type penalties and the L_0 penalty in identifying true models.

The contribution of this frequentist method is three-fold:

(i) We propose a novel penalty function for variable selection, which overcomes some drawbacks of the existing penalty functions. Compared to the Lasso-type penalties, it gives small coefficients large penalties and thus avoids to select overly dense models. Compared to the L_0-penalty, it is coefficient dependent and thus adaptive for different coefficients; in addition, as discussed at the end of Section 3, it allows the tuning parameter λ to take values at a lower order of $\log(p)$ and thus avoids to select overly sparse models.

(ii) We establish the consistency of the new method for variable selection and parameter estimation. Under the low dimensional setting, where p is fixed as n increases, we show that the new method possesses the oracle property. That is, the true model can be consistently selected and the coefficient estimator is as efficient as the ordinary least square (OLS) estimator. Under the high

dimensional setting, where $p \gg n$ and p increases with n, we show that the true model can also be consistently selected, and that the coefficient estimator can converge in L_2 and is thus consistent.

(iii) We propose an efficient Monte Carlo optimization algorithm based on the coordinate descent ([9], [24]), stochastic approximation Monte Carlo [50] and stochastic approximation annealing [49] algorithms to solve the optimization (1.2) with the proposed penalty function. The proposed optimization algorithm can deal with problems with multiple local minima and is thus also applicable for other L_0-regularization methods.

The remaining part of this dissertation is organized as follows. We describe the Bayesian variable selection approach in Section 2. Section 2.1 establishes consistency of the SaM approach for variable selection. Section 2.2 summarizes the SaM approach and discusses some implementation issues. Section 2.3 illustrates the performance of the SaM approach using two simulated examples along with comparisons with penalized likelihood approaches. Section 2.4 compares the SaM approach with penalized likelihood approaches on two real data examples. In Section 3, we describe our penalized likelihood methods with the new proposed penalty function. In Section 3.1, we establish the consistency of the new method for variable selection and parameter estimation under the low dimensional setting. In Section 3.2, we extend the low dimensional results to the high dimensional case. In Section 3.3, we describe the Monte Carlo optimization algorithm. In Section 3.4, we present some numerical results to illustrate the performance of the new method. In Section 4, we conclude the SaM algorithm and rLasso method with a brief discussion, We present related proofs of the theorems in Section 2 and Section 3 in the Appendix A and Appendix B respectively, and provide some related miscellaneous material in Appendix C.

2. BAYESIAN APPROACH: SPLIT AND MERGE STRATEGY AND BIG DATA ANALYSIS

In this section, we address the proposed Split and Merge method for high dimensional regression analysis, especially when dealing with massive data set. Before describe the detailed algorithm, we first lay the theoretical background of the Bayesian variable selection consistency in Section 2.1.

2.1 Variable Selection Consistency of the SaM Approach

This section is organized as follows. In Section 2.1.1, we establish the posterior consistency for correctly specified models; that is, if the model is correctly specified with all true predictors included in the set of candidate predictors, the true density of the model (1.1) can be estimated consistently by the density of the models sampled from the posterior as the sample size becomes large. In Section 2.1.2, we establish the sure screening property for the correctly specified models based on the marginal inclusion probability; that is, the marginal inclusion probability of the true predictors will converge to 1 in probability as the sample size becomes large. In this section, we also establish the consistency of the maximum *a posteriori* (MAP) model; that is, the MAP model is a consistent estimator of the true model. In Section 2.1.3, we show that for the misspecified models for which some true predictors are missed in the set of candidate predictors, the sure screening property based on the marginal inclusion probability still holds. Combining the theoretical results established in Sections 2.1.1 to Section 2.1.3 leads to the variable selection consistency of the SaM approach.

2.1.1 Posterior Consistency for Correctly Specified Models

Let $D_n = \{\boldsymbol{y}; \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{p_n}\}$ denote a data set of n observations drawn from model (1.1). Let $x = (x_1, \ldots, x_{p_n})^T$ denote a generic observation of the p_n predictors, and let y denote a generic observation of the response variable corresponding to x. Let ν_x denote the probability measure of x. Then the true probability density of (y, x) is given by

$$f^*(y,x)dxdy = \frac{1}{\sqrt{2\pi}\sigma^*}\exp\left\{-\frac{(y-x^T\boldsymbol{\beta}^*)^2}{2\sigma^{*2}}\right\}\nu_x(dx)dy = \phi(y; x^T\boldsymbol{\beta}^*, \sigma^*)\nu_x(dx)dy, \tag{2.1}$$

where $\phi(\cdot; \mu, \sigma)$ denotes a Gaussian density function with mean μ and standard deviation σ.

As in conventional studies of high dimensional variable selection, we consider the asymptotics of the SaM approach under the assumption that p_n increases as n increases. To be specific, we assume that $p_n \succ n^\theta$ for some $\theta > 0$, where $b_n \succ a_n$ means $\lim_{n\to\infty} a_n/b_n = 0$. In addition, we assume that the true model is sparse. Two sparseness conditions are considered in this work. One states that most components of $\boldsymbol{\beta}^*$ are very small in magnitude such that

$$\lim_{n\to\infty}\sum_{j=1}^{p_n}|\beta_j^*| < \infty, \tag{2.2}$$

where β_j^* denotes the jth entry of $\boldsymbol{\beta}^*$. The other one is stronger but more usual, which requires that most components of $\boldsymbol{\beta}^*$ are zero such that

$$\lim_{n\to\infty}\sum_{j=1}^{p_n}1\{\beta_j^* \neq 0\} < \infty. \tag{2.3}$$

Let $\xi \subset \{1, \ldots, p_n\}$ denote a subset model, which includes the indices of all

selected predictors. Let $|\xi|$ denote the number of predictors included in ξ. Thus, the strict sparsity condition (2.3) implies that the true model $\boldsymbol{t} = \{i : \beta_i^* \neq 0, 1 \leq i \leq p_n\}$ has a finite model size. For a model $\xi = \{i_1, i_2, \ldots, i_{|\xi|}\}$, we let $\boldsymbol{X}_\xi = (\boldsymbol{x}_{i_1}, \boldsymbol{x}_{i_2}, \ldots, \boldsymbol{x}_{i_{|\xi|}})$ denote the design matrix, and let $\boldsymbol{\beta}_\xi = (\beta_{i_1}, \ldots, \beta_{i_{|\xi|}})^T$ denote the regression coefficient vector. Let $L(D_n|\xi, \boldsymbol{\beta}_\xi, \sigma^2)$ denote the likelihood function, and let $\pi(\xi, d\boldsymbol{\beta}_\xi, d\sigma^2)$ denote the joint prior distribution imposed on ξ and the parameters σ^2 and $\boldsymbol{\beta}_\xi$. Then the posterior distribution is given by

$$\pi(\xi, d\boldsymbol{\beta}_\xi, d\sigma^2|D_n) = \frac{L(D_n|\xi, \boldsymbol{\beta}_\xi, \sigma^2)\pi(\xi, d\boldsymbol{\beta}_\xi, d\sigma^2)}{\sum_\xi \int_{\boldsymbol{\beta}_\xi, \sigma^2} L(D_n|\xi, \boldsymbol{\beta}_\xi, \sigma^2)\pi(\xi, d\boldsymbol{\beta}_\xi, d\sigma^2)}. \tag{2.4}$$

2.1.1.1 Prior Specification

For the model ξ, we assume that each predictor has a prior probability $\lambda_n = r_n/p_n$, independent of other predictors, to be included in the model ξ. Further, we impose a constraint on the model size such that $|\xi| \leq \bar{r}_n$, where the model size upper bound $\bar{r}_n$ is pre-specified. Then we have

$$\pi(\xi) \propto \lambda_n^{|\xi|}(1 - \lambda_n)^{p_n - |\xi|} I[|\xi| \leq \bar{r}_n]. \tag{2.5}$$

The constraint on $|\xi|$ is not crucial to the data analysis. Since the true model is sparse, there is not much loss as long as the size of the true model is less than $\bar{r}_n$. From the perspective of computation, such a constraint facilitates the posterior simulation, as the computational complexity of the likelihood evaluation is $O(|\xi|^3)$, which involves inverting a covariance matrix of size $|\xi|$ by $|\xi|$.

We let the variance σ^2 be subject to an Inverse-Gamma prior distribution with the hyper-parameters a_0 and b_0; i.e., $\sigma^2 \sim IG(a_0, b_0)$. Conditioned on the model ξ

and σ^2, we let $\boldsymbol{\beta}_\xi$ be subject to a Gaussian prior,

$$\boldsymbol{\beta}_\xi|\xi, \sigma^2 \sim N_{|\xi|}(0, \sigma^2 V_\xi),$$

where V_ξ is a positive definite matrix depending on ξ.

To facilitate computation, we integrate out $\boldsymbol{\beta}_\xi$ and σ^2 from the posterior distribution (2.4). The resulting posterior of ξ is given by

$$\pi(\xi|D_n) \propto (r_n/p_n)^{|\xi|}(1 - r_n/p_n)^{p_n-|\xi|} \frac{\sqrt{\det(V_\xi^{-1})}}{\sqrt{\det(\boldsymbol{X}_\xi^T \boldsymbol{X}_\xi + V_\xi^{-1})}}$$
$$\times \left\{2b_0 + y^T(I - \boldsymbol{X}_\xi(\boldsymbol{X}_\xi^T \boldsymbol{X}_\xi + V_\xi^{-1})^{-1}\boldsymbol{X}_\xi^T)y\right\}^{-n/2-a_0} I[|\xi| \leq \bar{r}_n]. \quad (2.6)$$

For a fully Bayesian variable selection approach, λ_n is usually subject to a Beta prior. As pointed out by [61], placing such a prior on λ_n will induce an automatic multiplicity adjustment for variable selection: The penalty for adding an extra variable increases as p_n increases. The multiplicity control is crucial for variable selection when p_n is greater than n. Otherwise, the resulting model tends to be liberal. For example, the BIC criterion places a flat prior for the model ξ, and the resulting models are overly liberal for high dimensional regression [10]. In this work, we assume $\lambda_n \to 0$ as $p_n \to \infty$, which corresponds to condition (2.12) in Theorem 2.1.1. This condition provides an automatic mechanism of multiplicity control for the SaM approach.

The posterior (2.6) is also closely related to the EBIC criterion [12], which is to choose a model ξ by minimizing the penalized likelihood function

$$EBIC(\xi) = -\hat{l}(\xi) + \frac{|\xi|}{2}\log n + \gamma|\xi|\log p_n,$$

where $\hat{l}(\xi)$ is the loglikelihood function of ξ evaluated at the maximum likelihood estimates of $\boldsymbol{\beta}_\xi$ and σ^2. If we set $V_\xi^{-1} = \boldsymbol{X}_\xi^T \boldsymbol{X}_\xi / n$, $a_0 = b_0 \approx 0$, and $r_n = p_n^{1-\gamma}$ for a constant $\gamma \in (0,1)$, then $\boldsymbol{X}_\xi^T \boldsymbol{X}_\xi + V_\xi^{-1} \approx \boldsymbol{X}_\xi^T \boldsymbol{X}_\xi$ and $\log(r_n/p_n)^{|\xi|}(1-r_n/p_n)^{p_n-|\xi|} \approx -\gamma|\xi| \log p_n + p_n^{1-\gamma}$. Therefore, $\log \pi(\xi|D_n) \approx -EBIC(\xi) + C$ for some constant C independent of ξ. A similar approximation between the posterior and EBIC can be found in [52] for generalized linear models.

2.1.1.2 Posterior Consistency

In what follows, we study the asymptotics of the posterior distribution with respect to the Hellinger distance under appropriate conditions of r_n, $\bar{r}_n$ and V_ξ. For two density functions $f_1(x,y)$ and $f_2(x,y)$, the Hellinger distance is defined as

$$d(f_1, f_2) = \sqrt{\int\int (f_1^{1/2}(x,y) - f_2^{1/2}(x,y))^2 dxdy}.$$

For a matrix Σ, we let $ch_i(\Sigma)$ denote the ith largest eigenvalue, and let $ch_i'(\Sigma)$ denote the ith smallest eigenvalue. Define $B(r_n) = \sup_{\xi:|\xi|=r_n} ch_1(V_\xi^{-1})$, $\bar{B}(r_n) = \sup_{\xi:|\xi|=r_n} ch_1(V_\xi)$, $\tilde{B}_n = \sup_{\xi:|\xi|\le\bar{r}_n} ch_1(V_\xi)$, and $\Delta(p) = \inf_{\xi:|\xi|=p} \sum_{i\notin\xi} |\beta_i^*|$.

Theorem 2.1.1 follows from [35]. Here we consider one of the simplest situations that all predictors have been standardized and uniformly bounded, and the sparsity condition (2.2) holds.

Theorem 2.1.1 (Posterior Consistency)**.** *Assume that the data set D_n is drawn from model (1.1) and all predictors $|x_i| \le 1$. If there exists a sequence $\{\epsilon_n\} \in (0,1)$ such*

that $n\epsilon_n^2 \succ 1$*, and the following conditions hold:*

$$\Delta(r_n) \prec \epsilon_n, \tag{2.7}$$

$$\bar{r}_n \ln(1/\epsilon_n^2) \prec n\epsilon_n^2, \tag{2.8}$$

$$\bar{r}_n \ln(p_n) \prec n\epsilon_n^2, \tag{2.9}$$

$$\bar{r}_n \ln(n\epsilon_n^2 \tilde{B}_n) \prec n\epsilon_n^2, \tag{2.10}$$

$$1 \leq r_n \leq \bar{r}_n < p_n, \tag{2.11}$$

$$r_n \prec p_n, \tag{2.12}$$

$$B(r_n) \prec n\epsilon_n^2, \tag{2.13}$$

then the posterior consistency holds, i.e., there exists a constant $c_1 > 0$ *such that*

$$P^*\{\pi[d(f, f^*) > \epsilon_n | D_n] > e^{-c_1 n\epsilon_n^2}\} < e^{-c_1 n\epsilon_n^2}, \tag{2.14}$$

for sufficiently large n*, where the Hellinger distance is*

$$d^2(f, f^*) = \int\int |\phi^{1/2}(y; x^T\boldsymbol{\beta}^*, \sigma^*) - \phi^{1/2}(y; x^T\boldsymbol{\beta}, \sigma)|^2 \nu_x(dx)dy,$$

where f^* *is as defined in (2.1),* f *is the density function for a model simulated from the posterior, and* P^* *denotes the probability measure of data generation.*

The proof of Theorem 2.1.1 can be found in the Appendix. The validity of this theorem depends on the choice of the prior covariance matrix V_ξ, or more precisely, the largest eigenvalues of V_ξ and V_ξ^{-1}. For most choices of V_ξ, $B(r_n) \leq Br_n^v$ and $\tilde{B}_n \leq Br_n^v$ hold for some positive constants B and v. For example, if $V_\xi = cI_{|\xi|}$, where I_p is an identity matrix of order p, then both $B(r_n)$ and $\tilde{B}_n$ are constants. One can also set $V_\xi^{-1} = cE(x_\xi^T x_\xi)$ as in [68] and [51], where the expectation is

replaced by its sample empirical estimator. If all x_i's are standardized with zero mean and equal variance, then $B(r_n) \leq r_n$. Furthermore, if x_i has some specific covariance structure, then $\tilde{B}_n$ can be shown to be bounded. For example, if all pairwise covariance $cov(x_i, x_j) = \rho \in (0, 1)$, then $\tilde{B}_n = 1 - \rho$; if x_i is generated from a finite order autoregressive (AR) or moving average (MA) model without root for the characteristic polynomial on the unit circle, then $\tilde{B}_n$ is also bounded by some constant [7]. In general, for standardized x_i's, we can choose a "ridge prior" as suggested by [31]: Let $V_\xi^{-1} = E(x_\xi^T x_\xi) + c_n I_{|\xi|}$, then $B(r_n) \leq r_n + c_n$ and $\tilde{B}(n) \leq c_n^{-1}$, where c_n is proportional to $1/n$, i.e., $c_n \propto n^{-1}$.

Theorem 2.1.1 assumes that all the predictors are uniformly bounded. However, this condition might not be satisfied for continuous predictors. In what follows, we relax this condition:

(A_1) There exist some constants $M > 0$ and $\delta > 0$ such that for any subvector of x, $(x_{j_1}, ..., x_{j_r})^T$ with $1 \leq j_1 < \cdots < j_r \leq p_n$, $r \leq \bar{r}_n$, and $|a_i| \leq \delta$ for $i = 1, \ldots, r$, we have

$$E \exp\left\{ \left(\sum_{i=1}^{r} a_i x_{j_i} / r \right)^2 \right\} \leq \exp(M),$$

where the expectation is taken with respect to the distribution of the random variables x_{j_i}'s.

The condition (A_1) controls the tail distribution of x such that the marginal distribution of y does not change dramatically in the sense of d_1 divergence (defined in the Appendix) with respect to the change of $\boldsymbol{\beta}$. If x has a heavy tail distribution, the prior information around f^* will be too weak to induce good posterior asymptotics. If all the variables are bounded by some constant m, as assumed in Theorem 2.1.1, then $E \exp\left\{ (\sum_{i=1}^{r} a_i x_{j_i} / r)^2 \right\} \leq \exp(\delta^2 m)$ and thus (A_1) holds. If x follows a Gaussian process with zero mean and $\mathrm{Var}(x_i) = 1$, then $(\sum_{i=1}^{r} a_i x_{j_i})^2$ follows the $s\chi_1^2$

distribution, where the scale $s \leq \delta^2 r^2$. Therefore, by the moment generating function of χ_1^2 distribution, $E \exp\left\{\left(\sum_{i=1}^r a_i x_{j_i}\right)^2 / r^2\right\} \leq (1-2\delta^2)^{-1/2}$, condition (A_1) holds.

Under the condition (A_1), we have the following theorem whose proof can be found in the Appendix.

Theorem 2.1.2. *Assume that the data set D_n is generated from model (1.1) and all the predictors satisfy condition (A_1). Given a sequence $\{\epsilon_n\}$, $\epsilon_n \to 0$, and $n\epsilon_n^2 \to \infty$. If the conditions (2.8) to (2.13) of Theorem 2.1.1 hold, the model is strictly sparse, i.e. condition (2.3) holds, and $r_n \succ 1$, then the posterior consistency still holds, i.e., there exists a constant $c_1 > 0$ such that for sufficiently large n,*

$$P^*\{\pi[d(f, f^*) > \epsilon_n | D_n] > e^{-c_1 n\epsilon_n^2}\} < e^{-c_1 n\epsilon_n^2}.$$

Note that condition (2.7) of Theorem 2.1.1 is redundant here, as for large n, $r_n > |\boldsymbol{t}|$ and $\Delta(r_n) = 0$ hold under the sparseness condition (2.3). The next corollary concerns the convergence rate of the posterior distribution.

Corollary 2.1.1. *Consider a strict sparse model, i.e. condition (2.3) holds. Suppose that $p_n < e^{Cn^\alpha}$ for some $\alpha \in (0,1)$, and the prior specification in section 2.1.1.1 is used such that $B(r_n) \leq Br_n^v$ and $\tilde{B}_n \leq Bn^v$ for some positive constants B and v. Take*

$$r_n < \bar{r}_n \prec \log^k(n), \ \ or \quad r_n < \bar{r}_n \prec n^q,$$

for some $k > 0$ or some $0 < q < \min\{1-\alpha, 1/v\}$. Then the convergence rate in Theorem 2.1.2 can be taken as

$$\epsilon_n = O(n^{-(1-\alpha)/2} \log^{k/2} n), \ \ or \quad \epsilon_n = O(n^{-(\min\{1-\alpha-q, 1-vq\})/2}).$$

Remark for condition A_1:

The condition A_1 imposes a stringent requirement for the distribution of the predictors, although in real world, one may treat any predictor as bounded in practice. If any of the predictor is not heavy tailed distributed, condition A_1 fails. A much weaker condition is proposed in below:

(A_2) There exist some constants $M > 0$ and $\delta > 0$ such that for any subvector of x, $(x_{j_1}, ..., x_{j_r})^T$ with $1 \leq j_1 < \cdots < j_r \leq p_n$, $r \leq \bar{r}_n$, and $|a_i| \leq \delta$ for $i = 1, \ldots, r$, we have

$$E\left(\sum_{i=1}^{r} a_i x_{j_i}/r\right)^2 \leq M.$$

(A_2) only imposes condition on the second moment structure of the predictors, and if all predictors have uniformly bounded second moment, condition (A_2) holds. It can be showed (in the Appendix), if condition (A_2) holds instead of (A_1), a slightly weaker result of theorem 2.1.2 can be still obtained as below,

$$\lim_{n\to\infty} P^*\{\pi[d(f, f^*) > \epsilon_n | D_n] > e^{-c_1 n\epsilon_n^2}\} = 0. \tag{2.15}$$

In the appendix, it is showed that under condition (A_1), $\pi[d(f, f^*) > \epsilon_n | D_n]$ converges in L_1 exponentially, while (2.15) only impliesconvergence in probability.

2.1.2 Variable Screening and Selection for Correctly Specified Models

To serve the purpose of variable selection for the case that the true model $\boldsymbol{t}$ exists, i.e. condition (2.3) holds, we need a consistent variable selection procedure. In this section, we consider two procedures, which are based on the marginal inclusion probability and maximum a *posteriori* (MAP) model, respectively.

2.1.2.1 Marginal Inclusion Probability

Variable selection by marginal inclusion probability has been used in high dimensional Bayesian analysis, see e.g., [5] and [52]. In this section, we study the property of marginal inclusion probability based on the posterior convergence result (2.14) or (2.15), which implies that $\pi[d(f, f^*) > \epsilon_n | D_n] \xrightarrow{p} 0$. Since our objective is to recover all the true predictors, we need to explore the relationship between the distance in distributions and the difference in models. Let q_j denote the marginal inclusion probability of the predictor $\boldsymbol{x}_j$, which is given by

$$q_j = \pi(j \in \xi | D_n) = \sum_{\xi \in \{\tilde{\xi}: j \in \tilde{\xi}\}} \pi(\xi | D_n).$$

Intuitively, we expect that a true predictor $\boldsymbol{x}_t$ ($t \in \boldsymbol{t}$) will have a high marginal inclusion probability. However, under high dimensional scenario, spurious multicollinearity is quite common phenomenon. [66] studied the relationship between low-rankness of design matrix and consistency of the pairwise model Bayes factors and global posterior probabilities given null model is true. If a true predictor can be well-approximated or even replaced by a couple of spurious predictors, hence inclusion probabilities of spurious predictors are inflated. Therefore, in order to consistently select the true variables, it is necessary to control the severeness of multicollinearity. To study the asymptotics of the marginal inclusion probability, we introduce the following identifiability condition:

(B_1) A predictor $\boldsymbol{x}_k$ is said to be identifiable among all other predictors, if, for any $1 \leq j_1, \ldots, j_{\bar{r}_n} \leq p_n$ ($j_i \neq k$ for all i) and $b_i \in \mathbb{R}$,

$$E \exp\left\{-(x_k + \sum_{i=1}^{\bar{r}_n} b_i x_{j_i})^2\right\} \leq 1 - \delta_n, \text{ and } \delta_n \succ \epsilon_n^2,$$

where x_l denotes a generic observation of the predictor $\boldsymbol{x}_l$.

This condition states that if a predictor $\boldsymbol{x}_k$ is identifiable, then there does not exist a linear combination of other predictors in $\boldsymbol{X}$ which can mimic it. The severeness of collinearity among predictors is controlled by sequence ϵ_n, and $\boldsymbol{x}_k$ can be distinguished from other predictors in $\boldsymbol{X}$.

If all the predictors have mean 0 and are bounded with $|x_i| \leq 1$, then

$$\begin{aligned} E\exp\left\{-(x_k+\sum_{i=1}^{\bar{r}_n} b_i x_{j_i})^2\right\} &\leq E\left(1-\frac{1-\exp\{-(1+\sum_i |b_i|)^2\}}{(1+\sum_i |b_i|)^2}(x_k+\sum_i b_i x_{j_i})^2\right) \\ &\leq 1-(1-e^{-1})\frac{\mathrm{Var}(x_k+\sum_i b_i x_{j_i})}{(1+\sum |b_i|)^2} \\ &\leq 1-\frac{1-e^{-1}}{(\bar{r}_n+1)} ch_1'(\mathrm{Var}(\tilde{x})), \end{aligned}$$

where $\tilde{x} = (x_k, x_{j_1}, \ldots, x_{j_{\bar{r}_n}})^T$, and $ch_1'(\mathrm{Var}(\tilde{x}))$ is the smallest eigenvalue of the covariance matrix of $\tilde{x}$. Hence, if $ch_1'(\mathrm{Var}(\tilde{x})) \succ \bar{r}_n \epsilon_n^2$ for any choice of $\tilde{x}$, x_t is identifiable. Furthermore, if $ch_1'(\mathrm{Var}(x)) \succ \bar{r}_n \epsilon_n^2$, then all the predictors are identifiable, here $x = (x_1, \ldots, x_{p_n})^T$ denotes a generic observation of all predictors.

If all the predictors follow the standard normal distribution, by the moment generating function of the chi-square distribution, we have

$$\begin{aligned} E\exp\{-(x_k+\sum b_i x_{j_i})^2\} &= (1+2\mathrm{Var}(x_k+\sum_i b_i x_{j_i}))^{-1/2} \\ &< \max\{0.5, 1-\mathrm{Var}(x_k+\sum_i b_i x_{j_i}))/8\}. \end{aligned}$$

Thus, if $\mathrm{Var}(x_k+\sum_i b_i x_{j_i}) \succ \epsilon_n^2$, then the identifiability condition is satisfied. Since

$$\mathrm{Var}(x_k+\sum_i b_i x_{j_i}) \geq \mathrm{Var}(x_k|\tilde{x}\backslash x_k) \geq ch_1'(\mathrm{Var}(\tilde{x})) \geq ch_1'(\mathrm{Var}(x)),$$

by the same arguments, all predictors are identifiable if $ch_1'(\mathrm{Var}(x)) \succ \epsilon_n^2$. Further, it can be shown (in the proof of Theorem 2.1.3) that if a true predictor $\boldsymbol{x}_t$ is identifiable, then $d(f^*, f) \geq \epsilon_n$ for any density f that does not select $\boldsymbol{x}_t$. Therefore, we have the following sure screening property.

Theorem 2.1.3 (Sure Screening)**.** *Assume that all the conditions of Theorem 2.1.2 hold. If a true predictor $\boldsymbol{x}_t$ ($t \in \boldsymbol{t}$) is identifiable, then $P^*\{\pi[t \in \xi|D_n] < 1 - e^{-c_1 n\epsilon_n^2}\} < e^{-c_1 n\epsilon_n^2}$, where ξ denotes a model sampled from the posterior distribution. Furthermore if condition (A_2) holds instead of (A_1), then weaker convergence holds:* $\lim_n P^*\{\pi[t \in \xi|D_n] < 1 - e^{-c_1 n\epsilon_n^2}\} = 0$.

The proof of Theorem 2.1.3 can be found in the Appendix. Let $\xi_q = \{i : 1 \leq i \leq p_n, q_i > q\}$ denote a model including all the predictors with the marginal inclusion probability greater than a threshold value of $q \in (0, 1)$. Then it follows from Theorem 2.1.3 that $\pi(\boldsymbol{t} \subset \xi_q) \xrightarrow{p} 1$ if all true predictors are identifiable. To determine the threshold value q, we adopt the multiple hypothesis testing procedure proposed in [52], where the procedure is used for selecting variables for a logistic regression model. The procedure can be briefly described as follows.

Let $z_i = \Phi^{-1}(q_i)$ denote the marginal inclusion score (MIS) of the predictor $\boldsymbol{x}_i$, where $\Phi(\cdot)$ is the CDF of the standard normal distribution. To select predictors with large MISs, we model the MISs by a two-component mixture exponential power distribution by

$$g(z|\vartheta) = \omega\varphi(z|\mu_1, \sigma_1, \alpha_1) + (1 - \omega)\varphi(z|\mu_2, \sigma_2, \alpha_2), \tag{2.16}$$

where $\vartheta = (\omega, \mu_1, \mu_2, \sigma_1, \sigma_2, \alpha_1, \alpha_2)^T$ is the vector of parameters of the distribution, and

$$\varphi(z|\mu_i, \sigma_i, \alpha_i) = \frac{\alpha_i}{2\sigma_i \Gamma(1/\alpha_i)} \exp\{-(|z-\mu_i|/\sigma_i)^{\alpha_i}\}, \\ -\infty < \mu_1 < \mu_2 < \infty,\ \sigma_i > 0,\ \alpha_i > 1, \tag{2.17}$$

where μ_i, σ_i, α_i are the location parameter, dispersion parameter and decay rate of the distribution, respectively. If $\alpha = 2$, then the exponential power distribution is reduced to a normal distribution.

The parameters of (2.16) can be estimated as in [53] by minimizing $KL(\vartheta)$ using the stochastic approximation algorithm, where $KL(\vartheta)$ is the Kullback-Leibler divergence between $g(z|\vartheta)$ and the unknown true distribution $g(z)$:

$$KL(\vartheta) = -\int \log(g(z|\vartheta)/g(z))g(z)dz.$$

For a given selection rule $\Lambda = \{Z_i > z_0\}$, the false discovery rate (FDR) of true predictors can be estimated by

$$FDR(\Lambda) = \frac{p_n \omega[1 - F(z_0|\hat{\mu}_1, \hat{\sigma}_1, \hat{\alpha}_1)]}{\#\{z_i : z_i > z_0\}},$$

where $F(\cdot)$ denote the CDF of the distribution (2.17), and $\#\{\cdot\}$ denotes the number of elements in a set. Similar to [64], we define the q-value as

$$Q(z) = \inf_{\{\lambda : z \in \Lambda\}} FDR(\Lambda). \tag{2.18}$$

Then a cutoff value of z_0, which corresponds to the marginal inclusion probability $\Phi(z_0)$, for the MIS can be determined according to a pre-specified FDR level α, e.g.,

$\alpha = 0.01$ and 0.05. That is, all the predictors belonging to the set $\{\boldsymbol{x}_i : Q(\Phi^{-1}(q_i)) \leq \alpha\}$ will be selected as "true" predictors.

Compared to other FDR control procedures, one advantage of this procedure is that it is applicable under general dependence between marginal inclusion probabilities. Refer to [53] for the details of the procedure and its generalization to the m-component case.

2.1.2.2 MAP model

Many Bayesian studies suggest that the MAP model is potentially a good estimator of the true model, see e.g., [23] and [30]. In what follows, we establish the consistency of the MAP model with two different choices of V_ξ: $V_\xi^{-1} = I$ and $V_\xi^{-1} = \boldsymbol{X}_\xi^T \boldsymbol{X}_\xi / n$.

Theorem 2.1.4. *(Consistency of the MAP model) Let D_n be a data set generated from model (1.1) with a sparse true model $\boldsymbol{t}$. Assume condition (2.11) and the following eigen-structure conditions hold: For any model ζ with size $|\zeta| = \min\{|\boldsymbol{t}| + \bar{r}_n, n\}$, there exist a non-increasing sequence $\{l_n\}$ and a non-decreasing sequence $\{l'_n\}$ such that*

$$nl_n < ch'_1(\boldsymbol{X}_\zeta^T \boldsymbol{X}_\zeta) \leq ch_1(\boldsymbol{X}_\zeta^T \boldsymbol{X}_\zeta) < nl'_n, \tag{2.19}$$

$$\bar{r}_n \log p_n \prec nl_n. \tag{2.20}$$

If, further, one of the following two condition holds:

1. *$nl_n \succ \log l'_n$, $\log \frac{\sqrt{nl_n}}{r_n} \succ \sqrt{\log p_n}$, and we set $V_\xi^{-1} = I$ for any ξ; or*
2. *$nl_n \succ l'_n$, $\log \frac{\sqrt{n}}{r_n} \succ \sqrt{\log p_n}$, and we set $V_\xi^{-1} = \boldsymbol{X}_\xi^T \boldsymbol{X}_\xi / n$ for any ξ,*

then, for any model $\boldsymbol{k}$*, the posterior probability (2.6) satisfies:*

$$\max_{\boldsymbol{k}\neq\boldsymbol{t}} \frac{\pi(\boldsymbol{k}|D_n)}{\pi(\boldsymbol{t}|D_n)} \xrightarrow{p} 0. \tag{2.21}$$

This theorem implies that the true model $\boldsymbol{t}$ can be identified asymptotically by just finding out the MAP model under (2.6). Regarding the eigen structure condition (2.19) and the choice of V_ξ, we have the following remarks:

- It is worth noting that the eigenvalue condition (2.19) is hardly satisfied, although similar conditions have been used in the literature, see e.g., [37]. A feasible condition can be as follows: There exist two sequences $\{l_n\}$ and $\{l'_n\}$ and a constant C such that for any model $|\zeta| = \bar{r}_n + |\boldsymbol{t}|$,

$$Pr\left\{nl_n > ch'_1(\boldsymbol{X}_\zeta^T\boldsymbol{X}_\zeta) \text{ or } ch_1(\boldsymbol{X}_\zeta^T\boldsymbol{X}_\zeta) > nl'_n\right\} < e^{-nC}. \tag{2.22}$$

  Since the condition $\bar{r}_n \ln p_n \prec n$ holds, (2.22) implies that, for sufficiently large n,

$$Pr\left\{\exists\ |\zeta| = \bar{r}_n + |\boldsymbol{t}|,\ nl > ch'_1(\boldsymbol{X}_\zeta^T\boldsymbol{X}_\zeta) \text{ or } ch_1(\boldsymbol{X}_\zeta^T\boldsymbol{X}_\zeta) > nl'\right\} < e^{-nC/2}.$$

  Hence, if (2.19) is replaced by (2.22), Theorem 2.1.4 still holds. The condition (2.22) does not require the eigenvalues of the sample covariance matrix to be bounded, but restricts the tail distribution of eigenvalues. Hence, it is more reasonable and acceptable.

- The condition (2.22) is similar to but somewhat weaker than the concentration condition used in [19]. The concentration condition constrains the eigenstructure of any $n \times \tilde{p}$ sub-datamatrix for any $\tilde{p} > cn$ for some constant c, while

the condition (2.22) only constrains the eigen-structure of $n \times (\bar{r}_n + |\boldsymbol{t}|)$ sub-datamatrix. In particular, condition (2.22) holds when x follows a Gaussian process, and the eigenvalues of $\mathrm{Var}(x)$ are bounded by l_n and l'_n (see Appendix A.7 of [19]). For a further study of the limit behavior of the spectral structure, see e.g. [63], [4], and [46].

- Choosing V_ξ^{-1} as a function of the sample covariance matrix makes use of the data information in the prior, and as a result, a slightly weaker condition is required for the consistency of variable selection. It follows from the theory of least square regression, a natural choice of V_ξ^{-1} is $V_\xi^{-1} = \boldsymbol{X}_\xi^T \boldsymbol{X}_\xi / n$. However, for binary data, even if $n > |\xi|$, one may still encounter the problem of singularity of $\boldsymbol{X}_\xi^T \boldsymbol{X}_\xi$. To address this issue, we set $V_\xi^{-1} = (\boldsymbol{X}_\xi^T \boldsymbol{X}_\xi + \tau I)/n$ for some small τ in our simulations. It can be shown that this choice leads to the same consistency result as with the sample covariance matrix (see Lemma A.2 of the Appendix). This choice also automatically meets the eigenvalue condition of V_ξ in Corollary 2.1.1 for a standardized dataset.

Corollary 2.1.2. *Consider a linear regression with a sparse true model* $\boldsymbol{t}$*. Assume that* $p_n < n^{\kappa \log n}$ *for some* $\kappa \in (0, 1/4)$*, and the prior in Section 2.1.1.1 is used with* $V_\xi^{-1} = \boldsymbol{X}_\xi^T \boldsymbol{X}_\xi / n$ *or* $V_\xi^{-1} = \boldsymbol{X}_\xi^T \boldsymbol{X}_\xi / n + \tau I / n$*. Let*

$$\bar{r}_n \prec \log^{k_1}(n), \ \textit{or} \quad \bar{r}_n \prec n^q,$$

for some $k_1 > 0$ *or* $q \leq 1/2 - \sqrt{\kappa}$*, and*

$$l'_n = l_n^{-1} = \sqrt{n} / \log^{k_2}(n),$$

for some $k_2 > 0$*. Then the consistency of the MAP model holds.*

Theorem 2.1.3 and Theorem 2.1.4 provide two strategies for variable selection, either by marginal inclusion probability or by the MAP model. Theorem 2.1.4 states that the MAP model is consistent, while Theorem 2.1.3 states only the sure screening property (sensitivity is asymptotically 1, but specificity is not controlled). In this work, we suggest to use marginal inclusion probability as the criterion of variable selection. Under this criterion, one may select more predictors, which are worthy of further investigation through a possibly different and expensive experiment. Since the sample size n can be small for a real problem, the MAP model might not capture all true predictors although it is consistent in theory.

2.1.3 Variable Screening for Misspecified Models

Let $D_n^{\boldsymbol{s}} = \{\boldsymbol{y}, \boldsymbol{X}_{\boldsymbol{s}}\}$ denote a fixed subset of observations, where $\boldsymbol{s} \subset \{1, 2, \ldots, p_n\}$ and $\boldsymbol{X}_{\boldsymbol{s}}$ contains only $s = |\boldsymbol{s}| < p_n$ predictors with the indices belonging to $\boldsymbol{s}$. Assume that $\boldsymbol{X}_{\boldsymbol{s}}$ does not include all of the true predictors of model (1.1). Therefore, the model

$$\boldsymbol{y} = \boldsymbol{X}_{\boldsymbol{s}}\boldsymbol{\beta}_{\boldsymbol{s}} + \boldsymbol{\varepsilon}$$

is misspecified. Under the Bayesian framework, some work for misspecified models have been done by [6], [62], [40], [15], among others. Let $\mathcal{P}_s$ denote a set of parameterized densities for all possible models formed with $\boldsymbol{X}_{\boldsymbol{s}}$. It is obvious that $f^* \notin \mathcal{P}_s$. Let f_0 denote the minimization point of the Kullback-Leibler divergence in $\mathcal{P}_s$, i.e.

$$f_0 = \arg\min_{f \in \mathcal{P}_s} \int \ln(f^*/f) f^*. \tag{2.23}$$

With a slight abuse of notation, we use, as in previous sections, ξ to denote a model selected among the predictors in $\boldsymbol{X}_{\boldsymbol{s}}$. Further, the same prior distributions

are specified for the parameters σ^2 and $\boldsymbol{\beta}_\xi$; that is,

$$\pi(\sigma^2) \sim IG(a_0, b_0), \qquad \pi(\boldsymbol{\beta}_\xi|\xi, \sigma^2) \sim N(0, \sigma^2 V_\xi).$$

Under these priors, we show in Theorem 2.1.5 that the density f_0 can be consistently estimated by the models sampled from the posterior $\pi(\xi|D_n^{\boldsymbol{s}})$ as $n \to \infty$. Note that to prove Theorem 2.1.5, $\pi(\xi)$ does not need to be specified, which is only required to be positive for all possible models, i.e., $\pi(\xi) > 0$ for all ξ.

Parallel to condition (A_1), we introduce the following condition regarding the range of x_i's.

(A_1') There exists a constant $\delta > 0$ such that for any $\boldsymbol{a} = (a_1, \ldots, a_s)^T \in \mathbb{R}^s$, with $|a_i| \leq \delta$ for $i = 1, \ldots, s$, we have

$$E[(\boldsymbol{a}^T x_{\boldsymbol{s}})^2] < \infty,$$

where $x_{\boldsymbol{s}}$ denote a generic observation of the predictors included in $\boldsymbol{X}_{\boldsymbol{s}}$.

Theorem 2.1.5. *(Posterior consistency for misspecified models) Assume the condition A_1' holds for a given subset $\boldsymbol{s}$. Under the prior setting as described above, for any $\epsilon > 0$,*

$$\pi(\{f \in \mathcal{P}_s : d(f_0, f) > \epsilon\}|D_n^{\boldsymbol{s}}) \to 0, \qquad a.s. \tag{2.24}$$

as $n \to \infty$, where f_0 is as defined in (2.23), and f is a parameterized density proposed from posterior.

It follows from Equation (A.7) of the Appendix that $f_0 = \phi(y; \boldsymbol{\beta}_0^T x_{\boldsymbol{s}}, \sigma_0)\nu(x_{\boldsymbol{s}})$,

where $\nu(x_{\boldsymbol{s}})$ denotes the probability measure of $x_{\boldsymbol{s}}$, and $\boldsymbol{\beta}_0$ is given by

$$\boldsymbol{\beta}_0 = \arg\min_{\boldsymbol{\beta}_{\boldsymbol{s}}} E(\boldsymbol{\beta}_{\boldsymbol{s}}^T x_{\boldsymbol{s}} - \boldsymbol{\beta}^{*T} x)^2,$$

i.e., $\boldsymbol{\beta}_0^T x_{\boldsymbol{s}}$ is the projection of $\boldsymbol{\beta}^{*T} x$. Let $\tilde{\boldsymbol{s}}$ denote the subset of $\boldsymbol{s}$ corresponding to nonzero entries of $\boldsymbol{\beta}_0$. Following the proof of Theorem 2.1.3, we have the following corollary:

Corollary 2.1.3. *Assume the conditions of Theorem 2.1.5 hold. If $x_{\boldsymbol{s}}$ does not have exact multicollinearity between variables, i.e., there does not exist a nonzero vector $\boldsymbol{a} \in \mathbb{R}^s$ such that $P(\boldsymbol{a}^T x_{\boldsymbol{s}} = 0) = 1$, then for the posterior probability of model ξ, conditioned on the subset $D_n^{\boldsymbol{s}}$, we have*

$$\pi(\tilde{\boldsymbol{s}} \subset \xi | D_n^{\boldsymbol{s}}) \xrightarrow{p} 1.$$

Corollary 2.1.3 implies that the marginal inclusion probability criterion, described in Section 2.1.2.1, is still applicable for selection of predictors in $\boldsymbol{X}_{\boldsymbol{s}}$. For any true predictor $\boldsymbol{x}_t$ with $t \in \boldsymbol{t} \cap \boldsymbol{s}$, it will be asymptotically selected under this criterion if and only if $\beta_{0,t} \neq 0$, where $\beta_{0,t}$ is the entry of $\boldsymbol{\beta}_0$ corresponding to $\boldsymbol{x}_t$. Let y denote a generic observation of the response variable, let $x_{\boldsymbol{t}}$ denote a generic observation of the true predictors, and let $\boldsymbol{\beta}_{\boldsymbol{t}}^*$ denote the regression coefficient vector of the true predictors. Since

$$\boldsymbol{\beta}_0 = \Sigma_{\boldsymbol{s}}^{-1} \Sigma_{\boldsymbol{s},\boldsymbol{t}} \boldsymbol{\beta}_{\boldsymbol{t}}^* = \Sigma_{\boldsymbol{s}}^{-1} \mathrm{Cov}(x_{\boldsymbol{s}}, y), \tag{2.25}$$

where $\Sigma_{\boldsymbol{s}} = \mathrm{Var}(x_{\boldsymbol{s}})$ and $\Sigma_{\boldsymbol{s},\boldsymbol{t}} = \mathrm{Cov}(x_{\boldsymbol{s}}, x_{\boldsymbol{t}})$, $\tilde{\boldsymbol{s}}$ is determined by the correlation structure of $x_{\boldsymbol{s}}$, $x_{\boldsymbol{t}}$ and the value of true regression coefficients. When $\boldsymbol{t} \nsubseteq \boldsymbol{s}$, it is unclear whether $\beta_{0,t}$ is exactly zero or not. However, from equation (2.25), if $\mathrm{Cov}(x_{\boldsymbol{s}}, y) \neq 0$, then $\boldsymbol{\beta}_0 \neq 0$ and thus at least one predictor in $\boldsymbol{s}$ will be selected ac-

cording to the marginal inclusion probability criterion. This motivates us to propose an iterative variable selection method for the subset data, provided the following condition holds:

(D_1) $|\mathrm{Cov}(x_t, y)| > 0$ for any true predictor $\boldsymbol{x}_t$.

The iterative variable selection method can be described as follows: First select predictors from $\boldsymbol{X}_{\boldsymbol{S}}$ based on the marginal inclusion probability estimated from the posterior $\pi(\xi|D_n^{\boldsymbol{S}})$; denote the set of unselected predictors by $\boldsymbol{X}_{\boldsymbol{S}'}$, and then select predictors again from $\boldsymbol{X}_{\boldsymbol{S}'}$ based on the marginal inclusion probability estimated from the posterior $\pi(\xi|\boldsymbol{y}, \boldsymbol{X}_{\boldsymbol{S}'})$; and keep running the iterative procedure until no predictors can be selected. Denote all the selected predictors by $\boldsymbol{X}_{\tilde{\boldsymbol{S}}}$. Under the assumption (D_1), this procedure ensures that all true predictors contained in $\boldsymbol{X}_{\boldsymbol{S}}$ will be selected.

There are a few remarks on this iterative procedure:

- Other than true predictors, how many other predictors will be selected? Generally speaking, all the predictors that are correlated with $\boldsymbol{X}_{\boldsymbol{t}}$ will be selected, because they are correlated with $\boldsymbol{y}$, where $\boldsymbol{X}_t$ denotes the set of all true predictors in D_n. This means if all predictors are linearly correlated, it is futile to apply this iterative procedure, as all predictors will be selected. The rational underlying the split-and-merge strategy lies on the belief that for a big data set with millions or more of predictors, there are only a small proportion of predictors linearly correlated with $\boldsymbol{X}_{\boldsymbol{t}}$. It is easy show that, if there are p_n predictors correlated with the true predictor $\boldsymbol{X}_{\boldsymbol{t}}$, $p_n \prec p_n$, then it can be shown that there are at most cp_n predictors to be selected into $\boldsymbol{X}_{\tilde{\boldsymbol{S}}}$ for some constant c.

- To ensure all true predictors in $\boldsymbol{X}_{\boldsymbol{s}}$ to be selected, a liberal way, as what we proposed, is to iteratively repeat the selection until no predictors can be selected. In practice, it is rare to have the case that $\beta_{0,t}$ is 0 exactly. Hence, the iterative procedure is not always necessary, and one may set an upper bound for the iteration number.

- Theorem 2.1.5 requires the prior probability $\pi(\xi) > 0$ for all possible models. Hence, the independent prior $\pi(\xi) = \lambda_{\boldsymbol{s}}^{|\xi|}(1-\lambda_{\boldsymbol{s}})^{|\boldsymbol{s}|-|\xi|}$ for some $\lambda_{\boldsymbol{s}} \in (0,1)$ is still applicable here. When $|\boldsymbol{s}|$ is large, we suggest to impose an upper bound for the model size to avoid expansive computation caused by inverting high order matrices. Then the posterior should asymptotically concentrate on the models with the density function minimizing the K-L divergence under the restriction of model size. In this case, although Corollary 2.1.3 does not hold any more, the marginal inclusion probabilities for correlated and uncorrelated predictors are distinct and the sure screening property of the marginal inclusion probability criterion should still hold.

- For data splitting, one extreme choice is $s = 1$. In this case, the SaM approach will perform like the SIS algorithm [19] with only the marginal utility being used in variable screening. The SaM approach generally does not perform well under this setting. The authors' numerical experience suggests to set s around 500 or larger. On one hand, this allows the joint information of predictors to be used in variable screening. On the other hand, a relatively large value of s improves the accuracy of the marginal inclusion probability-based variable screening procedure, which, as described in Section 2.1.2.1, is multiple hypothesis test-based.

2.2 SaM Approach and Its Implementation

2.2.1 SaM Approach

Based on the priors given in Section 2.1, the SaM approach can be summarized as follows:

1. Split the p_n predictors into K_n groups, $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{K_n}$, with $\max_{i=1,\ldots,K_n} |\boldsymbol{s}_i| \leq s$ for a pre-specified value of s.

2. **Stage I:** Apply the iterative procedure, proposed in Section 2.1.3, to select predictors from each of the sub-datasets $D_n^{\boldsymbol{s}_i} = \{\boldsymbol{y}, \boldsymbol{X}_{\boldsymbol{s}_i}\}$, $i = 1, \ldots, K_n$, at a FDR level of α_1. At each iteration, the Bayesian variable selection is subject to the prior setting: $\pi(\sigma) \sim \text{IG}(a_0, b_0)$, $\pi(\xi) = \lambda_{\boldsymbol{s}}^{|\xi|}(1-\lambda_{\boldsymbol{s}})^{|\boldsymbol{s}|-|\xi|} I(|\xi| \leq \bar{s})$, and $\pi(\boldsymbol{\beta}_\xi|\xi, \sigma) \sim \text{N}(0, \sigma^2(\boldsymbol{X}_\xi^T \boldsymbol{X}_\xi + \tau I)/n)$, where $\bar{s}$ denotes the upper bound of the size of the models considered for each subset data. Let $\tilde{\boldsymbol{s}}_1, \ldots, \tilde{\boldsymbol{s}}_{K_n}$ denote the sets of indices of the selected predictors from the K_n subsets, respectively.

3. **Stage II:** Merge the sets $\tilde{\boldsymbol{s}}_1, \ldots, \tilde{\boldsymbol{s}}_{K_n}$ into a single set $\tilde{\boldsymbol{S}} = \cup_{i=1}^{K_n} \tilde{\boldsymbol{s}}_i$, and define $p_n = |\tilde{\boldsymbol{S}}|$. Perform Bayesian variable selection on the subset data $D_n^{\tilde{\boldsymbol{S}}} = \{\boldsymbol{y}, \boldsymbol{X}_{\tilde{\boldsymbol{S}}}\}$ at a FDR level of α_2. The Bayesian variable selection is subject to the prior setting: $\pi(\sigma) \sim \text{IG}(a_0, b_0)$, $\pi(\xi) = (r_n/p_n)^{|\xi|}(1-r_n/p_n)^{p_n-|\xi|} I(|\xi| \leq \bar{r}_n)$, and $\pi(\boldsymbol{\beta}_\xi|\xi, \sigma) \sim \text{N}(0, \sigma^2(\boldsymbol{X}_\xi^T \boldsymbol{X}_\xi + \tau I)/n)$.

Clearly, it follows from the theory developed in Section 2.1 that the SaM approach will lead to a consistent selection of true predictors. In Step 3, the predictors can also be selected according to the MAP model.

2.2.2 Simulation and Hyperparameter Setting

Since different predictors can be highly correlated, the posterior (2.6) can have a rugged energy landscape for some problems. Conventional MCMC algorithms, such as the Metropolis-Hasting algorithm, tend to get trapped into a local mode in simulations. To avoid this problem, the stochastic approximation Monte Carlo (SAMC) algorithm [50] was used in the following numerical studies. SAMC belongs to the class of adaptive MCMC algorithms. However, unlike conventional adaptive MCMC algorithms, e.g., the adaptive Metropolis algorithm [32], which adapt the proposal distribution, SAMC adapts the invariant distribution at each iteration. As explained in [50], the self-adjusting mechanism of the invariant distribution enables SAMC to be immune to local trap problems. As shown in [48], SAMC is essentially a dynamic importance sampling algorithm, and quantities of interest can be estimated by weighted averaging its importance samples. Refer to the supplementary material of [52] for the implementation of the SAMC algorithm for regression.

Another issue related to the implementation of SaM is how to set hyperparameters. The prior distribution of SaM contains an important hyperparameter, namely, r_n. According to Theorem 2.1.2 and Theorem 2.1.4, r_n can be slowly growing with p_n for a good convergence rate of the posterior and the MAP model. In Section 2.1.1.1, we establish the connection between the posterior (2.6) and EBIC by choosing $r_n = p_n^{1-\gamma}$. [12] showed that a choice of $\gamma > 0.5$ is preferred for high dimensional regression. In [52], for a similar hyperparameter, the authors suggest the following rule:

$$\gamma = \inf \left\{ \tilde{\gamma} : \arg\max_{|\xi|} \pi(|\xi||D_n, \tilde{\gamma}) = |\xi_{MAP,\tilde{\gamma}}| \right\}, \tag{2.26}$$

that is, choosing the minimum value of γ such that the mode of $\pi(|\xi||D_n)$ coincides with the size of the MAP model ξ_{MAP}. If the resulting value of γ is greater than or

equal to 1, truncate it to a value less than 1, say 0.99. We can also apply this rule to determine the value of γ and set $r_n = p_n^{1-\gamma}$. In practice, one may try a sequence of γ values, and then choose the smallest one for which the mode of posterior of model size coincides with the size of the MAP model. Our experience shows that when n is not too small and $\log p_n/n$ is not too large, the choice of γ doesn't affect the result very much. In this work, we set $\gamma = 0.75$ and $r_n = p_n^{1-\gamma}$ unless otherwise stated. To make the prior for the subset data regression compatible with the prior for the whole data set regression, we set $\lambda_{\boldsymbol{s}} = |\boldsymbol{s}|^{-\gamma}$. For other hyperparameters, we set $a_0 = b_0 = 1$ and $\tau = 0.01$.

2.3 Simulated Examples

2.3.1 Toy Examples

2.3.1.0.1 Example 1 This example confirms the theoretical results established in Theorem 2.1.3 and Theorem 2.1.4. For this purpose, the predictors are directly selected without data splitting.

This example consists of multiple data sets with different values of n ranging from 20 to 120. For each value of n, we set $p_n = n^{1.5}$ and simulated 100 data sets independently. For each data set, the design matrix $\boldsymbol{X}$ was generated from a multivariate normal distribution. The variance of each column of $\boldsymbol{X}$ was set to be 1, and the correlation coefficient between different columns of $\boldsymbol{X}$ was set to be 0.25, which represents a strong correlation for real gene expression data (see e.g., [37]). For each data set, we set $\sigma^* = 1$ and chose the first 5 columns of $\boldsymbol{X}$ as the true predictors with the regression coefficients being 0.7, 0.9, 1.1, 1.3, 1.5, respectively.

For each data set, SAMC was run for $270 \times p_n$ iterations, where the first $20 \times p_n$ were for the burn-in process. In the simulations, we set the prior hyperparameters $r_n = \sqrt{n}$ and $\bar{r}_n = 2\sqrt{n}$, and set the gain factor sequence as in (2.27) with $k_0 =$

$10 \times p_n$:

$$a_k = \frac{k_0}{\max\{k_0, k\}}, \tag{2.27}$$

where k indexes the number of iterations of SAMC.

Figure 2.1 shows the distributions of marginal inclusion probabilities for the five true predictors and one false predictor. It is easy to see that as n increases, the marginal inclusion probabilities of the true predictors converge toward 1, while the marginal inclusion probability of the false predictor stays close to 0. This confirms the sure screening property of the marginal inclusion probability as established in Theorem 2.1.3. Figure 2.2 shows that the probability of the MAP model catching the true model increases with n. This confirms Theorem 2.1.4 that the MAP model is consistent. A comparison of Figure 2.1 and Figure 2.2 show that from the perspective of variable selection, the marginal inclusion probability may work better than the MAP model, as the former seems to converge faster than the latter.

A further simulation study is conducted here to study the stability of our approach when the multicollinearity is extremely strong. Instead of 0.25, we let the pairwise correlation to be 0.9, and let n increase from 30 to 125, whilst keep the other settings unchanged. Under such extremely high multicollinearity, the MAP model almost never catch the true model. However, as showed in figure 2.3, the trend of increasing marginal inclusion probabilities for the true predictor is still clear.

2.3.1.0.2 Example 2 This example illustrates the SaM approach. It consists of 100 simulated data sets, each consisting of $n = 150$ observations and $p_n = 1000$ predictors. For each data set, the design matrix $\boldsymbol{X}$ was generated from a multivariate normal distribution. The first 100 columns of $\boldsymbol{X}$ are mutually correlated with an equal correlation coefficient of 0.25, and independent of the rest 900 columns. The rest 900 columns are mutually independent. The first three columns were chose as

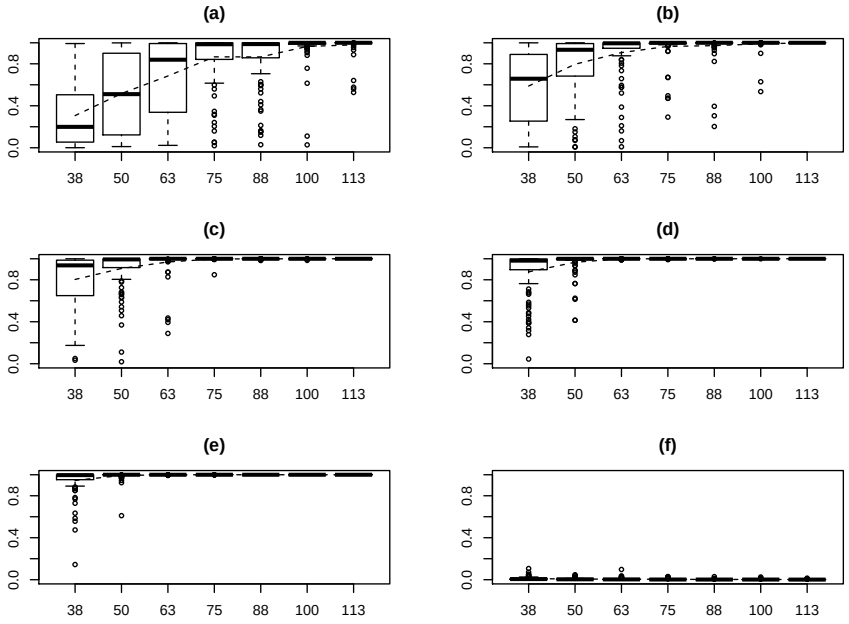

Figure 2.1: Simulation results for marginal inclusion probabilities. The six plots showed in this figure, (a)-(f), present the distributions of marginal inclusion probabilities of six predictors with the true regression coefficients 0.7, 0.9, 1.1, 1.3, 1.5 and 0, respectively. In each plot, the seven boxplots are for the sample size $n = 38, 50, 63, 75, 88, 100, 113$, respectively. Each boxplot shows the distribution of the marginal inclusion probabilities of one predictor calculated from 100 simulated data sets. The dashed line in each plot shows the mean value of the marginal inclusion probabilities.

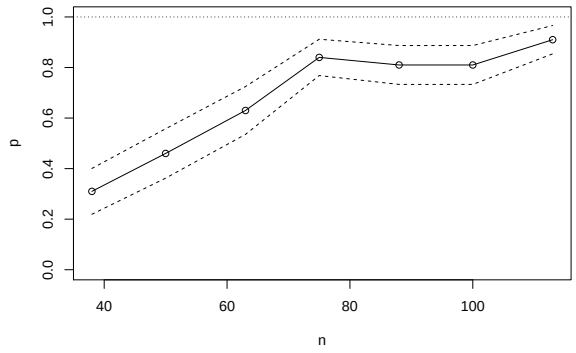

Figure 2.2: Simulation results for MAP model. The plot give the estimated probability that the MAP model coincides with the true model. For each value of n, the probability was estimated from 100 simulated data sets. The dashed lines show the 95% confidence interval of the probability.

the true predictors with the regression coefficients being 1.5, 3.0 and 4.5, respectively.

We randomly split each data set into 50 subsets with $s = 20$. In stage I, the predictors are iteratively selected twice for each subset. Each run of SAMC consists of $270 \times |\boldsymbol{s}|$ iterations, where the first $20 \times |\boldsymbol{s}|$ iterations were for the burn-in process, and $\boldsymbol{s}$ can be a subset directly split from the full data set or a remainder of a subset containing only unselected predictors. The gain factor sequence was set as $a_k = 10|\boldsymbol{s}|/\max\{10|\boldsymbol{s}|, k\}$, where k indexes the number of iterations of SAMC. The prior hyperparameter $\bar{s}$ was set to 20. The FDR level was set to $\alpha_1 = 0.15$, which is relatively large such that the variable selection is liberal and thus reducing the risk of losing important true predictors. In stage II, SAMC was run for the aggregated data set with $270 \times p_n$ iterations, where the first $20 \times p_n$ iterations were for the burn-in process. The gain factor sequence was set as $a_k = 10p_n/\max\{10p_n, k\}$, where k indexes the number of iterations of SAMC. The prior hyperparameter $\bar{r}_n$ was set to 35. The FDR level was set to $\alpha_2 = 0.01$, which is small such that the resulting model

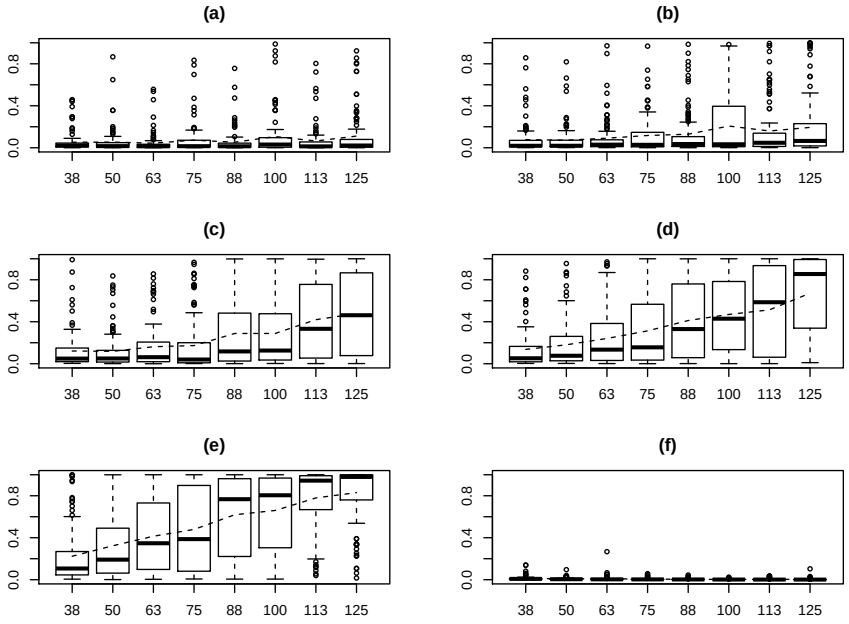

Figure 2.3: Simulation results for marginal inclusion probabilities under extremely high multicollinearity. The six plots showed in this figure, (a)-(f), present the distributions of marginal inclusion probabilities of six predictors with the true regression coefficients 0.7, 0.9, 1.1, 1.3, 1.5 and 0, respectively. In each plot, the seven boxplots are for the sample size $n = 38, 50, 63, 75, 88, 100, 113$ and 125, respectively. Each boxplot shows the distribution of the marginal inclusion probabilities of one predictor calculated from 100 simulated data sets. The dashed line in each plot shows the mean value of the marginal inclusion probabilities.

is sparse.

Stage I		Stage II	
size[a]	# of correlated predictors[b]	size[c]	# of true predictors[d]
166.6(.87)	98.2(.22)	3.2(.05)	3.0(0)

Table 2.1: Simulation results of SaM algorithm for the second toy example. [a] the average number of predictors selected in stage I; [b] the average number of predictors that are correlated with the true predictors and selected in stage I; [c] the average number of predictors selected in stage II; [d] the average number of true predictors selected in stage II. The values reported in parentheses are standard deviations of the corresponding estimate.

Table 2.1 summarizes the simulation results. It shows that almost all predictors that are correlated with the true predictors are selected in stage I, and all true predictors are selected in stage II. The SaM approach works perfectly for this example.

2.3.2 Massive Data Example

This example compares the SaM approach with several popular penalized likelihood approaches for massive data sets. We simulated 100 data sets with $\sigma^{*2} = 0.25$. Each data set was generated from a multivariate normal distribution with $n = 100$ observations and $p_n = 500,000$ predictors, among which 2,000 predictors are equally correlated with a correlation coefficient of $\rho = 0.25$ and the rest predictors are uncorrelated. Among those 2,000 correlated predictors, 5 of them were chosen as the true predictors with the regression coefficients (0.2, 0.3, 0.4, 0.5, 0.6).

Each data set was randomly split into 1,000 subsets with $s = 500$. SAMC was then run for each subset data and the aggregated subset data with the same setting as for Example 2 of Section 2.3.1. In simulations, we set $\bar{s} = 45$ as the upper bound of model size for the subset data and $\bar{r}_n$=35 for the aggregated subset data, and

set the FDR level $\alpha_1 = 0.15$ for variable selection in stage I and $\alpha_2 = 0.01$ in stage II. On average (over the 100 data sets), SaM reduced the dimension of the data to $18,057.7$ in stage I and then selected 5.3 predictors in stage II.

For comparison, several popular penalized likelihood approaches, including Lasso, elastic net, SIS and ISIS, were also applied to this example. Lasso was implemented for this example using the R package *glmnet* [24] with the tuning parameter λ determined through a 10-fold cross-validation procedure. Elastic net was also implemented using the R package *glmnet*, for which we set the penalty function $p_\lambda = \lambda(||\boldsymbol{\beta}||_{L_1} + ||\boldsymbol{\beta}||_{L_2}^2)$ and determined the value of λ via a 10-fold cross-validation procedure. We used the R package *SIS* to implement SIS-SCAD and ISIS-SCAD, which first reduce the number of predictors to $n/\log n$ by marginal utility, and then apply SCAD to refine the model. The results are summarized in Table 2.2.

Methods	SaM	Lasso	Elastic net	SIS	ISIS
$\overline{\lvert\xi\rvert}^a$	5.3(0.22)	44.62(0.98)	52.70(0.21)	11.47(0.29)	13.57(0.59)
$\overline{\lvert\xi \cap \boldsymbol{t}\rvert}^b$	3.53(0.073)	4.24(0.070)	4.07(0.074)	3.31(0.083)	3.37(0.086)
fsr(%)	25.5(2.30)	89.7(0.38)	92.3(0.15)	67.7(1.58)	65.8(2.50)
nsr(%)	29.4(1.46)	15.2(1.40)	18.6(1.48)	33.8(1.65)	32.6(1.72)
MSE^c	0.146(.013)^d	0.263(.011)	0.488(.011)	0.246(.016)	0.342(.022)

Table 2.2: Simulation results for half-million-predictor data sets. The table compares the SaM result with Lasso, elastic net, SIS and ISIS for the massive data example: ^{a} average number of predictors selected by different methods; ^{b} average number of true predictors selected by different methods; ^{c} mean squared error of $\boldsymbol{\beta}_\xi$, i.e., $\|\boldsymbol{\beta}_\xi - \boldsymbol{\beta}^*\|^2$, produced by different methods; ^{d} the posterior mean of $\boldsymbol{\beta}$ by model average is used as $\hat{\boldsymbol{\beta}}$ for SaM method; The numbers in the parentheses are the corresponding standard deviations.

To measure the performance of different methods for this example, we calculated the false selection rate (fsr) and negative selection rate (nsr). Let $\boldsymbol{t}$ denote the set of

true predictors, and let ξ denote the model selected by a method. Then, fsr and nsr can be defined as

$$\text{fsr} = \frac{|\xi \setminus \boldsymbol{t}|}{|\xi|}, \quad \text{nsr} = \frac{|\boldsymbol{t} \setminus \xi|}{|\boldsymbol{t}|},$$

where $|\cdot|$ denotes the size of a model. The smaller values of fsr and nsr are, the better the performance of the method is. From Table 2.2, it is easy to see that SaM yields a much smaller value of fsr than all other methods, and about the same value of nsr as SIS and ISIS. Lasso and elastic net yield a smaller nsr but at the expense of a high fsr. We have also compared the mean squared error of the estimator of $\boldsymbol{\beta}^*$. For SaM approach, we use the posterior mean of the second Bayesian analysis stage as the estimator of true coefficients. Strictly speaking, this estimator is not a sparse estimation, but it, as shown in Table 2.2, does indicate that SaM is more accurate than those selected by other methods.

A final remark for this massive data simulation is the computation cost. The same setting of Markov chain burn-in period and total iteration numbers are used as in section 2.3.1. A serial analysis of this big data set cost approximated 16 hours for Markov chain simulation on a single core of Intel® Xeon® CPU E5-2690(2.90Ghz). The computation time should be significant reduced if implemented in a parallel architecture. Since our approach has an embarrassingly parallel structure, the implementation of parallel computing does require communication among different nodes except collecting the selected variables' index at the end of first stage. Thus the bandwidth limit of the connection between nodes should not influence the computation time very much. Recently, there has been a development of scalable continuous shrinkage prior based on mixture of normal distribution (see e.g. [26], [59]), which aim at avoiding reversible jumps and reducing computation time. Gibbs sampler is always implemented, where the full condition of $\boldsymbol{\beta}$ follows multivariate normal

distribution, with computation complexity of order $O(p_n^3)$. In case of p_n equals half million, this is not feasible at all. The authors experience shows that it cost one day to update only 20 more iterations. In contrast, the computation complexity of proposed posterior is at most of the order of $O(\bar{r}_n^3)$. [3] show that such shrinkage priors lead to posterior consistency in the case of $p_n = o(n)$, and [59] also demonstrate its prediction outperformance compared to frequentist methods. However, to the authors' best knowledge, a general study of estimation consistency under shrinkage prior for the high dimensional regression is not established yet. In Appendix C, we present one simulation study that horseshoe prior fails for high dimensional variable selection.

2.4 Real Data Examples

2.4.1 mQTL Example

The first example is related to a metabolic quantitative trait loci (mQTL) experiment, which links SNPs data to metabolomics data [16]. The predictors come from a genome-wide analysis of candidate genes for ALAT enzyme elevation in liver with the Mass Spectroscopy metabolomics data as the response [38]. The spectra are divided by regions or bins to reduce the dimension of spectral data, and a log10-transformation is applied to normalize the signal. A total of 10,000 SNPs are preselected as candidate predictors based on the following criteria: no missing values, no monomorphic SNPs and close to known regulatory regions. The total number of subjects included in the dataset is 50. The genotype of each SNP is coded as 0,1 and 2 for homozygous rare, heterozygous and homozygous common allele, respectively. As in [8], one particular metabolite bin that discriminates between the disease status of the clinical trial's participants is chosen as the response.

The SaM approach was first applied to this example. The dataset was divided

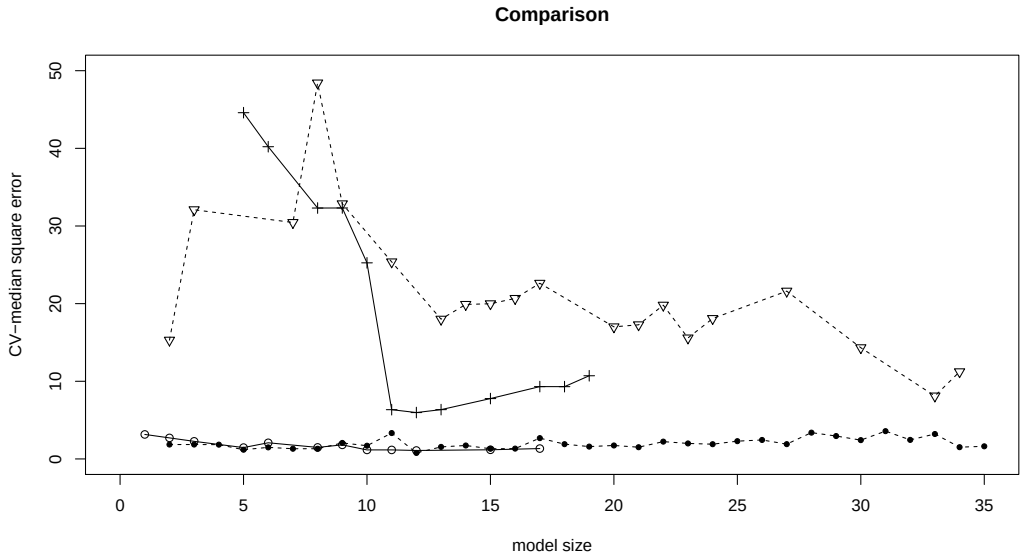

Figure 2.4: Results comparison for real mQTL data set. The plot compares the SaM result with Lasso, SIS-Lasso and ISIS-Lasso for the mQTL data: Leave-one-out cross validation median square error of the MAP_1 – MAP_{35} models produced by SaM (black dot with dashed line), the models selected by Lasso (triangle with dashed line), the models selected by SIS-Lasso ("+" with solid line), and the models selected by ISIS-Lasso (hollow dot with solid line).

into 20 subsets with $s = 500$ and SaM was then run with exactly the same setting as for the massive data example of Section 2.3.2. In Stage I SaM reduced the number of SNPs to 536 from 10,000, and in stage II SaM selected two SNPs, rs17041311 and rs17392161, whose marginal inclusion probabilities were 0.98 and 0.93, respectively. The two SNPs also compose the MAP model for this example. We note that in the dataset, the SNP rs7896824 has the same genotypes as the SNP rs17041311 across all 50 subjects. Similarly, the SNPs rs17390419, rs12328732, rs2164473, rs322664, rs17415876, rs16950829, rs6607364, rs829156, rs829157, rs2946537 share the same genotypes with the SNP rs17392161.

For comparison, SIS and ISIS were applied to this example. SIS-SCAD and ISIS-SCAD first reduced the number of SNPs from 10,000 to 25 and then applied SCAD to refine the selection, but both yielded the null model. To assess the performance of different methods, we used the leave-one-out crossing validation. The median cross-validation square error of the SaM model is 1.8, and that of the null model is 9.74.

We also compared the MAP_i models produced by SaM with those produced by Lasso, SIS-Lasso and ISIS-Lasso along their regularization paths. Here the MAP_i model refers to the maximum a *posteriori* model containing i SNPs. Figure 2.4 shows the leave-one-out median square error of these models. Lasso and SIS-lasso failed to select the SNP rs17041311, which is one of the two important SNPs identified by SaM, and thus yielded enormous median square errors. ISIS-Lasso successfully selected both rs17041311 and rs17392161, and thus yielded similar median square errors with SaM.

2.4.2 PCR Example

The second example relates to a PCR dataset. [45] conducted an experiment which examines the genetics of two inbred mouse population (C57BL/6J and BTBR). A total of 60 F2 samples, with 31 female and 29 male mice, were used to monitor the expression levels of 22,575 genes. Some physiological phenotypes, including numbers of phosphoenopyruvate carboxykinase (PEPCK), glycerol-3-phosphate acyltransferase (GPAT), and stearoyl-CoA desaturase 1 (SCD1) were measured by quantitative real-time PCR. In this example, we study the relationship between PEPCK (as responses) and the gene expression level. The gene expression data and the phenotype data are published at GEO (http://www.ncbi.nlm.nih.gov/geo; accession number GSE3330). The gene expression data have been normalized before statistical analysis.

SaM was first applied to this example. The data was divided into 45 subsets with $s = 502$. In stage I, SaM selected 1113 genes from 22575 genes. In stage II, SaM selected 6 genes under the setting $r_n = p_n^{0.3}$. The six selected genes are 1429089_s_at, 1430779_at, 1432745_at, 1437871_at, 1440699_at, 1459563_x_at. The first five genes also compose the MAP model. The leave-one-out cross validation mean square error of the model of the six genes is 0.084. In comparison, SIS-SCAD selected 17 genes with the leave-one-out mean square error 0.204, and the ISIS-SCAD selected 9 genes with the leave-one-out mean square error 0.112.

Figure 2.5 shows the leave-one-out mean square errors of the MAP_i models selected by SaM and the models selected by Lasso, SIS-Lasso and ISIS-Lasso along their regularization paths. For this example, SIS-Lasso and ISIS-Lasso first reduced the number of genes to 30, then applied Lasso to refine the selection. From Figure 2.5, it is easy to see that SaM significantly outperforms the penalized likelihood

approaches for this example.

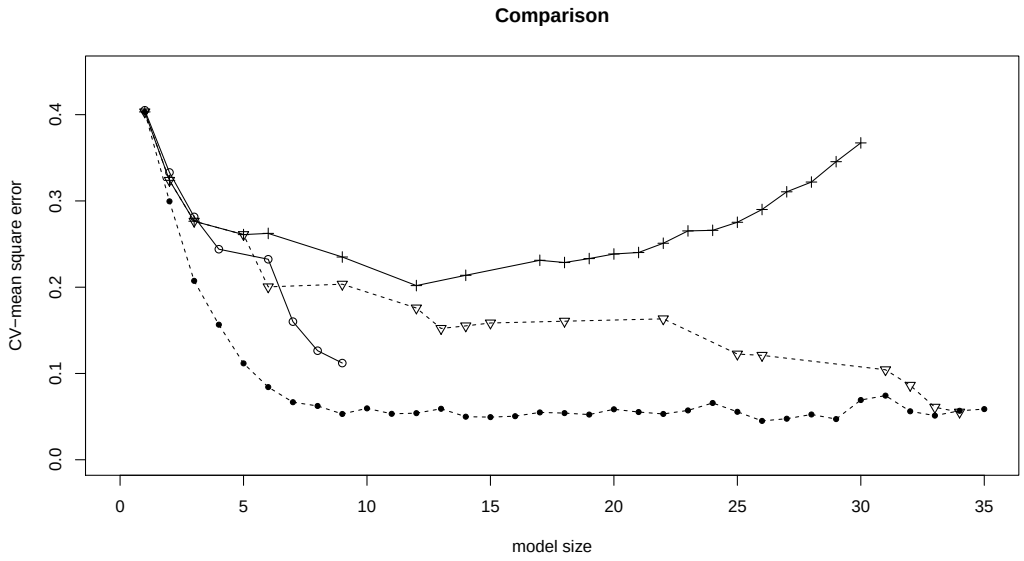

Figure 2.5: Results comparison for real PCR data set. The plot compares the SaM result with Lasso, SIS-Lasso, and ISIS-Lasso for the PCR data: leave-one-out cross validation median square error of the MAP_1 – MAP_{35} models produced by SaM (black dot with dashed line), the models selected by Lasso (triangle with dashed line), the models selected by SIS-Lasso ("+" with solid line), and the models selected by ISIS-Lasso (hollow dot with solid line).

3. FREQUENTIST APPROACH: RECIPROCAL LASSO PENALTY

In this section, we study the most popular frquentist approach for variable selection: penalized likelihood method. We begin our study by first looking at the classic low dimesinoal situation.

3.1 Low Dimensional Regression

In this section we study the property of the new class of penalty functions under the low dimension setting where $p < n$ and p is fixed as n increases. Let $\boldsymbol{\beta}^*$ denote the vector of true regression coefficients of model (1.1), let $\boldsymbol{t} = \{i, \beta_i^* \neq 0\}$ denote the true model, and let $|\boldsymbol{t}|$ denote the size of the true model. The design matrix $\boldsymbol{X}$ consists of n i.i.d. observations of p-dimensional predictors. Thus

$$\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} \to \Sigma, \tag{3.1}$$

where Σ is a $p \times p$ covariance matrix. Without loss of generality, we assume that the first $|\boldsymbol{t}|$ predictors are true predictors. Hence, Σ can be rewritten as

$$\Sigma = \begin{bmatrix} \Sigma_{\boldsymbol{t},\boldsymbol{t}} & \Sigma_{\boldsymbol{t},\boldsymbol{t}^c} \\ \Sigma_{\boldsymbol{t}^c,\boldsymbol{t}} & \Sigma_{\boldsymbol{t}^c,\boldsymbol{t}^c} \end{bmatrix}, \tag{3.2}$$

where $\Sigma_{\boldsymbol{t},\boldsymbol{t}}$ is the covariance matrix of the true model. We estimate $\boldsymbol{\beta}$ by minimizing the penalized residual sum of squares as given in (1.2). Let $\boldsymbol{\xi}(\hat{\boldsymbol{\beta}}) = \{i, \hat{\beta}_i \neq 0\}$ denote the selected model, where $\hat{\beta}_i$ is the ith component of $\hat{\boldsymbol{\beta}}$.

To motivate the design of the new class of penalty functions, we consider the model (1.1) under the simplest scenario where the columns of $\boldsymbol{X}$ are orthogonal and standardized such that $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle = 0$ and $\|\boldsymbol{x}_i\| = \sqrt{n}$ for $i \neq j$. Let $\boldsymbol{z} = \frac{1}{n}\boldsymbol{X}^T\boldsymbol{y}$ and

$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{z}$. With a slight abuse of notations, we write the penalty function in the form $P_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^{p} P_\lambda(|\beta_j|)$. This representation of the penalty function will also be used in the remaining part of the section. Then the penalized residual sum of squares in (1.2) can be rewritten as

$$\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + P_\lambda(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2 + \sum_{j=1}^{p}\left[\frac{1}{2}n(\beta_j - z_j)^2 + P_\lambda(|\beta_j|)\right], \tag{3.3}$$

where z_j denotes the jth component of $\boldsymbol{z}$. It is obvious that minimizing (3.3) is equivalent to componentwisely minimizing

$$\frac{1}{2}n(\beta_j - z_j)^2 + P_\lambda(|\beta_j|), \quad j = 1, \ldots, p. \tag{3.4}$$

The solution $\hat{\boldsymbol{\beta}}(\boldsymbol{z})$ of (3.4) (with respective to z) is usually called the thresholding function, which depends on the form of the penalty function $P_\lambda(\cdot)$ (see e.g., 18; 72). [18] claimed that a good penalty function should result in an estimator of $\boldsymbol{\beta}$ with three properties: unbiasedness, sparsity and continuity. Under Fan and Li's criterion, some penalty functions, although resulting in oracle properties for variable selection, should be considered suboptimal because the corresponding thresholding functions are discontinuous. Examples of such penalty functions include, for example, the L_0 penalty function used in AIC, BIC and EBIC, and the L_q $(0 < q < 1)$ penalty function used in bridge regression (22). The L_q penalty function is defined as $P_\lambda(\boldsymbol{\beta}) = \lambda \sum_{i=1}^{p} |\beta_j|^q$, where $\lambda = O(n^{q'})$ for some $q' \in (q/2, 1/2)$. Figure 3.1 shows the discontinuous threshold functions corresponding to the L_0 penalty (with $\lambda = \log(n)$) and L_q penalty $(q = 0.5, q' = 3/8)$.

Although the discontinuity may result in instability in model prediction, one should notice that as n grows, the discontinuity will vanish in the sense that the

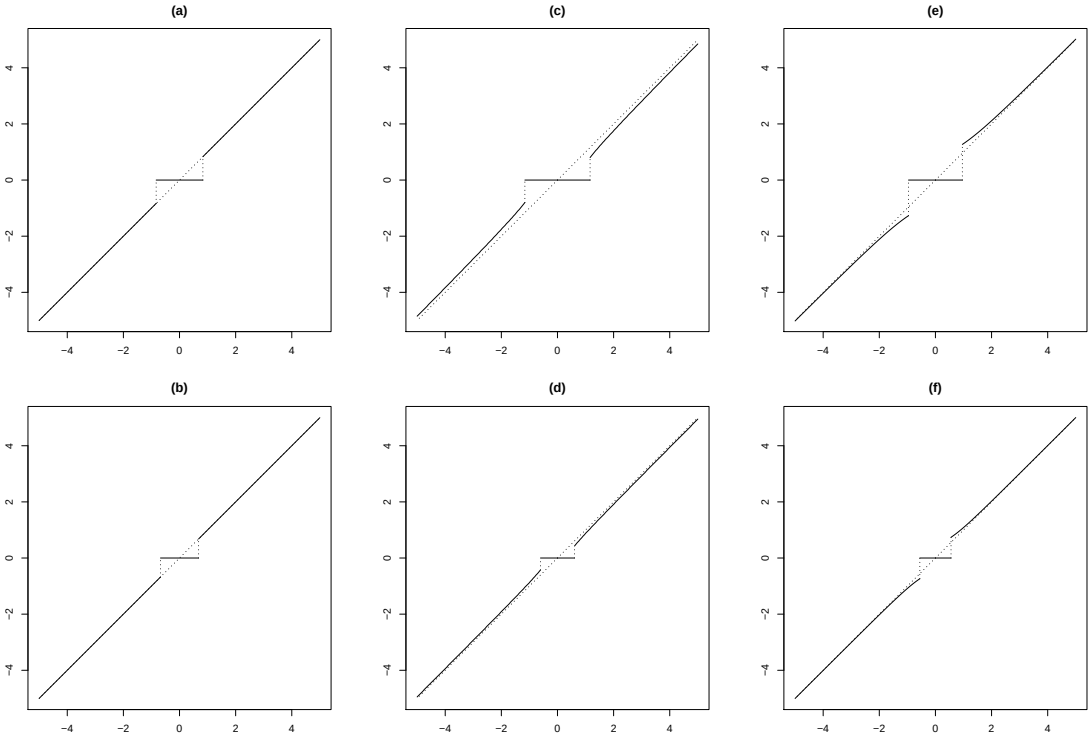

Figure 3.1: Discontinuous thresholding functions for three variable selection criteria. (a) BIC with L_0 penalty (first column), (b) bridge regression with $L_{.5}$ penalty (second column), and (c) rLasso (third column). The first row corresponds to the sample size $n = 2$, and the second row corresponds to the sample size $n = 10$. The discontinuity around the origin vanishes as n grows.

jump at the discontinuous point reduces to zero. Thus the authors conjecture that when n is sufficiently large, the continuity requirement is not crucial. On the other hand, the discontinuity around the origin can automatically avoid confusion in model interpretation when the coefficient estimate is tiny. Popular methods with continuous threshold functions, such as Adaptive Lasso and SCAD, can result in sparse models, but may still produce some petite coefficient estimates, say, at the order of 10^{-8}. In this case, we fall into a dilemma: whether or not to accept such small coefficient predictors.

The aim of regularization for variable selection is to control the model complexity and prevent the data from overfitting, where the penalty function penalizes against model complexity. From this perspective, the L_0 penalty works in a straightforward way by imposing a constant penalty on each predictor. The penalty functions used in Lasso, SCAD, MCP and many other methods are dependent on the value of $|\beta_j|$, which are to let the penalty $P_\lambda(|\hat{\beta}_j|)$ compete against the gain in residual sum of squares by adding an extra predictor $\boldsymbol{x}_j$. Take SCAD as the example. Loosely speaking, if one extra false predictor is added into the true model, the gain in residual sum of squares is a reduction from $\sigma^2\chi^2_{n-|\boldsymbol{t}|}$ to $\sigma^2\chi^2_{n-|\boldsymbol{t}|-1}$ with $\hat{\beta}_j = O(1/\sqrt{n})$, which is approximately $\sigma^2\chi^2_1$-distributed. Around the origin, the SCAD penalty function is $P_\lambda(|\beta|) = \lambda|\beta|$. To defeat the gain, the penalty $\lambda|\hat{\beta}_j|$ needs to be larger than the gain. Hence, to ensure the consistency of the method for variable selection, the condition $\lim_n(\lambda/\sqrt{n}) = \infty$ is required. The SCAD penalty works in an indirect way of regularization by competing the penalty against the gain, rather than directly imposing a penalty on model complexity.

From our point of view, controlling model complexity is to remove the predictors that do not contribute significantly to the response. Here the contribution should be understood as the partial contribution, i.e., the contribution conditional on other

predictors in the model. Provided that the response and predictors have been standardized in certain way such that $O(\|\boldsymbol{y}\|) = O(\|\boldsymbol{x}_i\|)$ for all $i = 1, \ldots, p$, which ensures that no coefficients can be extremely large (in magnitude), then the magnitude of a coefficient should serve as a good measure for the contribution of the corresponding predictor. Hence, it is reasonable to give a large penalty for a coefficient of small magnitude. In the literature, [37] proposed a non-local prior for Bayesian variable selection for model (1.1). For the case $n > p$, they achieved the so-called global consistency result that the posterior probability of the true model $\boldsymbol{t}$ will converge to 1 in probability as $n \to \infty$. Suppose that a predictor, say $\boldsymbol{x}_j$, has been selected into a model, the non-local prior imposes a zero prior density value for the regression coefficients of the model if $\beta_j = 0$. This translates into the penalty function that $\lim_{|\beta|\to 0} P_\lambda(|\beta|) = +\infty$.

Based on the above consideration, we propose a new class of penalty functions which satisfy the following conditions:

(C_1) $P_\lambda(0) = 0$.

(C_2) $P_\lambda(\cdot)$ is symmetric about 0 and $\lim_{|\beta|\to 0} P_\lambda(|\beta|) = \infty$.

For simplicity, we further assume

(C_3) $P_\lambda(\cdot)$ is continuously differentiable on $\mathbb{R}\backslash\{0\}$.

This class of penalty functions have a shape resembling to the shape of the absolute reciprocal function $1/|\beta|$. For this reason, we call the resulting penalized likelihood method the reciprocal Lasso, or shortly, the rLasso. For simplicity, we also call the new class of penalty functions the rLasso penalty functions. See Figure 3.2 for a comparison of shapes of different penalty functions.

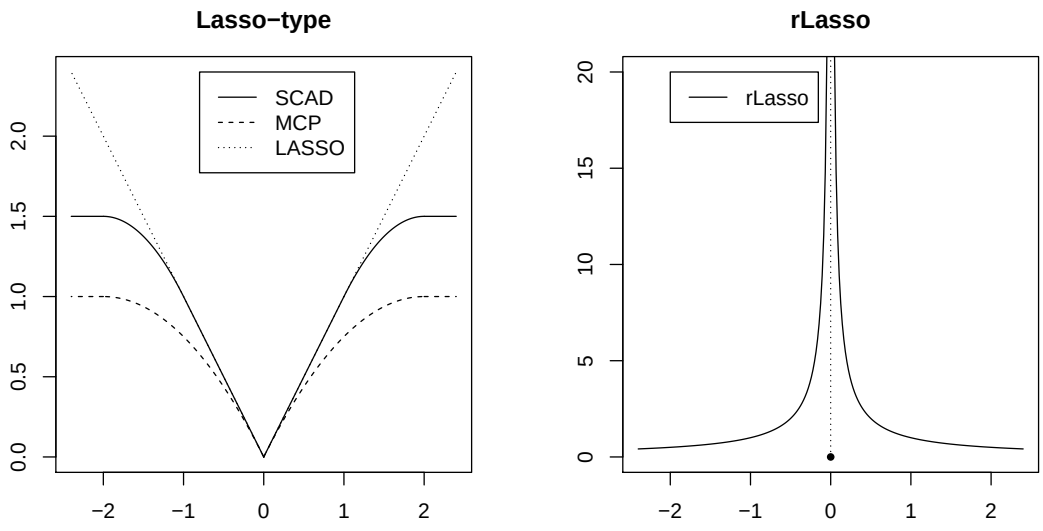

Figure 3.2: Comparison of shapes of different penalty functions.

Like the L_0 penalty function, the rLasso penalty function is discontinuous at 0 and yields a discontinuous thresholding function. See Figure 3.1, where we plot $P_\lambda(|\beta|) = 1/|\beta|1(\beta \neq 0)$. A key difference between the rLasso penalty function and the others is that the former is decreasing in $(0, \infty)$, while the others are not. This distinguishing feature makes the rLasso a heterodoxy for conventional regularization methods. However, rLasso still possesses the oracle property.

Theorem 3.1.1 (Oracle property)**.** *Consider a low dimensional linear regression, where p is fixed as n increases. If the penalty function $P_\lambda(\cdot)$ satisfies the conditions (C_1), (C_2) and (C_3), then the model selected by rLasso via minimizing (1.2) has the following properties:*

1. $\lim_n P(\boldsymbol{\xi}(\hat{\boldsymbol{\beta}}_n) = \boldsymbol{t}) = 1$, *where* $\boldsymbol{\xi}(\hat{\boldsymbol{\beta}}_n) = \{i : \hat{\beta}_i \neq 0\}$ *denotes the model corresponding to the vector* $\hat{\boldsymbol{\beta}}_n$ *and* $\hat{\beta}_i$ *denotes the ith element of* $\hat{\boldsymbol{\beta}}_n$*;*

2. $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\boldsymbol{t}} - \boldsymbol{\beta}^*_{\boldsymbol{t}}) \xrightarrow{d} N(0, \sigma^2 \Sigma^{-1}_{\boldsymbol{t},\boldsymbol{t}})$, *where* $\xrightarrow{d}$ *denotes convergence in distribution,*

and $\Sigma_{\boldsymbol{t},\boldsymbol{t}}$*, as defined in (3.2), is the covariance matrix of the true model.*

The proof of this theorem can be found in the Appendix. For this theorem, we have the following remarks:

1. For Lasso and the closely related methods, e.g., Lasso (Theorem 1 of 71), SCAD (Theorem 2 of 18) and Adaptive Lasso (Theorem 2 of 72), to achieve the oracle property, one usually needs to impose some conditions on the increasing order of λ with respect to n: λ shall neither increase too fast such that the penalty cannot beat the gain of selecting a true predictor, nor shall it increase too slowly such that the penalty can beat the gain of selecting a false predictor. For rLasso, such conditions on λ are not necessary as the assumption (C_2) has been strong enough to exclude false predictors. This makes the conditions of Theorem 3.1.1 look very neat.

2. Although the oracle result can be achieved asymptotically for rLasso with any fixed value of λ. In practice, one may still need a method to determine the optimal value of λ, e.g., cross-validation as used in Lasso and many other penalized likelihood methods.

Figure 3.3 shows the regularization paths of Lasso, SCAD and rLasso for a synthetic dataset. The data set consists of $n = 200$ observations and $p = 30$ predictors, where all predictors are generated from the standard multivariate Gaussian distribution $N(0, I_n)$ and the random errors are generated from $N(0, \sigma^2 I_n)$ with $\sigma = 1.5$. The first 8 predictors are true with the coefficients given by (2.63, 2.28, -1.43, 2.16, 1.73, 1.06, -1.7, -2.43). The rest 22 predictors are manipulated to have high random correlations with the response variable (we random generated 1,000 predictors from the standard multivariate Gaussian distribution and then selected the top 22

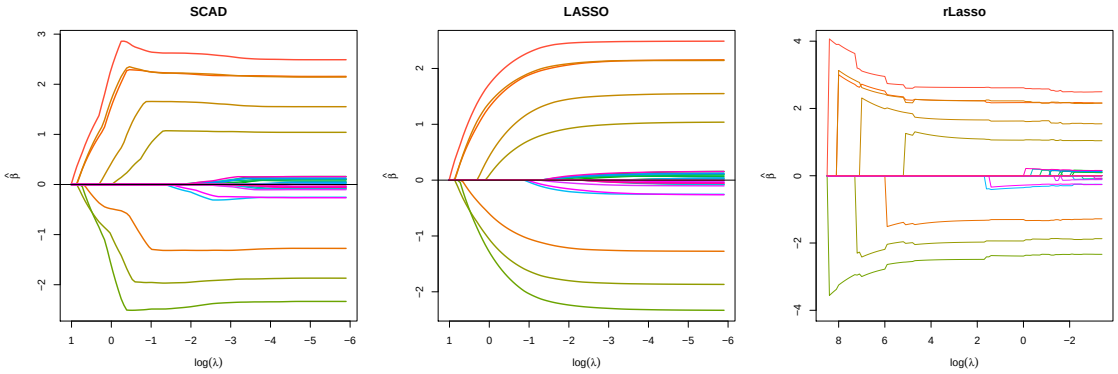

Figure 3.3: Regularization paths of SCAD, LASSO and rLasso for a simulated example.

which have the highest correlation with the response variable). Among the three regularization paths, the Lasso path is most smooth and continuous. SCAD, which uses a non-convex penalty, has also a continuous path. As a consequence of the discontinuity of the penalty and thresholding functions, the rLasso path is rugged and often jumps from a large value to zero.

Figure 3.4 zooms the core part of the regularization paths shown in Figure 3.3. For each method, we plot three vertical lines which indicate, respectively, the following values of λ:

- λ_1: dashed line, obtained by minimizing the prediction error of a 10-fold cross validation;
- λ_2: long-dashed line, obtained by minimizing $\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|^2$;
- λ_3: dash-dot line, the minimum value of λ at which the true model is selected and the coefficients of false predictors shrink to zero.

Hence, λ_2 represents the oracle best λ for prediction, λ_3 represents the oracle best λ for model selection, and λ_1 can be viewed as an estimator of λ_2. For rLasso,

$\lambda_2 = \lambda_3$ exactly. In contrast, for SCAD and LASSO, $\lambda_2 < \lambda_3$; that is, the oracle best prediction model includes some false predictors. Because the regularization paths of LASSO and SCAD are continuous, some false predictors with tiny coefficient estimates have a chance to survive through cross-validations. However, for rLasso, its path is discontinuous, so its coefficient estimates are either zero or somewhat bounded away from zero, and this discontinuity gives rLasso a power to rule out false predictors.

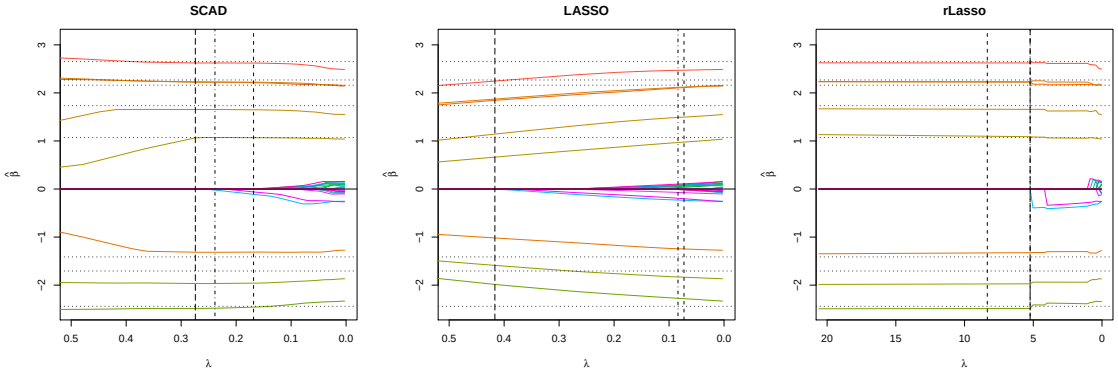

Figure 3.4: Zoomed regularization paths of SCAD, LASSO and rLasso for a simulated example. The dotted horizontal lines indicate true values of the non-zero coefficients, and three vertical lines indicate λ_1 (dashed line), λ_2 (long-bashed line), and λ_3 (dash-dot line), respectively. For rLasso, the long-dashed line overlaps with the dash-dot line.

3.2 High Dimensional Regression

In this section we study the property of the new class of penalty functions under the high dimension setting where $p \gg n$ and p can increase with n. To indicate the dependence of p on n, we rewrite p as p_n in what follows. To accommodate the high dimensional setting, we assume that the penalty function satisfies, in addition

to (C_1)–(C_3), the following conditions:

(C_4) $P_\lambda(|\beta|)$ is increasing and convex in $(-\infty, 0)$, and decreasing and convex in $(0, +\infty)$;

(C_5) $\lim_{|\beta|\to\infty} P_\lambda(|\beta|) = c_\lambda$, where $c_\lambda \geq 0$;

(C_6) There exist some constants b_λ and $a_\lambda \geq 0$, such that $P_\lambda(b_\lambda) \leq c_\lambda + a_\lambda$.

Under the high dimensional setting, we estimate $\boldsymbol{\beta}$ by minimizing the penalized residual sum of squares subject to the constraint $|\boldsymbol{\xi}(\boldsymbol{\beta})| \leq r_n$, i.e.,

$$\hat{\boldsymbol{\beta}}_n = \arg \min_{|\boldsymbol{\xi}(\boldsymbol{\beta})| \leq r_n} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p_n} P_{\lambda_n}(|\beta_j|) \right\}, \tag{3.5}$$

where $\boldsymbol{\xi}(\boldsymbol{\beta}) = \{i : \beta_i \neq 0\}$ denotes the model corresponding to the vector $\boldsymbol{\beta}$, $|\boldsymbol{\xi}(\boldsymbol{\beta})|$ denotes the size of the model $\boldsymbol{\xi}(\boldsymbol{\beta})$, and r_n is a user-specified parameter. Here, it is assumed that r_n can also increase with n. Let P_{λ_n} denote a penalty function with the tuning parameter λ_n, where λ_n depends on the value of n. We use the symbol $\prec$ for comparing the orders of two sequences; $a_n \prec b_n$ means $\lim_{n\to\infty} a_n/b_n = 0$. Given the above notations, we have the following theorem concerning the consistency of rLasso for variable selection and coefficient estimation.

Theorem 3.2.1. *Consider the linear regression (1.1) under the high dimensional setting. Suppose that the predictors have been standardized such that* $\|\boldsymbol{x}_i\| = \sqrt{n}$, $p_n < \exp(Cn^\alpha)$ *for some constants* $C > 0$ *and* $\alpha < 1$, *and the conditions* (C_1)–(C_6) *and the following conditions hold:*

1. *The true model* $\boldsymbol{t}$ *is fixed with the fixed regression coefficients* $\beta^*_{\boldsymbol{t}}$ *for all* n.

2. $r_n \succ 1$ *and* $r_n \log(p_n) \prec n$.

3. *There exist constants l_* and l^* such that $l_* < l^*$ and the following eigenstructure condition*

$$nl_* \leq \min_{|\boldsymbol{\zeta}| \leq |\boldsymbol{t}|+r} ch_1(\boldsymbol{X}_{\boldsymbol{\zeta}}^T \boldsymbol{X}_{\boldsymbol{\zeta}}) \leq \max_{|\boldsymbol{\zeta}| \leq |\boldsymbol{t}|+r} ch_1'(\boldsymbol{X}_{\boldsymbol{\zeta}}^T \boldsymbol{X}_{\boldsymbol{\zeta}}) \leq nl^*, \quad \text{for any subset model } \boldsymbol{\zeta}, \tag{3.6}$$

holds for all n, where $ch_1(\Sigma)$ and $ch_1'(\Sigma)$ denote the smallest and largest eigenvalues of the matrix Σ, respectively.

4. *For some sequences $\{K_n\}$ and $\{\kappa_n\}$ satisfying $K_n \geq \log(p_n)$, $r_n K_n \prec n$ and $\kappa_n \succ \sqrt{K_n/n}$, the penalty function $P_{\lambda_n}(\cdot)$ satisfies the conditions*

$$P_{\lambda_n}(\kappa_n) \succ K_n, \quad \limsup_n b_{\lambda_n} < \min_{i \in \boldsymbol{t}} \{\beta_i^*\}, \quad c_{\lambda_n} \prec n, \quad \text{and} \quad a_{\lambda_n} \leq O(K_n).$$

Then the model $\boldsymbol{\xi}(\hat{\boldsymbol{\beta}}_n)$ obtained via minimizing (3.5) satisfies

$$Pr(\boldsymbol{\xi}(\hat{\boldsymbol{\beta}}_n) = \boldsymbol{t}) > 1 - O(Ce^{-K_n}),$$

for some positive constant C, and

$$\hat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}^*,$$

where $\xrightarrow{p}$ denotes the convergence in probability. Furthermore, the mean squared estimation error of $\hat{\boldsymbol{\beta}}_n$ is bounded by

$$E\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*\|^2 \leq C_1 r_n \exp(-K_n) + C_2 \frac{K_n}{n},$$

where C_1 and C_2 are positive constants. Note that all the probabilities and expec-

tations in this theorem are with respect to the distribution of $\boldsymbol{y}$ *conditional on the design matrix* $\boldsymbol{X}$.

The proof of Theorem 3.2.1 can be found in the Appendix. Concerning the conditions and results of this theorem, we have the following remarks:

1. For the penalty function, c_λ can be viewed as a L_0-factor incorporated into the rLasso. The conditions on a_{λ_n} and b_{λ_n} constrain the values of $P_{\lambda_n}(\cdot)$ in the region bounded away from 0 such that the penalty cannot be too large at $\min_{i\in\boldsymbol{t}}\{\beta_i^*\}$. These conditions, together with condition (C_2), suggest that $P_{\lambda_n}(\cdot)$ should be sufficiently steep around zero to ensure the sparsity of the selected model.

2. Under the high dimensional setting $p \gg n$, the random matrix theory implies the possibility of $r_n \sim O(n/\log(p_n/n))$ such that (3.6) holds. For example, if all predictors follow the multivariate Gaussian distribution with $E(\boldsymbol{X}) = 0$ and $l_1 \leq E\|\boldsymbol{X}\boldsymbol{b}\|/n \leq l_2$ for any $\|\boldsymbol{b}\| = 1$, then $P\{(3.6) \text{ holds}\} \to 1$ with $l_* = l_1(1-\delta)^2$, $l^* = l_2(1+\delta)^2$ and $\log\binom{p_n}{r_n} < n\delta^2/2$ for any fixed $0 < \delta < 1$. See [14], [70] and [69] for the detail.

3. If $r_n \prec n^{\alpha_1}$ for some $\alpha_1 \in (0,1)$, then we can choose $K_n = n^\gamma \geq \log(p_n)$, where $0 < \gamma < 1 - \alpha_1$. In this case, the oracle result can be rewritten as $Pr(\boldsymbol{\xi}(\hat{\boldsymbol{\beta}}_n) = \boldsymbol{t}) > 1 - O(C\exp(-n^\gamma))$.

4. The optimization in (3.5) is subject to the model size constraint. Such a L_0-style constraint is necessary, especially when c_{λ_n} is small. In the extreme situation where $c_{\lambda_n} \equiv 0$, one can always select any more than n predictors with arbitrarily large coefficients such that the residual sum of squares is exactly 0. Hence the rLasso will fail without such a constraint. However, if $c_{\lambda_n} \succ n/r_n$,

the optimization is asymptotically equivalent to the unconstrained one. Refer to the Remarks of Lemma B.2.1 for more discussions on this issue.

Corollary 3.2.1. *Assume that* $p_n < \exp(Cn^{\alpha_2})$, $r_n < n^{\alpha_1}$, *for some positive constants* α_1 *and* α_2 *with* $\alpha_1+\alpha_2 < 1$, *and the condition (3.6) holds for all n. Let* $P(\cdot)$ *be a function which is symmetric about 0, decreasing and convex in* $(0, +\infty)$, $P(0) = 0$, *and has the limit* $\lim_{|\beta|\to 0} P(|\beta|) = \infty$. *Let*

$$P_{\lambda_n}(|\beta|) = P(|\beta|) + \lambda_n 1(\beta \neq 0), \quad \textit{with } \log p_n \prec \lambda_n \prec n; \tag{3.7}$$

or let

$$P_{\lambda_n}(|\beta|) = \lambda_n P(|\beta|), \quad \textit{with } \lambda_n = O(\log p_n). \tag{3.8}$$

Then, as $n \to \infty$, *the following convergence results hold for the solution of optimization (3.5):*

$$Pr(\boldsymbol{\xi}(\hat{\boldsymbol{\beta}}_n) = \boldsymbol{t}) > 1 - O(C\exp(-n^{\alpha_2})),$$

and

$$\hat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}^*.$$

More specifically, if we let $P(|\beta|) = |\beta|^{-1}1(\beta \neq 0)$, $P_{\lambda_n}(|\beta|) = \lambda_n P(|\beta|)$, *and* $(\log p_n)^{3/2}/\sqrt{n} \prec \lambda_n \leq O(\log p_n)$, *then the above convergence results still hold.*

Remarks for the results of Corollary 3.2.1:

1. The penalty function of form (3.7) can be viewed as a fixed rLasso penalty plus an L_0 penalty, where the L_0 term is the dominated one. Such a penalty function is comparable with the L_0 penalty used in EBIC.

2. The penalty function of form (3.8) is very interesting, and it has been used in all examples of this section. From Theorem 3.2.1, the order of λ_n can be

chosen as $K_n/P(\kappa_n) \prec \lambda_n = O(K_n)$, where κ_n can be chosen as $\kappa_n \to 0$ with $K_n = \log p_n$. Since $\lim_{|\beta|\to 0} P(|\beta|) = \infty$, λ_n can always be admitted with a smaller order than $O(\log(p_n))$. In MCP and EBIC, the consistency requires $P_{\lambda_n}(\min_{i\in \boldsymbol{t}} \beta_i^*) \geq O(\log p_n)$. In rLasso, we could set the penalty at a lower order of $\log(p_n)$ to avoid missing true predictors due to over-large penalties.

3.3 Computational Strategy for rLasso

3.3.1 Monte Carlo Optimization

The optimization in (3.5) can be a big challenge for rLasso. Many of the existing penalized likelihood methods use a penalty function which is concave or non-concave, but at least continuous. Hence the penalty function can be approximated by a local linear or quadratic function (18; 74), then the objective function can be optimized using an efficient algorithm such as LARS (17) or coordinate descent (9; 24).

For rLasso, the penalty function is discontinuous at 0. Hence, the local linear or quadratic approximation is not available. Furthermore, the objective function has multiple local optimal solutions. However, on any subspace or hyper-plain $\{\boldsymbol{\beta} \in \mathbb{R}^p : \text{sign}(\beta_i) = \omega_i, i = 1, \ldots, p\}$, where $\omega_i = -1, 0$ or 1 for $i = 1, \ldots, p$, the objective function in (3.5) is convex and has an unique local minimum. This observation helps us to design a Monte Carlo optimization algorithm for solving the optimization problem.

Let $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_p)^T \in \{-1, 0, 1\}^p \triangleq \mathcal{W}$. Define

$$L(\boldsymbol{\omega}) = \min_{\text{sign}(\beta_i)=\omega_i} \{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p} P_\lambda(|\beta_j|)\},$$

$$\boldsymbol{\beta}(\boldsymbol{\omega}) = \arg \min_{\text{sign}(\beta_i)=\omega_i} \{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p} P_\lambda(|\beta_j|)\}.$$

For any given $\boldsymbol{\omega} \in \mathcal{W}$, by convexity, $\boldsymbol{\beta}(\boldsymbol{\omega})$ can be easily obtained using an efficient algorithm such as Newton-Raphson or coordinate descent. Hence the optimization in (3.5) is equivalent to find

$$\hat{\boldsymbol{\omega}} = \arg \min_{\sum |\omega_i| \le r} L(\boldsymbol{\omega}), \tag{3.9}$$

and then the solution for rLasso is $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}(\hat{\boldsymbol{\omega}})$.

To find $\hat{\boldsymbol{\omega}}$, one possible way is to simulate from a Boltzmann distribution $f_\tau(\boldsymbol{\omega})$ defined at a small value of τ using the stochastic approximation Monte Carlo (SAMC) algorithm (50), where

$$f_\tau(\boldsymbol{\omega}) \propto \exp\left(-\frac{L(\boldsymbol{\omega})}{\tau}\right) I\left(\sum |\omega_i| \le r\right),$$

where $I(\cdot)$ is the indicator function, τ is called the temperature, and $L(\boldsymbol{\omega})$ is called the energy function which can be evaluated using the coordinate descent algorithm. Note that $f_\tau(\boldsymbol{\omega})$ is defined on a discrete space, where each sample point corresponds to a local minimum of the penalized likelihood function. To simulate from $f_\tau(\boldsymbol{\omega})$, conventional MCMC algorithms, such as the Metropolis-Hasting algorithm, can perform very badly when the energy function $L(\boldsymbol{\omega})$ is rugged. SAMC is an adaptive MCMC algorithm, which is designed to be immune to local trap problems. SAMC can adapt the invariance distribution at each iteration such that the Markov chain will asymptotically reach an equilibrium distribution under which the subspaces of $\mathcal{W}$ associated with different energy levels can be equally visited. This self-adjusting mechanism of the invariance distribution makes the SAMC chain capable to travel among different energy levels freely without being trapped. The temperature τ should be set to a small value, say .01 or .001, such that the majority mass of $f_\tau(\boldsymbol{\omega})$

is around $\hat{\boldsymbol{\omega}}$.

One may also combine SAMC with simulated annealing (39). The latter simulates a sequence of $f_{\tau_i}(\boldsymbol{\omega})$'s along a decreasing temperature sequence $\tau_1 > \tau_2 > \tau_3 > \cdots$. [49] proposed such a combination algorithm, the so-called stochastic approximation annealing (SAA) algorithm, for global optimization. SAA is different from simulated annealing in two respects. First, SAA can work with a square root cooling schedule $\tau_t \propto \sqrt{1/t}$, where t is the iteration number, while still guaranteeing that a sequence of samples can converge to the global energy minima in probability. It is known that to achieve the global convergence, simulated annealing needs to work with a logarithmic cooling schedule. Second, SAA employs SAMC moves as its local moves, while simulated annealing employs Metropolis-Hastings moves. This difference makes SAA less likely get trapped into local energy minima compared to simulated annealing. [49] showed that SAA generally outperforms simulated annealing in optimization. In this work, the SAA algorithm is used. The implementation of SAA for rLasso is similar to the implementation of SAMC algorithm, except that the temperature is decreasing, and some more types of local move in the sign space. Refer to the supplementary material of [52]. As illustrated by Figure 3.6, SAA can converge very fast, usually within 10^5 iterations for a dataset with a few thousands of predictors and a few hundreds of observations.

Finally, we note that unlike deterministic optimization algorithms which depend only on some user-specified settings (e.g., starting point, step scale, *etc.*), finite iterations of SAA may produce slightly different models in different runs. Researchers may perform multiple runs and select the best model produced thereby.

3.3.2 Tuning Parameter λ

The asymptotic theory (see Lemma B.2.1 and Corollary 3.2.1) provides us some suggestions for the choice of λ. For example, $\hat{\sigma}^2 \log(p)$ can be a reasonable choice of λ if $P(|\beta|) = |\beta|^{-1} I(\beta \neq 0)$. In practice, it is still of interest to use some data-dependent criteria to select the optimal value of λ. As for other penalized likelihood methods, a general way to select λ is cross validation, which is to choose λ such that the prediction error of a k-fold cross-validation is minimized. For example, we can set $k = 5$ or $k = 10$.

In practice, to stabilize the value of λ with respect to the number of observations, one often rescales the residual sum of squares and sets the objective function of minimization as

$$\frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2}{n} + \sum_{j=1}^{p} P_\lambda(\beta_j). \tag{3.10}$$

For cross-validation, the authors suggest to try a sequence of points equally spaced between ϵ and λ_m in the logarithmic scale, where ϵ is a user-specified small number and λ_m is the smallest tuning parameter value such that the null model is selected. Under the choice $P_\lambda(\beta) = |\beta| I(\beta \neq 0)$, if the response variable and all predictors have been standardized, then it is easy to derive that

$$\lambda_m = \frac{64 n \rho^3}{54}, \tag{3.11}$$

where $\rho = \max_{1 \leq i \leq p} \{|\text{corr}(\boldsymbol{x}_i, \boldsymbol{y})|\}$ and $\text{corr}(\cdot, \cdot)$ denotes the correlation function.

3.4 Numerical Studies and Real Data Applications

To illustrate the performance of the rLasso for variable selection under the high dimensional setting, we present two simulated and one real data examples along with comparisons with MCP, EBIC, SIS-SCAD, ISIS-SCAD, Lasso and Elastic Net (73).

We use the R package *SIS* to implement the SIS-SCAD and ISIS-SCAD methods, use the R package *glmnet* (24) to implement the Lasso and Elastic Net, and use the R package *ncvreg* (9) to implement the MCP. The simulation studies are designed to compare the performance of different methods in terms of accuracy of variable selection and coefficient estimation. In both simulation studies, we fixed the number of predictors $p = 1000$ with the true model

$$\boldsymbol{y} = \sum_{i=1}^{8} \beta_i^* \boldsymbol{x}_i + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I_n)$ and $\sigma = 1.5$. The true coefficients β_i^*, i=1,...,8, were randomly set to -1 with probability 0.4 and 1 with probability 0.6. We varied the value of n. As n increases, all methods should perform better as more information is contained in the data set. The performance of different methods will also be evaluated with respect to n.

In applications of SIS and ISIS, we first reduced the dimension p to $\max\{34, n/\log(n)\}$, and then applied SCAD. In applications of Lasso and Elastic Net, we controlled the model size to be not greater than $n/2$ by setting the argument *dfmax*= $n/2$ in the R package *glmnet*. The MCP employs a minimax concave penalty function of the form

$$P_\lambda(\boldsymbol{\beta}) = \lambda \sum_{i=1}^{p} \int_0^{|\beta_i|} (1 - t/(\gamma\lambda))_+ dt,$$

where $(z)_+ = z$ if $z > 0$ and 0 otherwise, and γ is a user-specified parameter. For MCP, we set $\gamma = 1.4$, which is also the optimal value of γ in high dimensional

regression experiment 3 of [69]. For rLasso, we set the penalty function as

$$P_\lambda(|\beta|) = \frac{\lambda}{|\beta|} 1(\beta \neq 0), \tag{3.12}$$

and constrained the model size to be not greater than $r = 35$. For all the above methods, λ is determined via a 10-fold cross-validation. For EBIC, we used the proposed Monte Carlo optimization algorithm to minimize its objective function

$$\frac{n}{2}\log\left(\frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2}{n}\right) + |\boldsymbol{\xi}(\boldsymbol{\beta})|\left(\frac{1}{2}\log(n) + \gamma\log(p)\right),$$

where γ is a user-specified parameter taking values in $(0.5, 1]$. For the simulated examples, we set $\gamma = 0.75$. We have tried several different values of γ. If γ is too small, EBIC will fail to maintain the model sparsity; and if γ is too big, EBIC may miss some true predictors especially when n is small. It seems that $\gamma = 0.75$ gives a balanced performance for EBIC.

3.4.1 Study I: Independent Predictors

In this study, the predictors were independently generated from the Gaussian distribution $N(\mathbf{0}, I_n)$, where I_n is the $n \times n$ identity matrix, and n=80, 100, 120, 150, 170 and 200. For each value of n, $m = 100$ data sets were simulated. As pointed out by [18], even with independent Gaussian predictors, variable selection under the high dimensional setting is far from trivial.

To measure the performance of different methods, we considered four quantities:

- Selection rate of true models, which is defined by $CR = \sum_{i=1}^{m} I(\hat{\boldsymbol{\xi}}_i = \boldsymbol{t})/m$, where $\hat{\boldsymbol{\xi}}_i$ denotes the model selected for dataset i.
- Squared coefficient estimation error, i.e., $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|^2$.

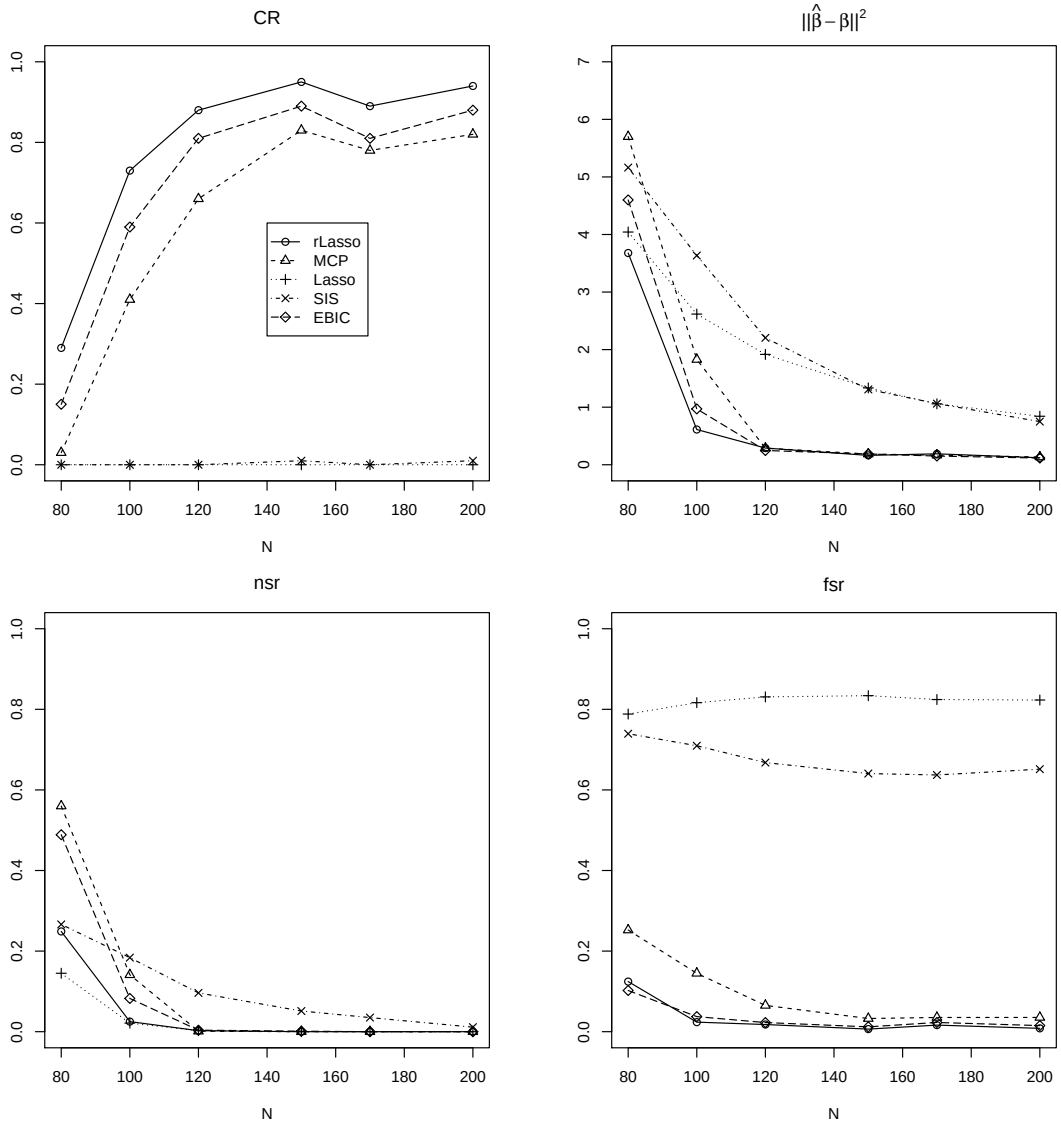

Figure 3.5: Results comparison under independent scenario between rLasso, MCP, EBIC, Lasso and SIS-SCAD for the datasets simulated in study I. The four plots show the selection rate of true models (upper left), squared coefficient estimation error (upper right), negative selection rate (down left) and false selection rate (down right), respectively.

- Negative selection rate, which is defined by $nsr = \sum_{i=1}^{m} |\boldsymbol{t} \backslash \hat{\boldsymbol{\xi}}_i| / (m|\boldsymbol{t}|)$.

- False selection rate, which is defined by $fsr = \sum_{i=1}^{m} |\hat{\boldsymbol{\xi}}_i \backslash \boldsymbol{t}| / (m|\hat{\boldsymbol{\xi}}|)$.

For a good method, we expect a higher value of CR, a lower coefficient estimation error, a smaller value of nsr and a smaller value of fsr.

Figure 3.5 plots these four quantities versus the sample size n for rLasso, MCP, Lasso, SIS and EBIC. ISIS and Elastic Net perform quite badly in this study. The readers of interest can refer to the Appendix C for their results. The comparison of the CR curves shows that rLasso has the best power to select true models, followed by EBIC and MCP. SIS and Lasso seldom select true models. The comparison of the nsr and fsr curves shows that the true predictors can be asymptotically selected by all the methods; nsr goes to zero as n becomes large. However, SIS and Lasso failed to control the sparsity of the selected models; they tend to produce over-dense models. In particular, Lasso is the best in terms of nsr, but the worst in terms of fsr. In overall performance, rLasso is clearly the champion among the five methods under comparison.

When n is large, say, greater than 120, MCP also yields good results with comparable coefficient estimation error, fsr and nsr as rLasso. As discussed in [9], although SCAD has a similar design to both MCP and Lasso, it is closer to Lasso than to MCP. Based on this study, we can also draw the same conclusion as SIS-SCAD produced more similar results to Lasso than to MCP. A comparison of rLasso and EBIC shows that when n is large, they perform equally well; however, when n is small, EBIC tends to produce over-sparse models by missing some true predictors. As aforementioned, this is due to that EBIC employs an overly large penalty when the ratio $\log(p)/n$ is large.

We also gave a full numerical summary of the study in Appendix C, where for

each of the methods, including rLasso, MCP, Lasso, SIS, ISIS, EBIC, and Elastic Net, and each value of n we report the average values of nsr, fsr and coefficient estimation error, and their standard deviations as well.

To give the readers some idea about the CPU cost of rLasso, we plot in Figure 3.6 the sample paths of the SAA algorithm in minimizing the objective function (3.5) for a dataset simulated in this study. The dataset consists of 200 observations. SAA was run on the whole dataset for 10 times with the fixed tuning parameter $\lambda = 1.5^2 \log(1000) \approx 15$ (without cross-validation). Figure 3.6 shows the minimum value of $L(\boldsymbol{\omega})$ found by SAA versus the number of iterations, along with the CPU time reported in seconds. The simulation was done on a single core of Intel® Xeon® CPU E5-2690 (2.90GHz). The plot shows that SAA can identify the true model very fast for this dataset: It usually converges to the true model within 3×10^4 iterations and costs less than 3 seconds. The performance of SAA is similar for other examples of this work. For each dataset of this and next studies, we let SAA run for 10^5 iterations for each of the optimization tasks involved in rLasso, which ensures that the global minimum of $L(\boldsymbol{\omega})$ can be found in a high probability. In summary, SAA together with the coordinate descent algorithm provides an efficient computational tool for rLasso.

3.4.2 Study II: Dependent Predictors

This study has a similar design with study I except that the predictors are now correlated with each other. In this study, the $p = 1,000$ predictors were generated from the multivariate normal distribution $\mathcal{N}(0, A)$, where $A_{ii} = 1$, $A_{ij} = 0.5$ for all i and $i \neq j$. Since the predictors are generally correlated, we tried slightly larger numbers of observations, n = 100, 120, 150, 170 and 200. Again, for each value of n, 100 datasets were randomly generated. Since the false predictors are highly

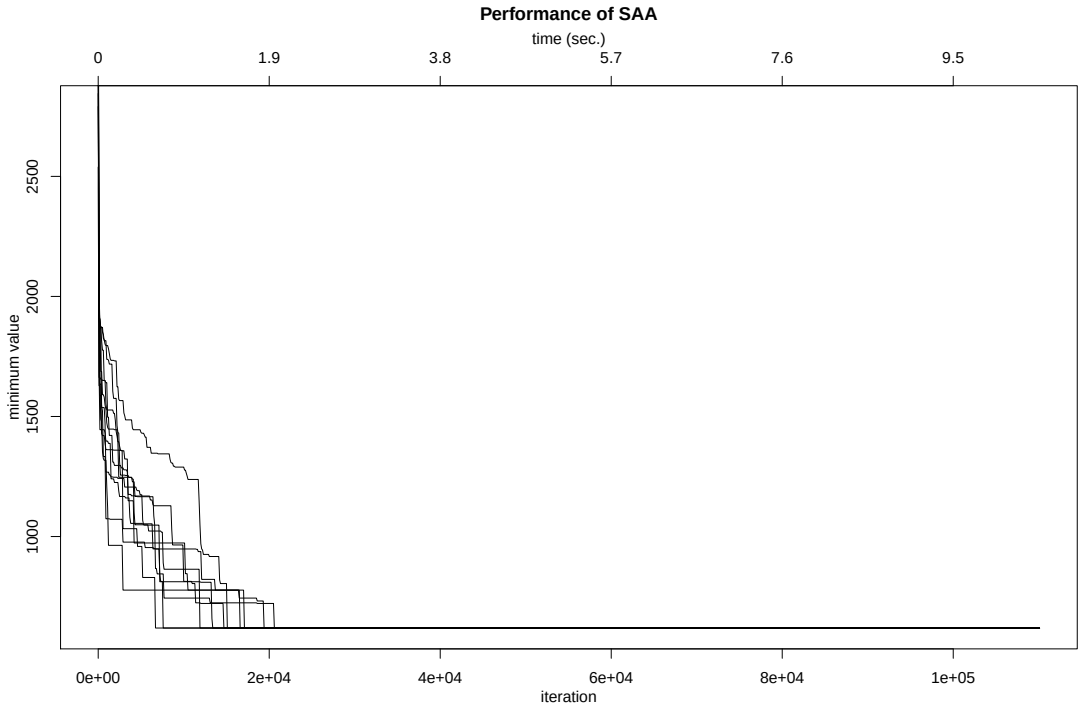

Figure 3.6: Illustration of SAA performance. The plot gives the sample paths of SAA for a dataset generated in study I: The curves show the minimum value of $L(\boldsymbol{\omega})$ found by SAA versus the number of iterations, along with the CPU time reported in seconds at the top horizontal axis. Each curve corresponds to an independent run of SAA.

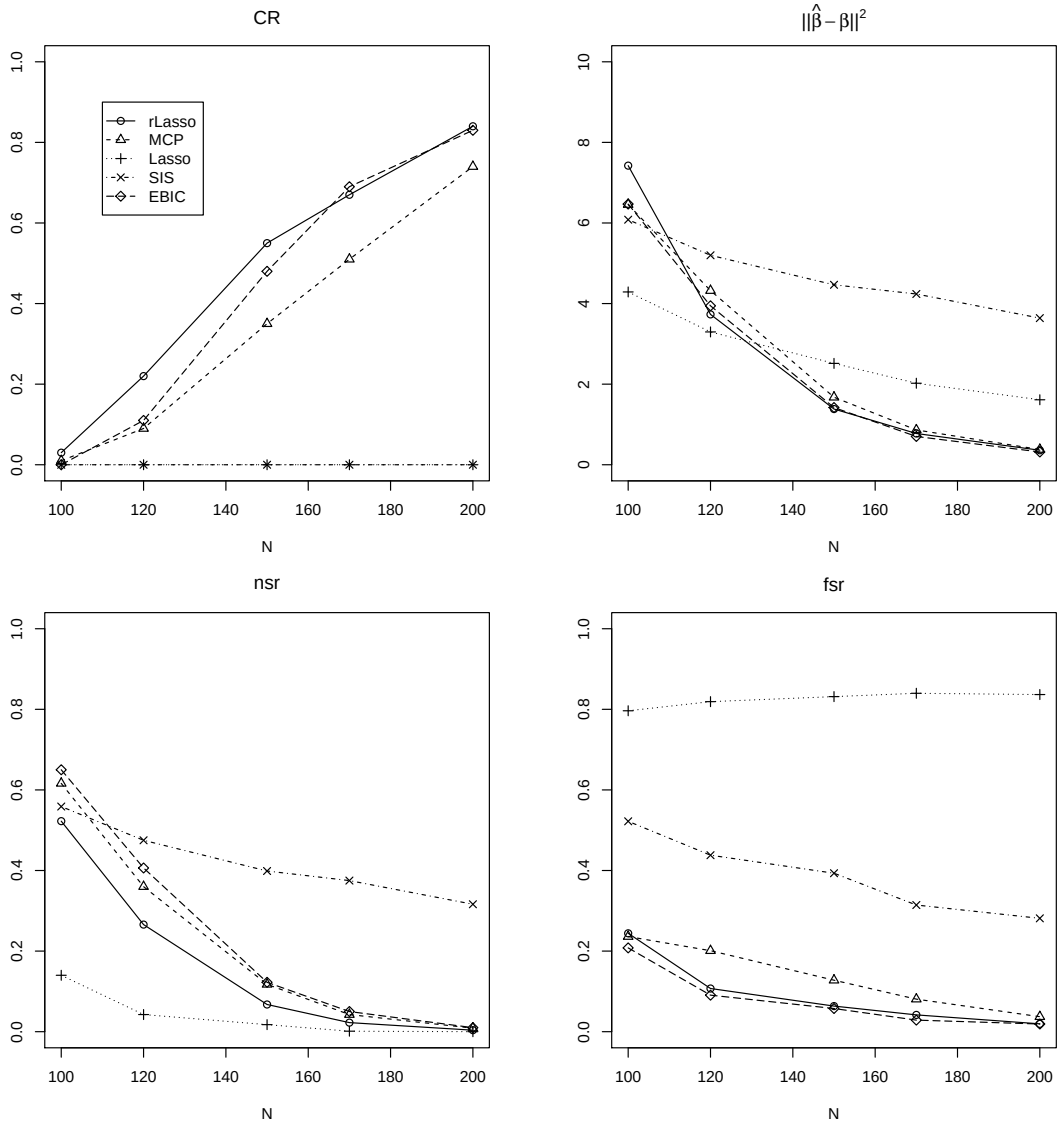

Figure 3.7: Results comparison under dependent scenario between rLasso, MCP, EBIC, Lasso and SIS-SCAD for the datasets simulated in study II. The four plots show the selection rate of true models (upper left), squared coefficient estimation error (upper right), negative selection rate (down left) and false selection rate (down right), respectively.

correlated with the response and true predictors, see Table 3.1 for a summary of the correlations, this study poses a great challenge on all variable selection methods.

Sample size (n)	80	100	120	150	170	200
Study I	0.38	0.34	0.31	0.28	0.26	0.24
	(.0036)	(.0032)	(.0028)	(.0027)	(.0027)	(.0023)
Study II	—	0.55	0.55	0.52	0.52	0.50
	—	(.0175)	(.0164)	(.0175)	(.0170)	(.0178)

Table 3.1: Severeness of multicollinearity of the simulated dependent data sets. The table contains the means and standard deviations of the maximum absolute sample correlations between the response and false predictors for studies I and II, where each entry is calculated based on 100 simulated datasets.

Figure 3.7 summarizes the results of this study (the detailed full numerical results are presented in Appendix C). It shows a similar pattern to Figure 3.5. When n is large, rLasso and EBIC perform similarly. They produce the highest selection rates of true models, very small coefficient estimation error, very small fsr, and very small nsr. When n is small, rLasso and EBIC have also the highest selection rates of true models and very low values of fsr, but higher values of nsr and coefficient estimation error. Note that in this scenario, rLasso outperforms EBIC. Compared to EBIC, rLasso has higher selection rates of true models and lower values of nsr; that is, rLasso misses less number of true predictors. This may be due to that the tuning parameter λ of rLasso has an optimal order lower than $\log(p)$.

The rLasso has a very different variable selection strategy from Lasso. The rLasso prefers larger coefficients, while Lasso prefers smaller ones. When the data information is not sufficient for identifying the true model, i.e., $\log(p)/n$ is too large, rLasso may overfit the model by selecting a few predictors with large coefficients. These

predictors may be false, but highly correlated with the response variable or even the true predictors. However, Lasso may select many predictors with small coefficients. Consequently, Lasso lowers the risk of missing true predictors and reduces the coefficient estimation error. As for other methods, MCP and EBIC perform similarly to rLasso, and SCAD and Elastic Net perform similarly to Lasso. The results of Elastic Net are presented in Appendix.

In summary of the two studies, we conclude that the rLasso tends to select sparse models with low false selection rates, and always outperforms the other methods in terms of selection rates of true models. For a dataset with sufficiently many observations, rLasso will work similarly to EBIC and outperform the methods like MCP, LASSO, SCAD and Elastic Net. In the case that the number of observations is not sufficiently large, rLasso may overfit the data by selecting some false predictors with large coefficients. In contrast, Lasso, SCAD and Elastic Net may overfit the data by selecting many false predictors with very small coefficients; and MCP and EBIC may result in over-spare models.

3.4.3 Real Data Analysis

In this section, we evaluate the performance of the rLasso on a real PCR data set. [45] conducted an experiment that examines the genetics of two inbred mouse populations. A total of 60 F2 samples, with approximately half males and half females, were used to monitor the expression level of a total of 22,575 genes. Some physiological phenotypes, including the numbers of Phosphoenolpyruvate carboxykinase (PEPCK), Glycerol-3-phosphate acyltransferase (GPAT) and stearoyl-CoA desaturase 1 (SCD1), were measured by quantitative real-time PCR. The gene expression data and phenotype data can be accessed at GEO (http://www.nci.nih.gov/geo; accession number GSE3330). In this illustrative example, we study the linear relation-

ship between PEPCK and gene expression levels. For the simplicity of computation, only 2000 genes which have the highest marginal correlation with the output PEPCK are considered as predictors. Hence, we have $n = 60$ and $p = 2001$ (including the intercept term) for this example.

The predictors were first standardized to have zero mean and unit variance, and the response variable was also subtracted by its sample mean. We then applied the rLasso, EBIC, Lasso, MCP, SIS-SCAD and ISIS-SCAD to this example. In rLasso, we constrained the model size to be less than 30 and ran SAA for 5×10^5 iterations for each of the involved optimization tasks. In SIS and ISIS, we first reduced the number of predictors to be no more than 30 in the variable screening stage, and then applied SCAD. In EBIC, we employed the same value of $\lambda = 0.75$ as in the simulation studies.

To evaluate the performance of different methods, we randomly split the PCR dataset into two parts: 6 observations serving as the test data, and the rest 54 observations serving as the training data. For each method, we estimated the regression on the training data, where the tuning parameter is determined via a 10-fold cross-validation procedure, and then calculated the mean squared prediction error (MSPE) for the test data. Table 3.2 provides a summary for the MSPEs and selected model sizes resultant from 100 random splits of the PCR dataset. It is easy to see that all the six methods produced almost the same MSPEs, but rLasso produced a significantly sparser model than did others. On average, rLasso selected only one gene, while EBIC, MCP, SIS and ISIS selected at least 5 genes, and Lasso selected even more than 30 genes.

Later, we applied rLasso and all other five methods to the whole PCR dataset. The rLasso selected a model with only one gene. All other five methods selected much larger models, but all including the gene selected by rLasso. This example illustrates

Methods	rLasso	EBIC	Lasso	MCP	SIS	ISIS
MSPE	0.583	0.591	0.523	0.573	0.606	0.568
	(.037)	(.036)	(.026)	(.033)	(.033)	(.031)
Model Size	1.06	8.05	38.73	9.35	15.59	5.28
	(0.14)	(0.66)	(1.51)	(0.71)	(0.33)	(0.34)

Table 3.2: Results comparsion for real PCR data set. The table gives averages of MSPEs and selected model sizes over 100 random splits of the PCR dataset, where the numbers in parenthesis represent the standard deviations of the averages.

the power of rLasso for selecting sparse models under the high dimensional setting.

4. CONCLUSIONS AND DISCUSSIONS

In Section 2, we propose a new Bayesian variable selection approach, the so-called SaM approach, for ultra-high dimensional linear regression. The SaM approach works in two stages. Stage I screens out the predictors that are uncorrelated with the response variable, and stage II refines the selection of predictors. Both stages work under a Bayesian framework. Compared to the SIS approach of Fan and Lv (2008), a significant advantage of the SaM approach is that it makes use of the joint information of multiple predictors in predictor screening, while SIS makes use of only the marginal utility of each predictor. Our numerical results, on both simulated and real datasets, show that the SaM approach can significantly outperform SIS and ISIS, and also significantly outperforms other penalized likelihood approaches such as Lasso and elastic net. The models selected by SaM are more sparse and closer to the true model.

The SaM approach possesses an embarrassingly parallel structure and can be easily implemented in a parallel architecture. Therefore, the SaM approach is ready to be applied to the big data problems with millions or more of predictors. This has been beyond the ability of conventional Bayesian approaches which directly work on the whole dataset.

To justify the SaM approach, we establish the Bayesian posterior consistency under both situations that the model is correctly specified and misspecified. We show that when the model is correctly specified, the marginal inclusion probability of the true predictor will converge to 1 as the sample size becomes large; and when the model is misspecified, the marginal inclusion probability of the predictor that is correlated with the response will converge to 1 as the sample size becomes large.

We also show the sure screening property for the predictors selected based on the marginal inclusion probability and that the MAP model is consistent as an estimator of the true model.

In this work, the theoretical results are established under a Gaussian prior setting for the regression coefficients. Although this prior setting has been widely used in the literature, e.g., [34] and [52], it is not natural from the perspective of marginal inclusion probability evaluation as it assigns the highest density value to the null point. It is of interest to use a nonlocal prior [37], which assigns a zero probability to the null point and may lead to a better convergence rate of the posterior distribution. A further extension of the SaM approach to generalized linear models is also of great interest.

In Section 3, we have proposed a new class of penalty functions, the so-called rLasso penalty functions, for high dimensional variable selection. The new penalty functions are different from conventional penalty functions, such as those used in Lasso, SCAD, MCP, and EBIC, in that they are decreasing in $(0, \infty)$, discontinuous at 0, and converge to infinity as the coefficients approach zero. By giving small coefficients large penalties, rLasso brings sparsity into the model and successfully avoids to select over-dense models. Such over-denseness problem can occur in Lasso, SCAD and other methods that employ a Lasso-type penalty function which gives nearly zero coefficients nearly zero penalties.

Theoretically, we establish the consistency of the rLasso for variable selection and coefficient estimation under both the low and high dimensional settings. We also show that the optimal order of the tuning parameter of rLasso can be smaller than $\log(p)$, and thus rLasso can avoid to select over-sparse models when the ratio $\log(p)/n$ is large. Such over-sparsity problems can occur in EBIC and other methods that employ a L_0 or L_0-like penalty function.

The rLasso has been tested on simulated and real data examples. The numerical results indicate that the rLasso outperforms other methods, such as Lasso, SCAD, MCP, SIS, ISIS, Elastic Net, and EBIC: It can produce sparser and more accurate coefficient estimates and have a higher probability to catch true models, especially when the ratio $\log(p)/n$ is large.

The rationale of the rLasso penalty function can also be explained as follows: Traditional penalty functions, such as those used in LASSO, SCAD and MCP, are singular at zero and give zero the largest derivative value such that the coefficients of false predictors can shrink faster than those of the true predictors. The rLasso penalty function brings sparsity into models in a different way: By giving a very large penalty around zero such that the model cannot afford a small coefficient for the false predictor.

Extending the rLasso to other models, such as generalized linear models, survival models and Gaussian graphical models, is of great interest. Conceptually we did not see any difficulties in these extensions, although some mathematical work are needed. The class of rLasso penalty functions can also be extended to the form $1/|\beta|^2$ or a higher order form, which satisfy all the conditions (C_1)–(C_6) and just make the penalty function steeper around zero. We note that such high order penalty forms have also been suggested in Johnson and Rossell (2012, p.659), although not studied theoretically.

In addition to the rLasso method, we have also proposed a Monte Carlo optimization algorithm, which is a combination of the coordinate descent and stochastic approximation annealing algorithms, for solving the minimization problem involved in rLasso. Compared to other stochastic optimization algorithms such as simulated annealing, the new algorithm has a few advantages, e.g., fast convergence and less local traps. The new algorithm provides an efficient computational tool for rLasso

and thus other L_0-regularization methods.

REFERENCES

[1] H. Akaike. Information theory and an extension of maximum likelihood principle. In B.N. Petrov and F. Csaki, editors, *Second International Symposium of Information theory*, pages 267–281. Akademiai Kiadao, Budapest, 1973.

[2] H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.

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

[4] Z.D. Bai and Y.Q. Yin. Limit of smallest eigenvalue of a large dimensional sample covariance matrix. *Annals of Probability*, 21(3):1275–1294, 1993.

[5] M.M. Barbieri and J.O. Berger. Optimal predictive model selection. *Annals of Statistics*, 32(3):870–897, 2004.

[6] R. H. Berk. Limiting behaviour of posterior distributions when the model is incorrect. *Annals of Mathematical Statistics*, 37(1):51–58, 1966.

[7] P.J. Bickel and E. Levina. Some thoery for fisher's linear discriminant function, 'naive bayes', and some alternatives when there are many more variable than observations. *Bernoulli*, 10(6):989–1010, 2004.

[8] L. Bottolo and S. Richardson. Evolutionary stochastic search for bayesian model exploration. *Bayesian Analysis*, 5(3):583–618, 2010.

[9] P. Breheny and J. Huang. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics*, 5(1):232–243, 2011.

[10] K.W. Broman and T.P. Speed. A model selection approach for the identification of quantitative trait loci in experimental crosses. *Journal of the Royal Statistical Society: Series B*, 64(4):641–656, 2002.

[11] E. Candes and T. Tao. The dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351, 2007.

[12] J. Chen and Z. Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

[13] J. Chen and Z. Chen. Extended bic for small-n-large-p sparse glm. *Statistica Sinica*, 22(2):555–574, 2012.

[14] K. Davidson and S. Szarek. Local operator theory, random matrics and banach spaces. In W.B. Johnson and J. Lindenstrauss, editors, *Handbook on the Geometry of Banach Spaces*, pages 317–366. North-Holland, Amsterdam, 2001.

[15] P. De Blasi and S.G. Walker. Bayesian asymptotics with misspecified models. *Statistica Sinica*, 23(1):169–187, 2013.

[16] M.E. Dumas, S.P. Wilder, M.T. Bihoreau, R.H. Barton, J.F. Fearnside, K. Arqoud, L. D'Amato, R.H. Wallis, C. Blancher, H.C. Keun, D. Baunsqaard, J. Scott, U.G. Sidelmann, J.K. Nicholson, and D. Gauquier. Direct quantitative trait locus mapping of mammalian metabolic phenotypes in diabetic and normoglycemic rat models. *Nature Genetics*, 39(5):666–672, 2007.

[17] B. Efron, T. Hastie, I. Johnston, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

[18] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[19] J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society, Series B*, 70(5):849–911, 2008.

[20] J. Fan and R. Song. Sure independence screening in generalized linear model with np-dimensionality. *Annals of Statistics*, 38(6):3567–3604, 2010.

[21] Jianqing Fan and Jinchi Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101–148, 2010.

[22] I.E. Frank and J.H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–148, 1993.

[23] G. Freeman and J.Q. Smith. Bayesian map model selection of chain event graphs. *Journal of Multivariate Analysis*, 102(7):1152–1165, 2011.

[24] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[25] E. George and D. Foster. Calibration and empirical bayes variable selection. *Biometrika*, 87(4):731–747, 2000.

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

[27] C.J. Geyer. On the asymptotics of constrained m-estimation. *Annals of Statistics*, 22(4):1993–2010, 1994.

[28] C.J. Geyer. On the asymptotics of convex stochastic optimization. Unpublished manuscript, 1996.

[29] Y. Guan and M. Stephens. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Annals of Applied Statistics*, 5(3):1780–1815, 2011.

[30] M. Gupta. Model selection and sensitivity analysis for sequence pattern models. In N. Balakrishnan, E.A. Peña, and M.J. Silvapulle, editors, *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen*, volume 1, pages 390–407. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2008.

[31] M. Gupta and J.G. Ibrahim. An information matrix prior for bayesian analysis in generalized linear models with high dimensional data. *Statistica Sinica*, 19(4):1641–1663, 2009.

[32] H. Haario, E. Saksman, and J. Tamminen. An adaptive metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001.

[33] T. Inglot. Inequalities for quantiles of the chi-square distribution. *Probability and Mathematical Statistics*, 30(2):339–351, 2010.

[34] W. Jiang. Bayesian variable selection for high dimensional generalized linear models. Technical Report 05-02, Department of Statistics, Northwestern University, Evanston, IL, USA, 2005.

[35] W. Jiang. Bayesian variable selection for high dimensional generalized linear models: Convergence rate of the fitted densities. *Annals of Statistics*, 35(4):1487–1511, 2007.

[36] V.E. Johnson. On numerical aspects of bayesian model selection in high and ultrahigh-dimensional settings. *Bayesian Analysis*, 8(4):741–758, 2013.

[37] V.E. Johnson and D. Rossel. Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 107(498):649–660, 2012.

[38] A. Kindmark, A. Jawaid, C.G. Harbron, B.J. Barrat, O.F. Bengtsson, T.B. Andersson, S. Carlsson, K.E. Cederbrant, N.J. Gibson, M. Armstrong, M.E. Lagerström-Fermér, A. Dellsén, E.M. Brown, M. Thornton, C. Dukes, S.C. Jenkins, M.A. Firth, G.O. Harrod, T.H. Pinel, S.M.E. Billing-Clason, L.R. Cardon, and R.E. March. Genome-wide pharmacogenetic investigation of a hepatic adverse event without clinical signs of immunopathology suggests an underlying immune pathogenesis. *The Pharmacogenomics Journal*, 8(3):186–195, 2008.

[39] Scott Kirkpatrick. Optimization by simulated annealing: Quantitative studies. *Journal of Statistical Physics*, 34(5-6):975–986, 1984.

[40] B.J.K. Kleijn and A.W. van der Vaart. Misspecification in infinite-dimensional bayesian statistics. *Annals of Statistics*, 34(2):837–877, 2006.

[41] A. Kleiner, A. Talwalkar, P. Sarkar, and M.I. Jordan. A scalable bootstrap for massive data. *arXiv:1112.5016*, 2011.

[42] K. Knight. Epi-convergence in distribution and stochastic equi-semicontinuity. Unpublished manuscript, 1999.

[43] K Knight. Limiting distributions of linear programming estimators. *Extremes*, 4(2):87–103, 2001.

[44] K. Knight and W. Fu. Asymptotics for lasso-type estimators. *Annals of Statistics*, 28(5):1356–1378, 2000.

[45] H. Lan, M. Chen, J.B. Flowers, B.S. Yandell, D.S. Stapleton, C.M. Mata, E.T-K. Mui, M.T. Flowers, K.L. Schueler, K.F. Manly, R.W. Williams, C. Kendziorski, and A. D. Attie. Combined expression trait correlations and expression quantitative trait locus mapping. *PLoS Genetics*, 2(1):e6, 2006.

[46] M. Ledoux. Deviation inequalities on largest eigenvalues. In V. Milman and G. Schechtman, editors, *Geometric aspects of functional analysis*, volume 1910 of *Lecture Notes in Mathematics*, pages 167–219. Springer-Verlag, 2007.

[47] F. Li and N.R. Zhang. Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American Statistical Association*, 105(491):1202–1214, 2010.

[48] F. Liang. On the use of stochastic approximation monte carlo for monte carlo intergration. *Statistics & Probability Letters*, 79(5):581–587, 2009.

[49] F. Liang, Y. Cheng, and G. Lin. Simulated stochastic approximation annealing for global optimization with a square-root cooling schedule. *Journal of the American Statistical Association*, (in press), 2013.

[50] F. Liang, C. Liu, and R.J. Carroll. Stochastic approximation in monte carlo computation. *Journal of the American Statistical Association*, 102(477):305–320, 2007.

[51] F. Liang, R. Paulo, G. Molina, M.A. Clyde, and J.O. Berger. Mixture of g priors for bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423, 2008.

[52] F. Liang, Q. Song, and K. Yu. Bayesian subset modeling for high dimensional generalized linear models. *Journal of the American Statistical Association*, 108(502):589–606, 2013.

[53] F. Liang and J. Zhang. Estimating fdr under general dependence using stochastic approximation. *Biometrika*, 95(4):961–977, 2008.

[54] N. Lin and R. Xi. Aggregated estimating equation estimation. *Statistics and Its Interface*, 4(1):73–83, 2011.

[55] L. Mackey, A. Talwalkar, and M.I. Jordan. Divide-and-conquer matrix factorization. *arXiv:1107.0789*, 2012.

[56] T. Park and G. Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[57] G.CH. Pflug. On an argmax-distribution connected to the poisson process. In P. Mandl and M. Hušková, editors, *Asymptotic Statistics*, Contributions to Statistics, pages 123–129. Physica-Verlag, 1994.

[58] G.CH. Pflug. Asymptotic stochastic programs. *Mathematics of Operations Research*, 20(4):769–789, 1995.

[59] Nicholas G Polson and James G Scott. Local shrinkage rules, lévy processes and regularized regression. *Journal of the Royal Statistical Society: Series B*, 74(2):287–311, 2012.

[60] G.E. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

[61] J.G. Scott and J.O. Berger. Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *Annals of Statistics*, (5):2587–2619, 2010.

[62] C.R. Shalizi. Dynamics of bayesian updating with dependent data and misspecified models. *Electronic Journal of Statistics*, 3:1039–1074, 2009.

[63] J.W. Sliverstein. The smallest eigenvalue of a large dimensional wishart matrix. *Annals of Probability*, 13(4):1364–1368, 1985.

[64] J.D. Storey. A direct approach to false discovery rate. *Journal of the Royal Statistical Society, Series B*, 64(3):479–498, 2008.

[65] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

[66] Melanie A Wilson. *Bayesian Model Uncertainty and Prior Choice with Applications to Genetic Association Studies*. PhD thesis, Duke University, Durham, NC, USA, 2010.

[67] R. Xi, N. Lin, and Y. Chen. Compression and aggregation for logistic regression analysis in data cubes. *Knowledge and Data Engineering, IEEE Transactions on*, 21(4):479–492, 2009.

[68] A. Zellner. On assessing prior distributions and bayesian regression analysis with g-prior distribution. In P.K. Goel and A. Zellner, editors, *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, pages 233–243. North-Holland/Elsevier, 1986.

[69] C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38(2):894–942, 2010.

[70] C.-H. Zhang and J. Huang. The sparsity and bias of the lasso selection in high-dimensional regression. *Annals of Statistics*, 36(4):1567–1594, 2008.

[71] P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7(Nov):2541–2563, 2006.

[72] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

[73] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67(2):301–320, 2005.

[74] H. Zou and R. Li. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36(4):1509–1533, 2008.

APPENDIX A

PROOF OF THEOREMS IN SECTION 2

A.1 Proofs of Theorem 2.1.1 and Theorem 2.1.2

Let $\mathcal{P}$ denote a set of probability densities which form the support of the prior. In case of Theorem 2.1.1 and Theorem 2.1.2, $\mathcal{P} = \{f(x, y|\xi, \boldsymbol{\beta}_\xi, \sigma^2) : |\xi| \leq \bar{r}_n, \boldsymbol{\beta}_\xi \in \mathbb{R}^{|\xi|}, \sigma^2 > 0\}$. Let $\{\mathcal{P}_n\}$ be a sequence of subsets of $\mathcal{P}$, and let $N(\epsilon_n, \mathcal{P}_n)$ be the minimal number of Hellinger balls of radius ϵ_n to cover $\mathcal{P}_n$. Define $\hat{\pi}(\epsilon) = \pi[d(f, f^*) > \epsilon | D_n]$, define the Kullback-Leibler divergence $d_0(f, f^*) = \int f^* \ln(f^*/f)$, and define $d_t(f, f^*) = t^{-1}(\int f^*(f^*/f)^t - 1)$ for any $t > 0$. It is easy to see that d_t decreases to d_0 as $t \to 0$.

The next lemma follows from Theorem 6 of [34]:

Lemma A.1.1. *Assume that there is a sequence $\{\epsilon_n\} \in (0, 1]$ such that $n\epsilon_n^2 \succ 1$. If, for all sufficiently large n, the priors satisfy the following conditions:*

1. $\ln N(\epsilon_n/4, \mathcal{P}_n) \leq n\epsilon_n^2/16$;
2. $\pi(\mathcal{P}_n^c) \leq ne^{-n\epsilon_n^2/8}$;
3. $\pi[f : d_t(f, f^*) \leq \epsilon_n^2/64] \geq e^{-n\epsilon_n^2/64}$ *for some* $t > 0$, *or,*

3' *for all small enough* b, $r > 0$, *there exits* $N_{b,r}$ *such taht for all* $n > N_{b.r}$, $\pi[f : d_0(f, f^*) \leq b\epsilon_n^2] \geq e^{-rn\epsilon_n^2}$,

then under (a) (b) and (c), we have

$$P^*[\hat{\pi}(\epsilon_n) \geq 2e^{-n\epsilon_n^2 \min\{1/32, t/64\}}] \leq 2e^{-n\epsilon_n^2 \min\{1/32, t/64\}}, \text{ and}$$

$$E^*\hat{\pi}(\epsilon_n) \leq 4e^{-n\epsilon_n^2 \min\{1/16, t/32\}};$$

under (a), (b) and (c'), we have

$$\lim_{n\to\infty} P^*[\hat{\pi}(\epsilon_n) \geq 2e^{-n\epsilon_n^2 \min\{1/16, b/32\}}] = 0,$$

where E^* denotes the expectation with respect to f^*.

To prove Theorem 2.1.1, it suffices to show $E^*\pi[d(f, f^*) > \epsilon_n | D_n] \leq e^{-2c_1 n\epsilon_n^2}$, and then (2.14) is implied by Markov inequality. This can be done through a direct application of Lemma A.1.1 with a choice of $c_1 < \frac{1}{2}\min\{1/32, t/64\}$. Now we check the conditions of Lemma A.1.1.

Checking condition 3 for $t = 1$:

Let $h_\xi = x_\xi^T \boldsymbol{\beta}_\xi$ and $h^* = x^T\boldsymbol{\beta}^*$. Thus, h^* is the true conditional mean of y. A direct calculation shows that

$$d_1(f, f^*) = \begin{cases} \frac{\sigma^2}{\sigma^*\sqrt{2\sigma^2 - \sigma^{*2}}} \int \exp\{\frac{(h_\xi - h^*)^2}{2\sigma^2 - \sigma^{*2}}\}\nu_x(dx) - 1, & \text{if } (2\sigma^2 - \sigma^{*2}) > 0, \\ +\infty, & \text{otherwise,} \end{cases}$$

where $\sigma^2/(\sigma^*\sqrt{2\sigma^2 - \sigma^{*2}})$ is a concave and continuously differentiable function. Since ϵ_n is smaller than 1, there exists a sufficiently small η_1 such that whenever $\sigma^2 \in [\sigma^{*2}, \sigma^{*2} + \eta_1\epsilon_n^2]$, $\sigma^2/(\sigma^*\sqrt{2\sigma^2 - \sigma^{*2}}) \in [1, 1 + \epsilon_n^2/192]$ holds.

On the other hand, $|h_{\xi_n} - h^*| \leq |\sum_{j\notin\xi_n} x_j\beta_j^*| + |\sum_{j\in\xi_n} x_j(\beta_j - \beta_j^*)| \leq \Delta(r_n) + r_n \max_{j\in\xi_n}\{|\beta_j - \beta_j^*|\}$, where ξ_n is the model which achieves $\Delta(r_n)$ and so $|\xi_n| = r_n$. Again, there exists a small enough η_2 such that if $\beta_j \in (\beta_j^* \pm \eta_2\sigma\epsilon_n/r_n)$ for all $j \in \xi_n$,

$\sigma \geq \sigma^*$, then, by condition (2.7),

$$\int \exp\{\frac{(h_{\xi_n} - h^*)^2}{2\sigma^2 - \sigma^{*2}}\}\nu_x(dx) \leq \exp\left\{\frac{(\Delta(r_n) + r_n \max_{j\in\xi_n}\{|\beta_j - \beta_j^*|\})^2}{2\sigma^2 - \sigma^{*2}}\right\}$$
$$\leq \exp\{\eta_2^2\epsilon_n^2\} \in [1, 1 + \epsilon_n^2/192]. \tag{A.1}$$

Since $(1 + \epsilon_n^2/192)^2 \leq (1 + \epsilon_n^2/64)$, we conclude that $\{f : d_t(f, f^*) \leq \epsilon_n^2/64\} \supset \{\xi = \xi_n,\, \sigma^2 \in [\sigma^{*2}, \sigma^{*2} + \eta_1\epsilon_n^2],\, \beta_j \in (\beta_j^* \pm \eta_2\sigma\epsilon_n/r_n),\ \text{for } j \in \xi_n\}$. Following from the conditions of Theorem 2.1.1, $\bar{r}_n \geq r_n \geq 1$, $r_n/p_n \prec 1$ and $\bar{r}_n \ln(p_n) \prec n\epsilon_n^2$, the prior probability of ξ_n satisfies the inequality

$$-\ln(\pi(\xi = \xi_n))I(|\xi_n| \leq \bar{r}_n) \leq -r_n \ln(r_n/p_n) - (p_n - r_n)\ln(1 - r_n/p_n)$$
$$< r_n \ln(p_n) + r_n \prec n\epsilon_n^2.$$

Since $\bar{r}_n \ln(1/\epsilon_n^2) \prec n\epsilon_n^2$,

$$-\ln \pi\{\sigma^2 \in [\sigma^{*2}, \sigma^{*2} + \eta_1\epsilon_n^2]\} \leq -\ln[\eta_1\epsilon_n^2 \min_{s\in[\sigma^{*2}, \sigma^{*2}+\eta_1\epsilon_n^2]} d(s)]$$
$$= constant + \ln[1/\epsilon_n^2] \prec n\epsilon_n^2,$$

where $d(s)$ is the prior density of σ^2. Since r_n, $r_n \ln \bar{B}(r_n)$ and $B(r_n)$ are all of a smaller order than $n\epsilon_n^2$, by the same arguments as used in the proof of Theorem 1 of [34],

$$-\ln \pi\{\beta_j \in [\beta_j^* \pm \eta_2\epsilon_n\sigma/r_n]_{j\in\xi_n} | \xi = \xi_n, \sigma^2\} \prec n\epsilon_n^2.$$

In conclusion, $-\ln \pi[f : d_t(f, f^*) \leq \epsilon_n^2/64] \prec n\epsilon_n^2$, condition 3 meets.

Checking condition 1:

Let $\mathcal{P}_n = \{f(y; \xi, \boldsymbol{\beta}_\xi, \sigma^2) : |\xi| \leq \bar{r}_n, \sigma^2 \in (\underline{\sigma}_n^2, \overline{\sigma}_n^2), |\beta_j|_{j\in\xi} \leq C_n\sigma\}$, where $\underline{\sigma}_n^2$ and $\overline{\sigma}_n^2$ denote the lower and upper bounds of σ^2, respectively. The choices of $\underline{\sigma}_n^2$ and $\overline{\sigma}_n^2$

will be given in the next section of *checking condition 2*. For any $f_1,\ f_2 \in \mathcal{P}_n$,

$$d^2(f_1, f_2) = 1 - \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}} \int_x e^{-\frac{(h_1-h_2)^2}{4(\sigma_1^2+\sigma_2^2)}} \nu_x(dx). \tag{A.2}$$

If $\sigma_1^2 \le \sigma_2^2$ and $\sigma_2^2/\sigma_1^2 \le [1/(1-\epsilon_n^2/16)]^2$, then $\sqrt{2\sigma_1\sigma_2/(\sigma_1^2+\sigma_2^2)} \ge \sqrt{1-\epsilon_n^2/16}$. Further, if $|h_1 - h_2| \le \sigma_2\epsilon_n/\sqrt{8}$, then

$$\int_x e^{-\frac{(h_1-h_2)^2}{4(\sigma_1^2+\sigma_2^2)}} \nu_x(dx) \ge 1 - \sigma_2^2\epsilon_n^2/32(\sigma_1^2+\sigma_2^2) \ge 1 - \epsilon_n^2/32. \tag{A.3}$$

Thus $d^2(f_1, f_2) \le 1 - \sqrt{1-\epsilon_n^2/16}(1-\epsilon_n^2/32) \le \epsilon_n^2/16$.

We first partition $\mathcal{P}_n$ according to the models. There are at most p_n^r models of size $|\xi| = r$ for $r = 0, \ldots, \bar{r}_n$, and there are at most $(\bar{r}_n+1)p_n^{\bar{r}_n}$ different models in total. We partition the space of σ^2 to $\ln(\overline{\sigma}_n^2/\underline{\sigma}_n^2)/\{2\ln[1/(1-\epsilon_n^2/16)]\}$ intervals, with the ratio of each interval's upper bound and lower bound being $[1/(1-\epsilon_n^2/16)]^2$. Given a model ξ and $\sigma^2 \in [\sigma_1^2, \sigma_2^2]$, we furthermore partition the space of β_j $(j \in \xi)$ into at most $2\sqrt{8}C_n|\xi|/\epsilon_n$ intervals, with the interval length being equal to $\sigma_2\epsilon_n/(\sqrt{8}|\xi|)$.

Within each partition, $\sigma_2^2/\sigma_1^2 \le (1/1-\epsilon_n^2/16)^2$, $|h_1 - h_2| \le \sigma_2\epsilon_n/\sqrt{8}$, which means that each partition is a subset of a d-$(\epsilon_n/4)$ ball. Then

$$N(\epsilon_n/4, \mathcal{P}_n) \le (\bar{r}_n+1)p_n^{\bar{r}_n} \frac{\ln(\overline{\sigma}_n^2/\underline{\sigma}_n^2)}{2\ln[1/(1-\epsilon_n^2/16)]} \left(\frac{2\sqrt{8}C_n\bar{r}_n}{\epsilon_n}\right)^{\bar{r}_n},$$

and thus

$$\begin{aligned}\ln[N(\epsilon_n/4, \mathcal{P}_n)] \le & \ln(\bar{r}_n+1) + \bar{r}_n\ln(p_n) + \ln\ln\frac{\overline{\sigma}_n^2}{\underline{\sigma}_n^2} - \ln\ln\frac{1}{1-\epsilon_n^2/16} - \ln 2 \\ & + \bar{r}_n\ln 2\sqrt{8} + \bar{r}_n\ln C_n\bar{r}_n + \frac{1}{2}\bar{r}_n\ln\frac{1}{\epsilon_n^2}.\end{aligned}$$

By noting $-\ln\ln\frac{1}{1-\epsilon_n^2/16} \leq \ln(16/\epsilon_n^2) \prec n\epsilon_n^2$, choosing $C_n = \sqrt{\eta_3 \tilde{B}_n n\epsilon_n^2}$ for some constant η_3, and choosing $\ln\ln(\overline{\sigma}_n^2/\underline{\sigma}_n^2) = n\epsilon_n^2/32$, condition 1 can be verified.

Checking condition 2:

Because our prior distribution of ξ assigns zero probability for the models with size exceeding $\bar{r}_n$, $\pi(\mathcal{P}_n^c) \leq \pi[\sigma^2 \notin (\underline{\sigma}_n^2, \overline{\sigma}_n^2)] + \max_{\xi:|\xi|\leq\bar{r}_n} \pi(\cup_{j\in\xi}[|\beta_j| > C_n\sigma]|\xi,\sigma)$. By Mill's ratio, $\pi(|\beta_j| > C_n\sigma | j \in \xi) \leq 2e^{-C_n^2/(2\tilde{B}_n)}/\sqrt{2\pi C_n^2/\tilde{B}_n}$, which can be smaller than $e^{-n\epsilon_n^2/4}$ for sufficiently large n and $\eta_3 > 1/2$. If $\pi[\sigma^2 \notin (\underline{\sigma}_n^2, \overline{\sigma}_n^2)] < e^{-n\epsilon_n^2/4}$ also holds, then $\pi(\mathcal{P}_n^c) \leq (1+\bar{r}_n)e^{-n\epsilon_n^2/4} < e^{-n\epsilon_n^2/8}$ following from $\ln(\bar{r}_n+1) \prec \bar{r}_n \ln(p_n) \prec n\epsilon_n^2$.

Hence, we only need to show there exist $\underline{\sigma}_n^2$ and $\overline{\sigma}_n^2$ such that $\ln\ln(\overline{\sigma}_n^2/\underline{\sigma}_n^2) = n\epsilon_n^2/32$ and $\pi(\mathcal{P}_n^c) \leq \pi[\sigma^2 \notin (\underline{\sigma}_n^2, \overline{\sigma}_n^2)] < e^{-n\epsilon_n^2/4}$. Let $\underline{\sigma}_n^2 = (1/2)e^{-n\epsilon_n^2/4}$. Since the Inverse-Gamma distribution has a continuously differentiable density taking value 0 at the origin, $\pi(\sigma^2 \leq \underline{\sigma}_n^2) \leq (1/2)e^{-n\epsilon_n^2/4}$ for sufficiently large n. Let $\overline{\sigma}_n^2 = (1/2)e^{-n\epsilon_n^2/4}\exp[\exp(n\epsilon_n^2/32)]$. Then $\pi\{\sigma^2 \geq (1/2)e^{-n\epsilon_n^2/4}\exp[\exp(n\epsilon_n^2/32)]\} = \pi\{\sigma^{-2} \leq 2\exp[n\epsilon_n^2/4 - \exp(n\epsilon_n^2/32)]\}$. Since σ^{-2} follows a gamma distribution with the density function upper bounded by d_c around the origin,

$$\pi(\sigma^2 > \overline{\sigma}_n^2) \leq 2d_c e^{n\epsilon_n^2/4 - \exp(n\epsilon_n^2/32)} \leq \frac{1}{2}e^{n\epsilon_n^2/4 - \exp(n\epsilon_n^2/32) + \ln(4d_c)}.$$

Since $n\epsilon_n^2/4 - \exp(n\epsilon_n^2/32) + \ln(4d_c) < -n\epsilon_n^2/4$ for sufficiently large n, $\pi(\sigma^2 > \overline{\sigma}_n^2) < (1/2)e^{-n\epsilon_n^2/4}$. Therefore, condition 2 meets.

Proof of Theorem 2.1.2. To show the posterior consistency, we can again apply Lemma A.1.1 by verifying the three conditions. Most of the arguments in proving Theorem 2.1.1 is applicable here except for some minor modifications for accommodating the condition changes. To avoid redundant replications of the proof, we only

point out those minor modifications without giving the full proof.

Checking condition 3 under condition (A_1):

Inequality (A.1) still holds. Let $\xi = \{x'_1, \ldots, x'_{r_n}\}$ be any model with $|\xi| = r_n$ and containing all the true predictors. Let $\sigma^2 \in [\sigma^{*2}, \sigma^{*2} + \eta_1\epsilon_n^2]$ (same η_1 as in proving Theorem 2.1.1), and $\beta_j \in (\beta_j^* \pm \eta_2'\sigma\epsilon_n/r_n)$ for some η_2' (the choice of η_2' will be described later). The condition (A_1) implies that, for sufficiently small ϵ_n,

$$\int \exp\left\{\frac{(h_\xi - h^*)^2}{2\sigma^2 - \sigma^{*2}}\right\}\nu_x(dx) = \int \exp\left\{\frac{(\sum_{i=1}^{r_n}\Delta\beta_i x'_i)^2}{2\sigma^2 - \sigma^{*2}}\right\}\nu_x(dx) \quad (\Delta\beta_i = \beta_i - \beta_i^*)$$
$$= \int \exp\left\{\sum_{i=1}^{r_n}\left[\frac{\eta_2'\sigma\epsilon_n}{\delta\sqrt{2\sigma^2 - \sigma^{*2}}}\frac{\delta r_n\Delta\beta_i}{\eta_2'\sigma\epsilon_n}\frac{x'_i}{r_n}\right]^2\right\}\nu_x(dx) \le \exp\left\{M\left(\frac{\eta_2'\sigma}{\delta\sqrt{2\sigma^2 - \sigma^{*2}}}\right)^2\epsilon_n^2\right\}$$
$$= 1 + 2M\left(\frac{\eta_2'\sigma}{\delta\sqrt{2\sigma^2 - \sigma^{*2}}}\right)^2\epsilon_n^2.$$

Choose η_2' to be small enough such that $2M(\eta_2'\sigma)^2/[\delta^2(2\sigma^2 - \sigma^{*2})] < 1/192$, then we still have $\{f : d_t(f, f^*) \le \epsilon_n^2/64\} \supset \{\xi = \xi_n,\ \sigma^2 \in [\sigma^{*2}, \sigma^{*2} + \eta_1\epsilon_n^2],\ \beta_j \in (\beta_j^* \pm \eta_2'\sigma\epsilon_n/r_n),\ \text{for } j \in \xi_n\}$.

Checking condition 3' under condition (A_2):

By basic calculus, we have

$$d_0(f, f^*) = \ln\sigma - \ln\sigma^* + \frac{\sigma^{*2} - \sigma^2}{2\sigma^2} + \int\frac{(h_\xi - h^*)^2}{2\sigma^2}\nu_x(dx).$$

For any give b, there exsit η_1' which depends on b, such that when $\sigma^2/\sigma^{*2} \in [1, 1+\eta_1'\epsilon_n^2]$, $|\ln\sigma - \ln\sigma^* + (\sigma^{*2} - \sigma^2)/(2\sigma^2)| \le b\epsilon_n^2/2$. Furthermore, let $\beta_j \in (\beta_j^* \pm \eta_2'\sigma\epsilon_n/r_n)$, then

by condition (A_2)

$$\int \frac{(h_\xi - h^*)^2}{2\sigma^2}\nu_x(dx) = \int \frac{(\sum_{i=1}^{r_n}\Delta\beta_i x_i')^2}{2\sigma^2}\nu_x(dx)$$
$$= \int \sum_{i=1}^{r_n}\left[\frac{\eta_2'\epsilon_n}{\delta}\frac{\delta r_n \Delta\beta_i}{\eta_2'\sigma\epsilon_n}\frac{x_i'}{r_n}\right]^2 \nu_x(dx) \leq \frac{M\eta_2'^2}{\delta^2}\epsilon_n^2.$$

Choose η_2' small enough such that $M(\eta_2'/\delta)^2 < b/2$, then we have $\{f : d_0(f, f^*) \leq b\epsilon_n^2\} \supset \{\xi = \xi_n,\, \sigma^2 \in [\sigma^{*2}, \sigma^{*2} + \eta_1'\sigma^{*2}\epsilon_n^2],\, \beta_j \in (\beta_j^* \pm \eta_2'\sigma\epsilon_n/r_n),\ \text{for } j \in \xi_n\}$ and by the same arguments, $-\ln \pi(\{f : d_0(f, f^*) \leq b\epsilon_n^2\}) \prec n\epsilon_n^2$.

Checking condition 1 under condition (A_1) or (A_2):

We show that inequality (A.3) will still be valid under condition (A_1) or (A_2). Given a model $|\xi| \leq \bar{r}_n$, $\sigma_2^2/\sigma_1^2 \in [1, 1/(1-\epsilon_n^2/16)^2]$, and the element-wise difference in $\boldsymbol{\beta}_\xi$, $|\Delta\beta_i| \leq K\epsilon_n\sigma_2/|\xi|$, where $K^2 = \delta^2/8M$,

$$\int \exp\left\{-\frac{(h_1 - h_2)^2}{4\sigma_1^2 + 4\sigma_2^2}\right\}\nu_x(dx) = \int \exp\left\{-\frac{K^2\sigma_2^2\epsilon_n^2}{4\delta^2(\sigma_1^2+\sigma_2^2)}\sum\left(\frac{|\xi|\delta\Delta\beta_i}{\epsilon_n\sigma_2 K}\frac{x_i}{|\xi|}\right)^2\right\}\nu_x(dx)$$
$$\geq \exp\left\{-\frac{MK^2\sigma_2^2}{4\delta^2(\sigma_1^2+\sigma_2^2)}\epsilon_n^2\right\} > 1 - \frac{MK^2\sigma_2^2}{4\delta^2(\sigma_1^2+\sigma_2^2)}\epsilon_n^2 > 1 - \epsilon_n^2/32.$$

Thus, we can use the same partition method as that used in proving Theorem 2.1.1 to partition $\mathcal{P}_n$, which gives the covering number satisfying condition 1.

□

A.2 Proofs of Theorem 2.1.3 and Theorem 2.1.4

Proof of Theorem 2.1.3. Assume that a model ξ does not include the true predictor $\boldsymbol{x}_t$, f is conditional density corresponding to this mode ξ with some arbitrary $\boldsymbol{\beta}_\xi$ and

σ_2^2. Then, for the Hellinger distance between true f^* and f, we have

$$d^2(f^*, f) = 1 - \sqrt{\frac{2\sigma^*\sigma_2}{\sigma^{*2} + \sigma_2^2}} \int_x \exp\{-\frac{\Delta h^2}{4(\sigma^{*2} + \sigma_2^2)}\}\nu_x(dx),$$

where $\Delta h = \boldsymbol{\beta}_\xi^T x_\xi - \boldsymbol{\beta}_{\boldsymbol{t}}^{*T} x_{\boldsymbol{t}}$ is a linear function of the vector x. Recall that x represents a generic observation of p_n predictors, $x_{\boldsymbol{t}}$ denotes a generic observation of the true predictors, $t \in \boldsymbol{t}$, and x_t denotes a generic observation of the predictor $\boldsymbol{x}_t$. Since $t \in \boldsymbol{t}$ and $t \notin \xi$, we can write $\Delta h = -\beta_t^* x_t + \sum_{i \neq t} \Delta\beta_i x_i$. Thus, by condition (B_1),

$$\int_x \exp(-\Delta h^2/\beta_t^{*2})\nu_x(dx) \leq 1 - \delta_n.$$

If $4(\sigma^{*2} + \sigma_2^2)/\beta_t^{*2} \leq 1$, then

$$\int_x \exp\{-\frac{\Delta h^2}{4(\sigma^{*2} + \sigma_2^2)}\}\nu_x(dx) \leq 1 - \delta_n \text{ , and thus } d^2(f^*, f) \geq \delta_n.$$

If $4(\sigma^{*2} + \sigma_2^2)/\beta_t^{*2} > 1$, by Jensen's Inequality,

$$\int_x \exp\{-\frac{\Delta h^2}{4(\sigma^{*2} + \sigma_2^2)}\}\nu_x(dx) \leq (1 - \delta_n)^{\beta_t^{*2}/(4\sigma_2^2 + 4\sigma^{*2})} \leq 1 - \frac{\beta_t^{*2}}{4(\sigma_2^2 + \sigma^{*2})}\delta_n.$$

One can show that

$$\sqrt{\frac{2\sigma^*\sigma_2}{\sigma^{*2} + \sigma_2^2}} \int_x \exp\{-\frac{\Delta h^2}{4(\sigma^{*2} + \sigma_2^2)}\}\nu_x(dx) \leq \max\left\{\sqrt{\frac{4}{5}}, 1 - \frac{\beta_t^{*2}}{20\sigma^{*2}}\delta_n\right\},$$

which implies $d^2(f^*, f) \geq \min\{\delta_n, \beta_t^{*2}\delta_n/20\sigma^{*2}, 1 - \sqrt{0.8}\}$. Combining with the facts that $\delta_n \succ \epsilon_n^2$ and $\epsilon_n \to 0$, we conclude $d(f^*, f) > \epsilon_n$, i.e., all the models which do not include the true predictor $\boldsymbol{x}_t$ are outside the ϵ_n-ball of f^*. Because $P^*\{\pi[d(f, f^*) >$

$\epsilon_n|D_n] > e^{-c_1 n\epsilon_n^2}\} < e^{-c_1 n\epsilon_n^2}$,

$$P^*\{\pi[t \in \xi|D_n] < 1 - e^{-c_1 n\epsilon_n^2}\} < e^{-c_1 n\epsilon_n^2}.$$

□

Proof of Theorem 2.1.4. Under the eigen-structure condition of $\boldsymbol{X}$, V_ξ^{-1}'s eigenvalues are bounded by two values λ_1 and λ_2. For example, $\lambda_1 = l_n$, $\lambda_2 = l'_n$ in the case $V_\xi^{-1} = \boldsymbol{X}_\xi^T \boldsymbol{X}_\xi/n$; and $\lambda_1 = \lambda_2 = 1$ in the case $V_\xi^{-1} = I$. Define $R_\xi = y^T(I - \boldsymbol{X}_\xi(\boldsymbol{X}_\xi^T\boldsymbol{X}_\xi + V_\xi^{-1})^{-1}\boldsymbol{X}_\xi^T)y$ and $R_\xi^* = y^T(I - \boldsymbol{X}_\xi(\boldsymbol{X}_\xi^T\boldsymbol{X}_\xi)^{-1}\boldsymbol{X}_\xi^T)y$. Let $\hat{\boldsymbol{\beta}}_\xi$ denote the OLS estimator of $\boldsymbol{\beta}_\xi$. Then we have $R_\xi^* < R_\xi < R_\xi^* + \omega_\xi$, where $\omega_\xi = \lambda_2 \hat{\boldsymbol{\beta}}_\xi' \hat{\boldsymbol{\beta}}_\xi$.

For any model $\boldsymbol{k}$ and the true model $\boldsymbol{t}$,

$$\begin{aligned} \frac{\pi(\boldsymbol{k}|D_n)}{\pi(\boldsymbol{t}|D_n)} &\leq \left(\frac{r_n}{p_n - r_n}\right)^{|\boldsymbol{k}|-|\boldsymbol{t}|} \frac{K_2^{|\boldsymbol{t}\setminus\boldsymbol{k}|}}{K_1^{|\boldsymbol{k}\setminus\boldsymbol{t}|}} \left\{\frac{2b_0 + R_{\boldsymbol{t}}}{2b_0 + R_{\boldsymbol{k}}}\right\}^{n/2+a_0} \\ &\leq \left(\frac{r_n}{p_n - r_n}\right)^{|\boldsymbol{k}|-|\boldsymbol{t}|} \frac{K_2^{|\boldsymbol{t}\setminus\boldsymbol{k}|}}{K_1^{|\boldsymbol{k}\setminus\boldsymbol{t}|}} \left\{\frac{1 + 2b_0/R_{\boldsymbol{t}}^* + \omega_{\boldsymbol{t}}/R_{\boldsymbol{t}}^*}{R_{\boldsymbol{k}}^*/R_{\boldsymbol{t}}^*}\right\}^{n/2+a_0}, \end{aligned} \tag{A.4}$$

where $K_1 = K_2 = \sqrt{n+1}$ if $V_\xi^{-1} = \boldsymbol{X}_\xi^T\boldsymbol{X}_\xi/n$; and $K_1 = \sqrt{nl_n + 1}$, $K_2 = \sqrt{nl'_n + 1}$ if $V_\xi = I$.

Since $R_{\boldsymbol{t}}^*$ follows $\sigma^{*2}\chi^2_{n-|\boldsymbol{t}|}$, we have $R_{\boldsymbol{t}}^* = \sigma^{*2}n(1+\varepsilon_n)$ as n becomes large, where $\varepsilon_n = o_p(1)$, since $R_{\boldsymbol{t}}^*/(n - |\boldsymbol{t}|) \xrightarrow{a.s.} 1$. The OLS estimator $\hat{\boldsymbol{\beta}}_{\boldsymbol{t}}$ follows a normal distribution with mean $\boldsymbol{\beta}_{\boldsymbol{t}}^*$ and covariance matrix $(\boldsymbol{X}_{\boldsymbol{t}}^T\boldsymbol{X}_{\boldsymbol{t}})^{-1}$. Since the eigenvalue of $(\boldsymbol{X}_{\boldsymbol{t}}^T\boldsymbol{X}_{\boldsymbol{t}})^{-1}$ is smaller than $1/(nl_n)$, we have $\mathrm{Var}(\omega_{\boldsymbol{t}})$ is of smaller order than $\lambda_2^2/n^2l_n^2$ and $\omega_{\boldsymbol{t}} = \lambda_2 \sum_{i\in\boldsymbol{t}} |\beta_i^*|^2 + o_p(1)$. Therefore, $\{1 + 2b_0/R_{\boldsymbol{t}}^* + \omega_{\boldsymbol{t}}/R_{\boldsymbol{t}}^*\}^{n/2+a_0} = O_p(1)$.

For the rest terms, we first consider the case that $\boldsymbol{t} \subset \boldsymbol{k}$ and $|\boldsymbol{k}| - |\boldsymbol{t}| = d$. Thus

$$\frac{R_{\boldsymbol{t}}^*}{R_{\boldsymbol{k}}^*} = 1 + \frac{Z_d^2(\boldsymbol{k})}{R_{\boldsymbol{t}}^*/\sigma^{*2} - Z_d^2(\boldsymbol{k})},$$

where $Z_d^2(\boldsymbol{k})$ depends on model $\boldsymbol{k}$ and follows a χ_d^2 distribution, and $R_{\boldsymbol{t}}^*/\sigma^{*2} = n(1+\varepsilon_n)$. By Bonferroni inequality and the quantile estimation of χ^2 distribution (Theorem 4.1 of 33), we have, with probability $(1-e_1^d)$,

$$\max_{\boldsymbol{k},\boldsymbol{t}\subset\boldsymbol{k},|\boldsymbol{k}|=|\boldsymbol{t}|+d} Z_d^2(\boldsymbol{k}) \le 2d\log(p_n/e_1) + d + 2d\sqrt{\log(p_n/e_1)}.$$

Then it is easy to derive that

$$\left(\frac{R_{\boldsymbol{t}}^*}{R_{\boldsymbol{k}}^*}\right)^{n/2+a_0} \le \exp\left\{(1+\frac{2a_0}{n})[d\log(p_n/e_1) + d/2 + d\sqrt{\log(p_n/e_1)}]/(1+\varepsilon_n)\right\}.$$

Thus,

$$\begin{aligned}&\max_{\boldsymbol{k},\boldsymbol{t}\subset\boldsymbol{k},|\boldsymbol{k}|=|\boldsymbol{t}|+d} \left(\frac{r_n}{p_n-r_n}\right)^d \frac{1}{K_1^d}\left(\frac{R_{\boldsymbol{t}}^*}{R_{\boldsymbol{k}}^*}\right)^{n/2+a_0}\\ &\le \exp\left\{d\frac{(1+2\frac{a_0}{n})\{\log\frac{p_n}{e_1}+\frac{1}{2}+\sqrt{\log\frac{p_n}{e_1}}\}}{1+\varepsilon_n} + d\log\frac{r_n}{K_1(p_n-r_n)}\right\},\end{aligned}$$

which implies that with probability $1-\sum_{i=1}^{\bar{r}_n} e_1^i$ (which is greater than $1-2e_1$), uniformly for all $d \le r_n - |\boldsymbol{t}|$,

$$\begin{aligned}\max_{\boldsymbol{k},\boldsymbol{t}\subset\boldsymbol{k},|\boldsymbol{k}|=|\boldsymbol{t}|+d} \log\frac{\pi(\boldsymbol{k}|D_n)}{\pi(\boldsymbol{t}|D_n)} \le& d\frac{(1+2\frac{a_0}{n})\{\log\frac{p_n}{e_1}+\frac{1}{2}+\sqrt{\log\frac{p_n}{e_1}}\}}{1+\varepsilon_n} + d\log\frac{r_n}{K_1(p_n-r_n)}\\ &+O_p(1).\end{aligned}$$

By $\log(K_1/r_n) \succ \sqrt{\log(p_n)}$ and

$$\left[\frac{(1+2\frac{a_0}{n})\{\log\frac{p_n}{e_1}+\frac{1}{2}+\sqrt{\log\frac{p_n}{e_1}}\}}{1+\varepsilon_n}+\log\frac{r_n}{K_1(p_n-r_n)}\right]-\left[\log\frac{r_n}{K_1e_1}+\sqrt{\log\frac{p_n}{e_1}}+\frac{1}{2}\right] \xrightarrow{p} 0,$$

we conclude that $\sup_{t\subset k}\pi(\boldsymbol{k}|D_n)/\pi(\boldsymbol{t}|D_n) \xrightarrow{p} 0$.

Now we consider the case $\boldsymbol{k} \nsubseteq \boldsymbol{t}$. Let $\boldsymbol{u} = \boldsymbol{t} \cup \boldsymbol{k}$ and $d' = |\boldsymbol{u}| - |\boldsymbol{k}| = |\boldsymbol{t}\backslash\boldsymbol{k}| < |\boldsymbol{t}|$, then

$$\frac{R^*_{\boldsymbol{t}}}{R^*_{\boldsymbol{k}}} = \frac{R^*_{\boldsymbol{t}}}{R^*_{\boldsymbol{u}}}\frac{R^*_{\boldsymbol{u}}}{R^*_{\boldsymbol{k}}}.$$

It suffices to show

$$\max_{k\nsubseteq t}[R^*_{\boldsymbol{u}}/R^*_{\boldsymbol{k}}]^{n/2+a_0}(p_nK_2/r_n)^{d'} = o_p(1).$$

We know that $(R^*_{\boldsymbol{k}} - R^*_{\boldsymbol{u}})/\sigma^{*2}$ and $R^*_{\boldsymbol{u}}/\sigma^{*2}$ are independent. The former follows $\chi^2_{d'}(C)$, and the latter follows $\chi^2_{n-|\boldsymbol{u}|}$, where the noncentrality parameter $C = \boldsymbol{\beta}^{*T}_{\boldsymbol{t}}\boldsymbol{X}^T_{\boldsymbol{t}}(P_{\boldsymbol{X}_{\boldsymbol{u}}} - P_{\boldsymbol{X}_{\boldsymbol{k}}})\boldsymbol{X}_{\boldsymbol{t}}\boldsymbol{\beta}^*_{\boldsymbol{t}} = \boldsymbol{\beta}^{*T}_{\boldsymbol{t}}\boldsymbol{X}^T_{\boldsymbol{t}}(I - P_{\boldsymbol{X}_{\boldsymbol{k}}})\boldsymbol{X}_{\boldsymbol{t}}\boldsymbol{\beta}^*_{\boldsymbol{t}} \geq nl_nc^2$, $P.$ is the projection matrix, nl_n is the lower bound of the eigenvalues of $\boldsymbol{X}^T_{\boldsymbol{u}}\boldsymbol{X}_{\boldsymbol{u}}$, and $c = \min_{i\in\boldsymbol{t}}|\beta^*_i|$. Therefore,

$$\frac{R^*_{\boldsymbol{u}}}{R^*_{\boldsymbol{k}}} = \frac{R^*_{\boldsymbol{u}}}{R^*_{\boldsymbol{u}} + (R^*_{\boldsymbol{k}} - R^*_{\boldsymbol{u}})} \leq \frac{R^*_{\boldsymbol{t}}}{R^*_{\boldsymbol{t}} + (R^*_{\boldsymbol{k}} - R^*_{\boldsymbol{u}})}.$$

By Theorem 2.1 of [33], with probability $(1-2e_2)$,

$$\max_{k\nsubseteq t}(R^*_{\boldsymbol{k}} - R^*_{\boldsymbol{u}})/\sigma^2 > \left\{\sqrt{nl_nc^2} - \sqrt{2\log\left(\frac{\bar{r}_n p_n^{\bar{r}_n}}{e_2}\right)}\right\}^2.$$

Because $nl_n \succ \log \bar{r}_n + \bar{r}_n \log p_n$, we have $\max_{k \not\subseteq t}(R^*_{\boldsymbol{k}} - R^*_{\boldsymbol{u}})/\sigma^2 > nlc^2(1+o(1))$. Therefore, with probability $(1-2e_2)$,

$$\max_{k \not\subseteq t} \left(\frac{K_2 p_n}{r_n}\right)^{d'} \left(\frac{R^*_{\boldsymbol{u}}}{R^*_{\boldsymbol{k}}}\right)^{n/2+a_0}$$
$$\leq \exp\left\{-(\frac{n}{2}+a_0)\log\left[1 + l_n c^2 \frac{1+o(1)}{1+\varepsilon_n}\right] + d'\log p_n - d'\log r_n + d'\log K_2\right\}.$$

Since $nl_n \succ \log p_n + \log K_2$, we conclude that $\max_{k \not\subseteq t}[R^*_{\boldsymbol{u}}/R^*_{\boldsymbol{k}}]^{n/2+a_0} = o_p\{(r_n/K_2 p_n)^{d'}\}$.

In summary, we have $\max_{k \neq t} \pi(\boldsymbol{k}|D_n)/\pi(\boldsymbol{t}|D_n) \xrightarrow{p} 0$. □

Sketch proof of the remark under Theorem 2.1.4. Here we consider the choice $V_\xi^{-1} = (\boldsymbol{X}_\xi^T \boldsymbol{X}_\xi + \tau I)/n$, and define R_ξ as in the proof of Theorem 2.1.4. For any model ξ with size $|\xi| = k$, we denote $\boldsymbol{X}_\xi^T \boldsymbol{X}_\xi$'s eigenvalues by $l_n \leq L_1 \leq \cdots \leq L_k \leq l'_n$. Then

$$\frac{det(V_\xi^{-1})}{det(\boldsymbol{X}_\xi^T \boldsymbol{X}_\xi + V_\xi^{-1})} = \prod_{i=1}^{k} \frac{nL_k + \tau}{n(n+1)L_k + \tau} > \left(\frac{1}{n+1}\right)^k.$$

Since $nl_n \succ \bar{r}_n$,

$$(n+1)^k \prod_{i=1}^{k} \frac{nL_k + \tau}{n(n+1)L_k + \tau} < \left(1 + \frac{\tau}{(n+1)l_n + \tau/n}\right)^{\bar{r}_n} < C,$$

for some positive C. Therefore,

$$\frac{\pi(\boldsymbol{k})}{\pi(\boldsymbol{t})} \leq C \left(\frac{r_n}{p_n - r_n}\right)^{|\boldsymbol{k}|-|\boldsymbol{t}|} (n+1)^{(|\boldsymbol{t}|-|\boldsymbol{k}|)/2} \left\{\frac{2b_0 + R_{\boldsymbol{t}}}{2b_0 + R_{\boldsymbol{k}}}\right\}^{-n/2-a_0},$$

which has the same form as equation(A.4). Thus, the proof of Theorem 2.1.4 is applicable here. □

A.3 Proof of Theorem 2.1.5

In the case that the subset does not contain all true predictors, the model is misspecified. In the frequentist approach, we expect that the estimator of f^* to converge to the minimum point of the Kullback-Leibler divergence. We expect the same result in the Bayesian analysis.

To show the posterior consistency for the misspecified models, we follow the work by [15]. The following lemma is Corollary 1 of [15].

Lemma A.3.1. *Let $\mathcal{P}_s$ be the support of the prior with the true density $f^* \notin \mathcal{P}_s$, and let $\mathcal{P}_1$ be the set of all the densities associated with the minimum Kullback-Leibler divergence. Assume $\mathcal{P}_1$ is not empty. If the prior π satisfies the condition*

$$\pi\left\{f \in \mathcal{P}_s : d_0(f, f^*) < d_0(f_0, f^*) + \eta\right\} > 0, \quad \text{for any } \eta > 0, \tag{A.5}$$

and for all $\rho \in (0, 1/2)$, $\epsilon_\rho = 2(\rho^2/2)^{1/2\rho}$, there exists a sequence of sets $(B_{j,\epsilon_\rho})_{j\geq 1}$, each of which is inside Hellinger balls of size ϵ_ρ, that cover $\mathcal{P}_s$ such that

$$\sum_j \pi(B_{j,\epsilon_\rho})^\rho < \infty, \tag{A.6}$$

then, as $n \to \infty$, $\pi(\{f : d(\mathcal{P}_1, f) > \epsilon\}|D_n^{\boldsymbol{s}}) \to 0$ almost surely for any positive ϵ.

The Kullback-Leibler divergence between the true distribution f^* and the distribution of a model in $\mathcal{P}_s$ is given by

$$d_0(f, f^*) = \int \ln(f^*/f) f^* = \ln\sigma - \ln\sigma^* + \frac{\sigma^{*2} - \sigma^2}{2\sigma^2} + \int \frac{(\boldsymbol{\beta}^{*T}x - \boldsymbol{\beta}^T x_{\boldsymbol{s}})^2}{2\sigma^2} d\nu_x(dx). \tag{A.7}$$

To minimize (A.7), the optimized values of $\boldsymbol{\beta}_0$ and σ_0 are $\boldsymbol{\beta}_0 = \arg\min E(\boldsymbol{\beta}^T x_{\boldsymbol{s}} - \boldsymbol{\beta}^{*T}x)^2$ and $\sigma_0^2 = \arg\min\{\ln(\sigma) + \sigma^{*2}/2\sigma^{*2} + E(\boldsymbol{\beta}_0^T x_{\boldsymbol{s}} - \boldsymbol{\beta}^{*T}x)^2/2\sigma^2\} = \sigma^2 + E(\boldsymbol{\beta}_0^T x_{\boldsymbol{s}} -$

$\boldsymbol{\beta}^{*T}x)^2$. Thus, $\boldsymbol{\beta}_0$ is uniquely determined by the covariance structure, and so does σ_0. Let f_0 denote the unique density which has the minimum Kullback-Leibler divergence, $\mathcal{P}_1 = \{f_0\}$. Thus, $f_0 = \phi(y; \boldsymbol{\beta}_0^T x_{\boldsymbol{s}}, \sigma_0)\nu_x(x_{\boldsymbol{s}})$, where $\phi(y; \mu, \sigma)$ denotes a normal density with mean μ and standard deviation σ.

Now we apply Lemma A.3.1 to show that the posterior is consistent with the minimized Kullback-Leibler divergence density f_0. From the formula of the Kullback-Leibler divergence (A.7), it is easy to see that the K-L divergence is continuous with respect to the parameters $\boldsymbol{\beta}$ and σ, which implies that the condition (A.5) is satisfied. To verify condition (A.6), we first divide $\mathcal{P}_s$ by models into $\{\mathcal{P}_s^\xi\}$, where $\mathcal{P}_s^\xi$ denotes a subset of $\mathcal{P}_s$ corresponding to the model ξ. Since the number of predictors s is fixed, there are only a fixed number of possible models, and $\pi(\mathcal{P}_s^\xi)$ is bounded away from zero for any model ξ. For each model ξ, we cover the whole $\mathcal{P}_s^\xi$ by $B^\xi_{m,m_1,\ldots,m_{|\xi|}} = \{\xi, \sigma^2 \in [s_\rho^m, s_\rho^{m+1}), \beta_i \in [C_\xi\sqrt{s_\rho^m} \times m_i, C_\xi\sqrt{s_\rho^m} \times (m_i+1)), i = 1, \ldots, |\xi|\}$, where $s_\rho = (1-\epsilon_\rho^2)^{-2} > 1$ and $C_\xi = C/|\xi|$ for some constant C, Let m, m_1, $\ldots$, $m_{|\xi|}$ take all integer values in $\mathbb{Z}$ such that

$$\bigcup_{m,m_1,\ldots,m_{|\xi|}} B^\xi_{m,m_1,\ldots,m_{|\xi|}} = \mathcal{P}_s^\xi.$$

By condition A_1', following the same arguments as used in verifying condition 1 in the proof of Theorem 2.1.2, each of these small sets are inside a ϵ_ρ-Hellinger ball and the union of these sets covers $\mathcal{P}_s^\xi$. Without losing generality, we consider $m_1, \ldots, m_{|\xi|} \geq 0$,

which stands for one of the $2^{|\xi|}$ subparts of the $|\xi|$-dimension space of $\boldsymbol{\beta}_{|\xi|}$,

$$\begin{aligned}
\sum_{m_1,\ldots,m_{|\xi|}\geq 0} \pi(B^{\xi}_{m,m_1,\ldots,m_{|\xi|}})^{\rho} &< \pi(\xi)^{\rho}\pi_{IG}(\sigma^2 \in [s_{\rho}^{m}, s_{\rho}^{m+1}))^{\rho} \\
&\times \sum_{m_1,\ldots,m_{|\xi|}\geq 0} \pi_N(\beta_i \in [C_{\xi}\sigma m_i/\sqrt{s_{\rho}}, C_{\xi}\sigma \times (m_i+1)), i=1,\ldots,|\xi|\big|\sigma)^{\rho} \\
&= \pi(\xi)^{\rho}\pi_{IG}(\sigma^2 \in [s_{\rho}^{m}, s_{\rho}^{m+1}))^{\rho} \sum_{m_1,\ldots,m_{|\xi|}\geq 0} \left(\int_{\Omega_{m_1,\ldots,m_{|\xi|}}} \phi_V(\boldsymbol{\beta})d\boldsymbol{\beta}\right)^{\rho},
\end{aligned}$$

where π_{IG} denotes the Inverse Gamma prior of σ^2, π_N denotes the conditional Gaussian prior of $\boldsymbol{\beta}$ given σ, $\Omega_{m,m_1,\ldots,m_{|\xi|}}$ denote the set $\{\beta_i \in [C_{\xi}m_i/\sqrt{s_{\rho}}, C_{\xi}(m_i+1)), i = 1,\ldots,|\xi|\}$ and ϕ_V is the normal density with mean 0, covariance matrix V_{ξ}.

Let $\boldsymbol{\beta}_{m_1,\ldots,m_{|\xi|}} = (\beta_i = C_{\xi}m_i/\sqrt{s_{\rho}})_{i=1}^{|\xi|}$ which is the nearest point to the origin in set $\Omega_{m,m_1,\ldots,m_{|\xi|}}$. And let $\psi_{|\xi|}(\cdot)$ be the $|\xi|-$dimension Gaussian density function corresponding to mean zero, covariance matrix $\lambda_0 I_{|\xi|}$, where $\lambda_0 > ch_1(V_{\xi})$, thus $\psi_{|\xi|}(\boldsymbol{\beta}) > \phi_V(\boldsymbol{\beta})$ for large $|\boldsymbol{\beta}|$. There exist some large N, such that,

$$\begin{aligned}
&\sum_{m_1,\ldots,m_{|\xi|}\geq N} \left(\int_{\Omega_{m_1,\ldots,m_{|\xi|}}} \phi_V(\boldsymbol{\beta})d\boldsymbol{\beta}\right)^{\rho} \\
&< C_{\xi}^{|\xi|\rho} \sum_{m_1,\ldots,m_{|\xi|}\geq N} \prod_{i=1}^{|\xi|} (1 + m_i(1/\sqrt{s_{\rho}} - 1))^{\rho}\psi_{|\xi|}(\boldsymbol{\beta}_{m_1,\ldots,m_{|\xi|}})^{\rho} \\
&= C_{\xi}^{|\xi|\rho} \left(\sum_{m_1\geq N} (1 + m_1(1/\sqrt{s_{\rho}} - 1))^{\rho}\psi_1(C_{\xi}m_1/\sqrt{s_{\rho}})^{\rho}\right)^{|\xi|} \\
&< \infty,
\end{aligned}$$

since $\sum_{m_1\geq N}(1+m_1)\exp\{-(C_{\xi}^2 m_1^2\rho/(2s_{\rho}\lambda_0))\} < \infty$. Hence,

$$\sum_{m_1,\ldots,m_{|\xi|}\geq 0} \left(\int_{\Omega_{m_1,\ldots,m_{|\xi|}}} \phi_V(\boldsymbol{\beta})d\boldsymbol{\beta}\right)^{\rho} < C_2,$$

for some constant C_2. Then for any sufficiently large N, there exists a constant C_2' such that

$$\begin{aligned}
&\sum_{m>N}\sum_{m_1,\ldots,m_\xi\geq 0}\pi(B^\xi_{m,m_1,\ldots,m_{|\xi|}})^\rho < C_2\pi(\xi)^\rho\sum_{m>N}\pi_{IG}(\sigma^2\in[s_\rho^m,s_\rho^{m+1}))^\rho\\
<&C_2'\sum_{m>N}\left(\int_{s_\rho^m}^{s_\rho^{m+1}}x^{-a_0-1}dx\right)^\rho = C_2'a_0^{-1}\sum_{m>N}\left(s_\rho^{-a_0m}-s_\rho^{-a_0(m+1)}\right)^\rho\\
=&C_2'a_0^{-1}(1-s_\rho^{-a_0})^\rho\sum_{m>N}s_\rho^{-a_0\rho m}<\infty.
\end{aligned}$$

On the other hand, for any sufficiently large N', there exist some postive constants a_1 and C_1' such that

$$\begin{aligned}
&\sum_{m<-N'}\sum_{m_1,\ldots,m_{|\xi|}\geq 0}\pi(B^\xi_{m,m_1,\ldots,m_\xi})^\rho < C_2\pi(\xi)^\rho\sum_{m<-N'}\pi_{IG}(\sigma^2\in[s_\rho^m,s_\rho^{m+1}))^\rho\\
<&C_1'\sum_{-m>N'}\left(\int_{s_\rho^{-m-1}}^{s_\rho^{-m}}\exp(-b_0x/2)dx\right)^\rho < C_1'\sum_{m>N'}\left(\int_{s_\rho^{m-1}}^{s_\rho^m}x^{-a_1-1}dx\right)^\rho\\
=&C_1'a_1^{-1}(1-s_\rho^{-a_1})^\rho\sum_{m\geq N'}s_\rho^{-a_1\rho m}<\infty.
\end{aligned}$$

Since there are only a finite number of possible models, we conclude that for any $\rho\in(0,1/2)$,

$$\sum_\xi\sum_{m,m_1,\ldots,m_{|\xi|}}\pi(B^\xi_{m,m_1,\ldots,m_{|\xi|}})^\rho<\infty.$$

Hence, the posterior is consistent.

APPENDIX B

PROOF OF THEOREMS IN SECTION 3

B.1 Proof of Theorem 3.1.1

Proof. We first prove the normality part. Following the proof of Theorem 2 of [44], we define

$$V_n(\boldsymbol{u}) = \sum_i^n [(\epsilon_i - \boldsymbol{u}^T \boldsymbol{x}_i/\sqrt{n})^2 - \epsilon_i^2] + \sum_{j=1}^p \left(P_\lambda(\beta_j^* + u_j/\sqrt{n}) - P_\lambda(\beta_j^*) \right),$$

where ϵ_i denotes the ith element of $\boldsymbol{\epsilon}$ as defined in (1.1). It is easy to see that $V_n(\boldsymbol{u})$ is minimized at $\hat{\boldsymbol{u}}_n = \sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$.

For the first term of $V_n(\boldsymbol{u})$, we have

$$I = \sum_i^n [(\epsilon_i - \boldsymbol{u}^T \boldsymbol{x}_i/\sqrt{n})^2 - \epsilon_i^2] = \boldsymbol{u}^T \left(\frac{1}{n} \boldsymbol{X}^T \boldsymbol{X} \right) \boldsymbol{u} - 2 \frac{\boldsymbol{\epsilon}^T \boldsymbol{X}}{\sqrt{n}} \boldsymbol{u} \xrightarrow{d} \boldsymbol{u}^T \Sigma \boldsymbol{u} - 2 \boldsymbol{u}^T \boldsymbol{W},$$

where Σ is defined in (3.1) and $\boldsymbol{W}$ is a normal random vector with the covariance matrix $\sigma^2 \Sigma$. It is also easy to show that

$$I \xrightarrow{e-d} \boldsymbol{u}^T \Sigma \boldsymbol{u} - 2 \boldsymbol{u}^T \boldsymbol{W},$$

by Theorem 5 of [42], where $\xrightarrow{e-d}$ denotes epi-convergence in distribution for a sequence of random lower-semicontinuous functions.

For the second term of $V_n(\boldsymbol{u})$, we have for any compact set $\boldsymbol{U} \in \mathbb{R}^p$,

$$II = P_\lambda(\beta_j^* + u_j/\sqrt{n}) - P_\lambda(\beta_j^*) \rightrightarrows \begin{cases} P_\lambda(u_j/\sqrt{n}) \to \infty, & \text{if } \beta_j^* = 0, \quad u_j \neq 0, \\ 0, & \text{if } \beta_j^* = 0, \quad u_j = 0, \\ u_j/\sqrt{n} P'_\lambda(\beta_j^*) \to 0, & \text{if } \beta_j^* \neq 0, \end{cases} \tag{B.1}$$

where $\rightrightarrows$ denotes uniformly convergence. Let

$$V(\boldsymbol{u}) = \begin{cases} -2\boldsymbol{u}_{\boldsymbol{t}}^T \boldsymbol{W}_{\boldsymbol{t}} + \boldsymbol{u}_{\boldsymbol{t}}^T \Sigma_{\boldsymbol{t}} \boldsymbol{u}_{\boldsymbol{t}}, & \text{if } u_j = 0 \ \forall j \notin \boldsymbol{t}, \\ \infty, & \text{Otherwise.} \end{cases}$$

Then, by Lemma 1 of [58], we have $V_n(\boldsymbol{u}) \xrightarrow{e-d} V(\boldsymbol{u})$, and $V(\boldsymbol{u})$ has the unique minimum $\hat{\boldsymbol{u}} = (\Sigma_{\boldsymbol{t}}^{-1} \boldsymbol{W}_{\boldsymbol{t}}, 0)^T$.

To show $\hat{\boldsymbol{u}}_n \to^d \hat{\boldsymbol{u}}$, it is sufficient to show that $\hat{\boldsymbol{u}}_n = O_p(1)$ (see Theorem 1 of [42]), where $O_p(1)$ denotes bounded in probability. Note that

$$V_n(\boldsymbol{u}) \geq \boldsymbol{u}^T \left(\frac{1}{n} \boldsymbol{X}^T \boldsymbol{X} \right) \boldsymbol{u} - 2 \frac{\boldsymbol{\epsilon}^T \boldsymbol{X}}{\sqrt{n}} \boldsymbol{u} - \sum_{j=1}^{p} P_\lambda(\beta_j^*) = \tilde{V}_n(\boldsymbol{u}).$$

Since $0 = V_n(0) \geq V_n(\hat{\boldsymbol{u}}_n) \geq \tilde{V}_n(\hat{\boldsymbol{u}}_n)$, $\tilde{V}_n(\boldsymbol{u})$ is convex, $\arg\min(\tilde{V}_n(\boldsymbol{u})) = O_p(1)$, and the eigenvalues of $\boldsymbol{X}^T \boldsymbol{X}/n$ is $O_p(1)$, it follows that $\hat{\boldsymbol{u}}_n = O_p(1)$. For more details of epi-convergence in distribution and limiting distribution of argmin estimators, see [57, 58], [27, 28] and [42, 43].

We now prove the model consistency part. For any $j \in \boldsymbol{t}$, the asymptotic normality result implies that $\hat{\beta}_j \xrightarrow{p} \beta_j^*$; which implies

$$P(j \in \boldsymbol{\xi}(\hat{\boldsymbol{\beta}}_n) | j \in \boldsymbol{t}) \to 1, \quad \text{as } n \to \infty. \tag{B.2}$$

For any $j \notin \boldsymbol{t}$, the asymptotic normality result implies $P(|(\hat{\boldsymbol{u}}_n)_j| < \delta) \to 1$ for any $\delta > 0$. In addition, we have

$$
\begin{aligned}
&P\{(\hat{\boldsymbol{u}}_n)_j \neq 0, |(\hat{\boldsymbol{u}}_n)_j| < \delta\} \leq P\{\inf_{(\hat{\boldsymbol{u}}_n)_j \neq 0} V_n(\boldsymbol{u}) \leq V_n(0), |(\hat{\boldsymbol{u}}_n)_j| < \delta\} \\
<&P\left\{-\left(\frac{\boldsymbol{\epsilon}^T\boldsymbol{X}}{\sqrt{n}}\right)^T\left(\frac{\boldsymbol{X}^T\boldsymbol{X}}{n}\right)^{-1}\left(\frac{\boldsymbol{\epsilon}^T\boldsymbol{X}}{\sqrt{n}}\right) - \sum_{j=1}^{p} P_\lambda(\beta_j^*) + P_\lambda(\delta/\sqrt{n}) \leq 0\right\} \\
\to&P\left\{P_\lambda(\delta/\sqrt{n}) - \sum_{j=1}^{p} P_\lambda(|\beta_j^*|) \leq \frac{1}{\sigma^2}\chi_p^2\right\} \to 0, \quad \text{as } n \to \infty,
\end{aligned}
$$

where the last row follows from the asymptotics $\left(\frac{\boldsymbol{\epsilon}^T\boldsymbol{X}}{\sqrt{n}}\right)^T\left(\frac{\boldsymbol{X}^T\boldsymbol{X}}{n}\right)^{-1}\left(\frac{\boldsymbol{\epsilon}^T\boldsymbol{X}}{\sqrt{n}}\right) \xrightarrow{d} \chi_p^2/\sigma^2$ (by Slutsky's theorem and continuous mapping theorem). Note that in this case, we have $\beta_j^* = 0$ and $P_\lambda(\delta/\sqrt{n}) \to \infty$. Therefore,

$$
\begin{aligned}
P\{(\hat{\boldsymbol{u}}_n)_j = 0\} &\geq P\{(\hat{\boldsymbol{u}}_n)_j = 0, |(\hat{\boldsymbol{u}}_n)_j| < \delta\} \\
&= P(|(\hat{\boldsymbol{u}}_n)_j| < \delta) - P\{(\hat{\boldsymbol{u}}_n)_j \neq 0, |(\hat{\boldsymbol{u}}_n)_j| < \delta\} \to 1,
\end{aligned}
$$

which implies

$$
P(j \notin \boldsymbol{\xi}(\hat{\boldsymbol{\beta}}_n) | j \notin \boldsymbol{t}) \to 1, \quad \text{as } n \to \infty. \tag{B.3}
$$

The consistency of the model selection can then be concluded by combining (B.2) and (B.3). □

B.2 Proof of Theorem 3.2.1

To prove Theorem 3.2.1, we first prove the following lemma.

Lemma B.2.1. *Considering the linear regression (1.1) and the following model se-*

lection criterion

$$\hat{\boldsymbol{\beta}} = \arg \min_{|\boldsymbol{\xi}(\boldsymbol{\beta})| \leq r} \{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p} P_\lambda(|\beta_j|)\}, \tag{B.4}$$

where $\boldsymbol{\xi}(\boldsymbol{\beta}) = \{i : \beta_i \neq 0\}$ *denotes the model corresponding to the vector* $\boldsymbol{\beta}$, $|\boldsymbol{\xi}(\boldsymbol{\beta})|$ *denote the size of the model* $\boldsymbol{\xi}(\boldsymbol{\beta})$, *and each column of* $\boldsymbol{X}$ *has been standardized such that* $\|\boldsymbol{x}_i\| = \sqrt{n}$ *for* $i = 1, \ldots, p$. *Suppose that the following conditions are satisfied*

(a) $|\boldsymbol{t}| \leq r < n - |\boldsymbol{t}|$;

(b) for any subset model $\boldsymbol{\zeta}$,

$$nl_* \leq \min_{|\boldsymbol{\zeta}| \leq |\boldsymbol{t}|+r} ch_1(\boldsymbol{X}_{\boldsymbol{\zeta}}^T \boldsymbol{X}_{\boldsymbol{\zeta}}) \leq \max_{|\boldsymbol{\zeta}| \leq |\boldsymbol{t}|+r} ch_1'(\boldsymbol{X}_{\boldsymbol{\zeta}}^T \boldsymbol{X}_{\boldsymbol{\zeta}}) \leq nl^*;$$

(c) $P_\lambda\left(2\sqrt{\frac{2\sigma^2(|\boldsymbol{t}|+1)\log(p/e_1)}{nl_*^2} + \frac{|\boldsymbol{t}|a_\lambda}{nl_*}}\right) \geq \sigma^2(2\log(p/e_2) + 1 + 2\sqrt{\log(p/e_2)}) + |\boldsymbol{t}|a_\lambda$;

(d) $\sqrt{nl_*\underline{\beta}^2} - \sigma\sqrt{2\log(rp^r/e_3)} \geq \sqrt{\sigma^2(2r\log\frac{p}{e_2} + r + 2r\sqrt{\log(p/e_2)}) + |\boldsymbol{t}|(c_\lambda + a_\lambda)}$;

(e) $b_\lambda \leq \underline{\beta} - \sigma\frac{\sqrt{2\log(1/e_4)}}{nl_*}$;

where $\boldsymbol{t}$ *denotes the true model,* $|\boldsymbol{t}|$ *denotes the size of* $\boldsymbol{t}$, e_1, e_2, e_3 *and* e_4 *are sufficiently small numbers,* $\underline{\beta} = \min_{i \in \boldsymbol{t}} \beta_i^*$, *and* $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)$ *denotes the true regression coefficient vector. Then*

$$Pr\left(\boldsymbol{\xi}(\hat{\boldsymbol{\beta}}) = \boldsymbol{t}, \|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^o\| \leq \sqrt{|\boldsymbol{t}|a_\lambda/nl_*}\right) > 1 - 2e_1 - 4e_2 - 2e_3 - 2|\boldsymbol{t}|e_4, \tag{B.5}$$

where $\hat{\boldsymbol{\beta}}^o$ *is equal to* $\hat{\boldsymbol{\beta}}_{\boldsymbol{t}}^o$ *for the components corresponding to the model* $\boldsymbol{t}$ *and 0 otherwise, and* $\hat{\boldsymbol{\beta}}_{\boldsymbol{t}}^o$ *is the OLS estimator of* $\boldsymbol{\beta}_{\boldsymbol{t}}$. *Furthermore, we have the following*

upper bound for the mean estimation error,

$$E(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|^2) \leq \frac{2|\boldsymbol{t}|a_\lambda}{nl_*} + \frac{2|\boldsymbol{t}|\sigma^2}{nl_*} + (2e_1 - 4e_2 - 2e_3 - 2|\boldsymbol{t}|e_4)* \left(3\|\boldsymbol{\beta}^*\|^2 + 3\frac{n\sigma^2 + \sum P_\lambda(\beta_j^*)}{nl_*} + \frac{6rl^*\|\boldsymbol{\beta}^*\|^2}{l_*^2} + \frac{6rn\sigma^2}{n^2 l_*^2}\right).$$

Proof. Define

$$L_\lambda(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p} P_\lambda(|\beta_j|).$$

Let $\boldsymbol{\xi}(\boldsymbol{\beta}) = \{i : \beta_i \neq 0\}$ be the model extractor of $\boldsymbol{\beta}$, and let

$$R_{\boldsymbol{\xi}} = \boldsymbol{y}^T(I - X_{\boldsymbol{\xi}}(\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{X}_{\boldsymbol{\xi}})^{-1}\boldsymbol{X}_{\boldsymbol{\xi}})\boldsymbol{y}$$

denote the residual sum of squares of the OLS estimator of the model $\boldsymbol{\xi} = \boldsymbol{\xi}(\boldsymbol{\beta})$, where $\boldsymbol{X}_{\boldsymbol{\xi}}$ denotes the submatrix of $\boldsymbol{X}$ with columns corresponding to the predictors selected by $\boldsymbol{\xi}$. Therefore,

$$L_\lambda(\hat{\boldsymbol{\beta}}^o) = R_{\boldsymbol{t}} + \sum_{j=1}^{p} P_\lambda(|\hat{\beta}_j^o|),$$

where $\hat{\beta}_j^o$ denotes the jth element of $\hat{\boldsymbol{\beta}}^o$. Since $\hat{\boldsymbol{\beta}}_{\boldsymbol{t}}^o \sim N(\beta_{\boldsymbol{t}}^*, \sigma^2(\boldsymbol{X}_{\boldsymbol{t}}^T\boldsymbol{X}_{\boldsymbol{t}})^{-1})$, by Theorem 2.1 of [33], condition (b) and (e), we have

$$P\left\{L_\lambda(\hat{\boldsymbol{\beta}}^o) < R_{\boldsymbol{t}} + |\boldsymbol{t}|(c_\lambda + a_\lambda)\right\} \geq 1 - 2|\boldsymbol{t}|e_4. \tag{B.6}$$

Next, we show that for all $\boldsymbol{\beta}$ with $\boldsymbol{\xi}(\boldsymbol{\beta})$ strictly including the true model $\boldsymbol{t}$,

$$P\left\{\min_{\boldsymbol{\beta}:\boldsymbol{\xi}(\boldsymbol{\beta})\supset\boldsymbol{t},|\boldsymbol{t}|<|\boldsymbol{\xi}(\boldsymbol{\beta})|\leq r} L_\lambda(\boldsymbol{\beta}) > R_{\boldsymbol{t}} + |\boldsymbol{t}|(c_\lambda + a_\lambda)\right\} \geq 1 - 2e_1 - 2e_2. \tag{B.7}$$

Since $\boldsymbol{X}$ has been standardized such that each column has a norm of $\sqrt{n}$, $\boldsymbol{X}^T\boldsymbol{\epsilon}$ is a p-vector with each entry following the Gaussian distribution N(0, $n\sigma^2$). Then, by Theorem 2.1 of [33],

$$P\left\{|(\boldsymbol{X}^T\boldsymbol{\epsilon})_j| \le \sqrt{n}\sigma\sqrt{2\log(p/e_1)}, \quad \text{for all } j = 1, \ldots, p\right\} \ge 1 - 2e_1,$$

where $(\boldsymbol{z})_j$ denotes the jth element of the vector $\boldsymbol{z}$.

If $\boldsymbol{\xi} \supset \boldsymbol{t}$, then

$$\begin{aligned} &\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 = \|\boldsymbol{X}_{\boldsymbol{\xi}}\boldsymbol{\beta}^* + \boldsymbol{\epsilon} - \boldsymbol{X}_{\boldsymbol{\xi}}\beta_{\boldsymbol{\xi}}\|^2 \\ =&\boldsymbol{\epsilon}^T(I - P_{\boldsymbol{\xi}})\boldsymbol{\epsilon} + (\boldsymbol{u}_{\boldsymbol{\xi}} - (\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{X}_{\boldsymbol{\xi}})^{-1}\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{\epsilon})^T\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{X}_{\boldsymbol{\xi}}(\boldsymbol{u}_{\boldsymbol{\xi}} - (\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{X}_{\boldsymbol{\xi}})^{-1}\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{\epsilon}), \end{aligned} \tag{B.8}$$

where $\boldsymbol{\beta}^*_{\boldsymbol{\xi}}$ denotes the subvector of $\boldsymbol{\beta}^*$ corresponding to the model $\boldsymbol{\xi}$, $\boldsymbol{u}_{\boldsymbol{\xi}} = \boldsymbol{\beta}_{\boldsymbol{\xi}} - \boldsymbol{\beta}^*_{\boldsymbol{\xi}}$, and $P_{\boldsymbol{\xi}} = \boldsymbol{X}_{\boldsymbol{\xi}}(\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{X}_{\boldsymbol{\xi}})^{-1}\boldsymbol{X}_{\boldsymbol{\xi}}^T$ is the projection matrix. If $\boldsymbol{\beta}$ is outside the ellipse $\{\boldsymbol{\beta} : \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^*\|^2 + |\boldsymbol{t}|a_\lambda = \|\boldsymbol{\epsilon}\|^2 + |\boldsymbol{t}|a_\lambda\}$, then

$$L_\lambda(\boldsymbol{\beta}) > \|\boldsymbol{\epsilon}\|^2 + |\boldsymbol{t}|a_\lambda + \sum_{j=1}^{p} p_\lambda(|\beta_j|) \ge R_{\boldsymbol{t}} + |\boldsymbol{t}|(a_\lambda + c_\lambda),$$

by the property of the OLS estimator and the conditions (C_1) and (C_5).

If $\boldsymbol{\beta}$ is inside the ellipse, it follows from (B.8) that

$$\boldsymbol{\epsilon}^T(I-P_{\boldsymbol{\xi}})\boldsymbol{\epsilon}+(\boldsymbol{u}_{\boldsymbol{\xi}}-(\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{X}_{\boldsymbol{\xi}})^{-1}\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{\epsilon})^T\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{X}_{\boldsymbol{\xi}}(\boldsymbol{u}_{\boldsymbol{\xi}}-(\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{X}_{\boldsymbol{\xi}})^{-1}\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{\epsilon}) \le \|\boldsymbol{\epsilon}\|^2+|\boldsymbol{t}|a_\lambda,$$

which implies by condition (b) that

$$\|\boldsymbol{u}_{\boldsymbol{\xi}}\| \le \|(\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{X}_{\boldsymbol{\xi}})^{-1}\boldsymbol{X}_{\boldsymbol{\xi}}^T\boldsymbol{\epsilon}\| + \frac{1}{\sqrt{nl_*}}\sqrt{\boldsymbol{\epsilon}^T P_{\boldsymbol{\xi}}\boldsymbol{\epsilon} + |\boldsymbol{t}|a_\lambda}. \tag{B.9}$$

When all entries of $\boldsymbol{X}^T\boldsymbol{\epsilon}$ are bounded by $\sqrt{n}\sigma\sqrt{2\log(p/e_1)}$, we have

$$\|\boldsymbol{u}_{\boldsymbol{\xi}}\| \le 2\sqrt{\frac{2\sigma^2|\boldsymbol{\xi}|\log(p/e_1)}{nl_*^2} + \frac{|\boldsymbol{t}|a_\lambda}{nl_*}}. \tag{B.10}$$

It is easy to show that $P_\lambda(\sqrt{\cdot})$ is convex and thus

$$\begin{aligned}\sum_{j=1}^{|\boldsymbol{\xi}|} P_\lambda(|\beta_{\boldsymbol{\xi},j}|) &\ge \sum_{j=1}^{|\boldsymbol{\xi}|} P_\lambda(|\beta^*_{\boldsymbol{\xi},j}| + |u_{\boldsymbol{\xi},j}|) \ge |\boldsymbol{t}|c_\lambda + \sum_{\{j:\beta^*_{\boldsymbol{\xi},j}=0\}} P_\lambda(\sqrt{|u_{\boldsymbol{\xi},j}|^2}) \\ &\ge |\boldsymbol{t}|c_\lambda + (|\boldsymbol{\xi}| - |\boldsymbol{t}|)P_\lambda\left(2\sqrt{\frac{2\sigma^2|\boldsymbol{\xi}|\log(p/e_1)}{(|\boldsymbol{\xi}| - |\boldsymbol{t}|)nl_*^2} + \frac{|\boldsymbol{t}|a_\lambda}{(|\boldsymbol{\xi}| - |\boldsymbol{t}|)nl_*}}\right) \\ &\ge |\boldsymbol{t}|c_\lambda + (|\boldsymbol{\xi}| - |\boldsymbol{t}|)P_\lambda\left(2\sqrt{\frac{2\sigma^2(|\boldsymbol{t}|+1)\log(p/e_1)}{nl_*^2} + \frac{|\boldsymbol{t}|a_\lambda}{nl_*}}\right),\end{aligned} \tag{B.11}$$

where $\beta_{\boldsymbol{\xi},j}$, $\beta^*_{\boldsymbol{\xi},j}$ and $u_{\boldsymbol{\xi},j}$ denote the jth elements of $\boldsymbol{\beta}_{\boldsymbol{\xi}}$, $\boldsymbol{\beta}^*_{\boldsymbol{\xi}}$ and $\boldsymbol{u}_{\boldsymbol{\xi}}$, respectively; the third inequality follows from (B.10) and the convexity of $P_\lambda(\sqrt{\cdot})$; and the last inequality follows from the facts that both $|\boldsymbol{\xi}|/(|\boldsymbol{\xi}|-|\boldsymbol{t}|)$ and $|\boldsymbol{t}|/(|\boldsymbol{\xi}|-|\boldsymbol{t}|)$ are decreasing functions of $|\boldsymbol{\xi}|$.

In addition, we have

$$\|\boldsymbol{y} - \boldsymbol{X}_{\boldsymbol{\xi}}\boldsymbol{\beta}_{\boldsymbol{\xi}}\|^2 \ge R_{\boldsymbol{t}} - (R_{\boldsymbol{t}} - R_{\boldsymbol{\xi}}) = R_{\boldsymbol{t}} - \sigma^2 Z^2_{|\boldsymbol{\xi}|-|\boldsymbol{t}|}(\boldsymbol{\xi}), \tag{B.12}$$

where $Z^2_{|\boldsymbol{\xi}|-|\boldsymbol{t}|}(\boldsymbol{\xi})$ follows a χ^2-distribution of degree of freedom $|\boldsymbol{\xi}|-|\boldsymbol{t}|$. By Theorem 4.1 of [33] and Bonferroni inequality, with probability greater than $1 - \sum_{i=1}^{r-|\boldsymbol{t}|} e_2^i$ (which is greater than $1 - 2e_2$), for all $\boldsymbol{\xi}$ with $\boldsymbol{\xi} \supset \boldsymbol{t}$,

$$Z^2_{|\boldsymbol{\xi}|-|\boldsymbol{t}|}(\boldsymbol{\xi}) \le 2(|\boldsymbol{\xi}| - |\boldsymbol{t}|)\log(p/e_2) + |\boldsymbol{\xi}| - |\boldsymbol{t}| + 2(|\boldsymbol{\xi}| - |\boldsymbol{t}|)\sqrt{\log(p/e_2)}. \tag{B.13}$$

Combining (B.11), (B.12), (B.13) and condition (c), one can show (B.7) by Bonferroni

inequality.

Third, we show that for all $\boldsymbol{\beta}$ with $\boldsymbol{\xi}(\boldsymbol{\beta}) \not\supseteq \boldsymbol{t}$,

$$P\left\{\min_{\boldsymbol{\beta}:\boldsymbol{\xi}(\boldsymbol{\beta})\not\supseteq\boldsymbol{t},|\boldsymbol{\xi}(\boldsymbol{\beta})|\le r} L_\lambda(\boldsymbol{\beta}) > R_{\boldsymbol{t}} + |\boldsymbol{t}|(c_\lambda + a_\lambda)\right\} \ge 1 - 2e_2 - 2e_3. \tag{B.14}$$

If $\boldsymbol{\xi} \not\supseteq \boldsymbol{t}$, let $\boldsymbol{\zeta} = \boldsymbol{t} \cup \boldsymbol{\xi}$, then

$$L_\lambda(\boldsymbol{\beta}) > R_{\boldsymbol{\xi}} = (R_{\boldsymbol{\xi}} - R_{\boldsymbol{\zeta}}) - (R_{\boldsymbol{t}} - R_{\boldsymbol{\zeta}}) + R_{\boldsymbol{t}}, \tag{B.15}$$

where $(R_{\boldsymbol{\xi}} - R_{\boldsymbol{\zeta}})/\sigma^2$ is noncentral $\chi^2_{|\boldsymbol{\zeta}|-|\boldsymbol{\xi}|}(C)$ distribution with noncentral parameter

$$C = \boldsymbol{\beta}^*_{\boldsymbol{t}} \boldsymbol{X}^T_{\boldsymbol{t}} (P_{\boldsymbol{\zeta}} - P_{\boldsymbol{\xi}}) \boldsymbol{X}_{\boldsymbol{t}} \boldsymbol{\beta}^*_{\boldsymbol{t}} / \sigma^2 \ge n l_* \underline{\beta}^2 / \sigma^2.$$

If $\sqrt{n l_* \underline{\beta}^2/\sigma^2} > \sqrt{2\log(rp^r/e_3)}$, then by Theorem 2.1 of [33], with probability greater then $1 - 2e_3$, for all possible $\boldsymbol{\xi}$ with $\boldsymbol{t} \not\subseteq \boldsymbol{\xi}$,

$$\begin{aligned} R_{\boldsymbol{\xi}} - R_{\boldsymbol{\zeta}} &> \left\{\sqrt{n l_* \underline{\beta}^2} - \sigma\sqrt{2\log(rp^r/e_3)}\right\}^2 \\ &\ge \sigma^2(2r\log(p/e_2) + r + 2r\sqrt{\log(p/e_2)}) + |\boldsymbol{t}|(c_\lambda + a_\lambda). \end{aligned} \tag{B.16}$$

Combining (B.15), (B.16) and(B.13), one can show (B.14) by Bonferroni inequality.

Finally, we combine (B.6), (B.7) and (B.14), and conclude that

$$\begin{aligned} P\left\{\boldsymbol{\xi}(\hat{\boldsymbol{\beta}}) = \boldsymbol{t}\right\} &\ge P\left\{L_\lambda(\hat{\boldsymbol{\beta}}^o) < R_{\boldsymbol{t}} + |\boldsymbol{t}|(c_\lambda + a_\lambda)\right\} \\ &\quad + P\left\{\min_{\boldsymbol{\beta}:\boldsymbol{\xi}(\boldsymbol{\beta})\supset\boldsymbol{t},|\boldsymbol{t}|<|\boldsymbol{\xi}(\boldsymbol{\beta})|\le r} L_\lambda(\boldsymbol{\beta}) > R_{\boldsymbol{t}} + |\boldsymbol{t}|(c_\lambda + a_\lambda)\right\} \\ &\quad + P\left\{\min_{\boldsymbol{\beta}:\boldsymbol{\xi}(\boldsymbol{\beta})\not\supseteq\boldsymbol{t},|\boldsymbol{\xi}(\boldsymbol{\beta})|\le r} L_\lambda(\beta) > R_{\boldsymbol{t}} + |\boldsymbol{t}|(c_\lambda + a_\lambda)\right\} - 2 \\ &\ge 1 - 2e_1 - 4e_2 - 2e_3 - 2|\boldsymbol{t}|e_4, \end{aligned}$$

by Bonferroni inequality.

Suppose that $\boldsymbol{\xi}(\hat{\boldsymbol{\beta}}) = \boldsymbol{t}$. Let $\hat{\boldsymbol{\beta}}_{\boldsymbol{t}} = \min_{\{\boldsymbol{\beta}:\boldsymbol{\xi}(\boldsymbol{\beta})=\boldsymbol{t}\}} L_\lambda(\boldsymbol{\beta})$. Then,

$$\|\boldsymbol{y} - \boldsymbol{X}_{\boldsymbol{t}}\hat{\boldsymbol{\beta}}_{\boldsymbol{t}}\|^2 + |\boldsymbol{t}|c_\lambda < R_{\boldsymbol{t}} + |\boldsymbol{t}|(a_\lambda + c_\lambda).$$

It follows from the decomposition $\|\boldsymbol{y} - \boldsymbol{X}_{\boldsymbol{t}}\hat{\boldsymbol{\beta}}_{\boldsymbol{t}}\|^2 = R_{\boldsymbol{t}} + \|\boldsymbol{X}_{\boldsymbol{t}}\hat{\boldsymbol{\beta}}^o_{\boldsymbol{t}} - \boldsymbol{X}_{\boldsymbol{t}}\hat{\boldsymbol{\beta}}_{\boldsymbol{t}}\|^2$ that

$$(\hat{\boldsymbol{\beta}}_{\boldsymbol{t}} - \hat{\boldsymbol{\beta}}^o_{\boldsymbol{t}})^T \boldsymbol{X}_{\boldsymbol{t}}^T \boldsymbol{X}_{\boldsymbol{t}} (\hat{\boldsymbol{\beta}}_{\boldsymbol{t}} - \hat{\boldsymbol{\beta}}^o_{\boldsymbol{t}}) \le |\boldsymbol{t}| a_\lambda,$$

which, by condition (b), implies

$$\|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^o\| = \|\hat{\boldsymbol{\beta}}_{\boldsymbol{t}} - \hat{\boldsymbol{\beta}}^o_{\boldsymbol{t}}\| \le \sqrt{|\boldsymbol{t}| a_\lambda / n l_*}.$$

This concludes (B.5).

Let $\boldsymbol{\xi} = \boldsymbol{\xi}(\hat{\boldsymbol{\beta}})$ for some $\|\boldsymbol{\xi}\| \le r$, and let $\hat{\boldsymbol{\beta}}_{\boldsymbol{\xi}} = \min_{\{\boldsymbol{\beta}:\boldsymbol{\xi}(\boldsymbol{\beta})=\boldsymbol{\xi}\}} L_\lambda(\boldsymbol{\beta})$. Consider the case that $\boldsymbol{\xi} \ne \boldsymbol{t}$, then

$$\begin{aligned} &(\hat{\boldsymbol{\beta}}_{\boldsymbol{\xi}} - (\boldsymbol{X}_{\boldsymbol{\xi}}^T \boldsymbol{X}_{\boldsymbol{\xi}})^{-1} \boldsymbol{X}_{\boldsymbol{\xi}}^T \boldsymbol{y})^T (\boldsymbol{X}_{\boldsymbol{\xi}}^T \boldsymbol{X}_{\boldsymbol{\xi}}) (\hat{\boldsymbol{\beta}}_{\boldsymbol{\xi}} - (\boldsymbol{X}_{\boldsymbol{\xi}}^T \boldsymbol{X}_{\boldsymbol{\xi}})^{-1} \boldsymbol{X}_{\boldsymbol{\xi}}^T \boldsymbol{y}) \\ &\le \|\boldsymbol{y} - \boldsymbol{X}_{\boldsymbol{\xi}} \hat{\boldsymbol{\beta}}_{\boldsymbol{\xi}}\|^2 < L_\lambda(\hat{\boldsymbol{\beta}}_{\boldsymbol{\xi}}) \le L_\lambda(\boldsymbol{\beta}^*) = \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} + \sum P_\lambda(\beta_j^*), \end{aligned} \tag{B.17}$$

where the first inequality follows from the decomposition

$$\|\boldsymbol{y} - \boldsymbol{X}_{\boldsymbol{\xi}} \hat{\boldsymbol{\beta}}_{\boldsymbol{\xi}}\|^2 = R_{\boldsymbol{\xi}} + \|\boldsymbol{X}_{\boldsymbol{\xi}} \hat{\boldsymbol{\beta}}_{\boldsymbol{\xi}} - \boldsymbol{X}_{\boldsymbol{\xi}} \hat{\boldsymbol{\beta}}^o_{\boldsymbol{\xi}}\|^2,$$

and $\hat{\boldsymbol{\beta}}^o_{\boldsymbol{\xi}} = (\boldsymbol{X}_{\boldsymbol{\xi}}^T \boldsymbol{X}_{\boldsymbol{\xi}})^{-1} \boldsymbol{X}_{\boldsymbol{\xi}}^T \boldsymbol{y}$ denotes the OLS estimator of $\boldsymbol{\beta}_{\boldsymbol{\xi}}$. Therefore, $\|\hat{\boldsymbol{\beta}}_{\boldsymbol{\xi}} - (\boldsymbol{X}_{\boldsymbol{\xi}}^T \boldsymbol{X}_{\boldsymbol{\xi}})^{-1} \boldsymbol{X}_{\boldsymbol{\xi}}^T \boldsymbol{y}\|^2 \le (\boldsymbol{\epsilon}^T \boldsymbol{\epsilon} + \sum P_\lambda(\beta_j^*))/n l_*$ and $\boldsymbol{X}_{\boldsymbol{\xi}}^T \boldsymbol{y} = \boldsymbol{X}_{\boldsymbol{\xi}}^T \boldsymbol{X} \boldsymbol{\beta}^* + \boldsymbol{X}_{\boldsymbol{\xi}}^T \boldsymbol{\epsilon}$, where $\|\boldsymbol{X}\boldsymbol{\beta}^*\|^2 \le n l^* \|\boldsymbol{\beta}^*\|^2$ and each row of $\boldsymbol{X}_{\boldsymbol{\xi}}^T$ has been standardized to have a norm of

$\sqrt{n}$. It follows that $(\boldsymbol{X}_{\boldsymbol{\xi}}^T \boldsymbol{X}\boldsymbol{\beta}^*)_j \leq n\sqrt{l^*}\|\boldsymbol{\beta}^*\|$ for $j = 1, \ldots, |\boldsymbol{\xi}|$. Furthermore,

$$\begin{aligned}\|(\boldsymbol{X}_{\boldsymbol{\xi}}^T \boldsymbol{X}_{\boldsymbol{\xi}})^{-1}\boldsymbol{X}_{\boldsymbol{\xi}}^T \boldsymbol{y}\|^2 &\leq \frac{1}{n^2 l_*^2}\|\boldsymbol{X}_{\boldsymbol{\xi}}^T \boldsymbol{y}\|^2 \leq \frac{2}{n^2 l_*^2}(\|\boldsymbol{X}_{\boldsymbol{\xi}}^T \boldsymbol{X}\boldsymbol{\beta}^*\|^2 + \|\boldsymbol{X}_{\boldsymbol{\xi}}^T \boldsymbol{\epsilon}\|^2) \\ &\leq \frac{2rn^2 l^*\|\boldsymbol{\beta}^*\|^2}{n^2 l_*^2} + \frac{2\boldsymbol{\epsilon}^T \boldsymbol{X}_{\boldsymbol{\xi}} \boldsymbol{X}_{\boldsymbol{\xi}}^T \boldsymbol{\epsilon}}{n^2 l_*^2}.\end{aligned} \tag{B.18}$$

Following from (B.18),

$$\begin{aligned}\|\hat{\boldsymbol{\beta}}_{\boldsymbol{\xi}} - \boldsymbol{\beta}^*\|^2 &\leq 3\|\boldsymbol{\beta}^*\|^2 + 3\|\hat{\boldsymbol{\beta}}_{\boldsymbol{\xi}} - (\boldsymbol{X}_{\boldsymbol{\xi}}^T \boldsymbol{X}_{\boldsymbol{\xi}})^{-1}\boldsymbol{X}_{\boldsymbol{\xi}}^T \boldsymbol{y}\|^2 + 3\|(\boldsymbol{X}_{\boldsymbol{\xi}}^T \boldsymbol{X}_{\boldsymbol{\xi}})^{-1}\boldsymbol{X}_{\boldsymbol{\xi}}^T \boldsymbol{y}\|^2 \\ &\leq 3\|\boldsymbol{\beta}^*\|^2 + 3\frac{\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} + \sum P_\lambda(\beta_j^*)}{nl_*} + \frac{6rn^2 l^*\|\boldsymbol{\beta}^*\|^2}{n^2 l_*^2} + \frac{6\boldsymbol{\epsilon}^T \boldsymbol{X}_{\boldsymbol{\xi}} \boldsymbol{X}_{\boldsymbol{\xi}}^T \boldsymbol{\epsilon}}{n^2 l_*^2}.\end{aligned} \tag{B.19}$$

Combining (B.5) and (B.19), we have

$$\begin{aligned}E(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|^2) &\leq \frac{2|\boldsymbol{t}|a_\lambda}{nl_*} + 2E(\|\hat{\boldsymbol{\beta}}^o - \boldsymbol{\beta}^*\|^2) + (2e_1 - 4e_2 - 2e_3 - 2|\boldsymbol{t}|e_4) \\ &\quad \times E\left(3\|\boldsymbol{\beta}^*\|^2 + 3\frac{\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} + \sum P_\lambda(\beta_j^*)}{nl_*} + \frac{6rn^2 l^*\|\boldsymbol{\beta}^*\|^2}{n^2 l_*^2} + \frac{6\boldsymbol{\epsilon}^T \boldsymbol{X}_{\boldsymbol{\xi}} \boldsymbol{X}_{\boldsymbol{\xi}}^T \boldsymbol{\epsilon}}{n^2 l_*^2}\right) \\ &\leq \frac{2|\boldsymbol{t}|a_\lambda}{nl_*} + \frac{2|\boldsymbol{t}|\sigma^2}{nl_*} + (2e_1 - 4e_2 - 2e_3 - 2|\boldsymbol{t}|e_4) \\ &\quad \times \left(3\|\boldsymbol{\beta}^*\|^2 + 3\frac{n\sigma^2 + \sum P_\lambda(\beta_j^*)}{nl_*} + \frac{6rl^*\|\boldsymbol{\beta}^*\|^2}{l_*^2} + \frac{6rn\sigma^2}{n^2 l_*^2}\right).\end{aligned}$$

This concludes the proof of the lemma. □

Remark:

1. The conditions (c) and (d) look very technical, but can be interpreted intuitively. In order to bring sparsity into the model, the shape of the penalty function around zero is crucial. Traditional penalty functions, such as those used in LASSO, SCAD or MCP, are singular at zero and have the largest derivative at

zero, such that the coefficients of false predictors can shrink faster than those of true predictors. rLasso brings sparsity into the model in a different way: By giving a very large penalty around zero (i.e. condition (c)) such that the model cannot afford a small coefficient for the false predictor. Condition (d) restricts the dimensionality and eigen-structure of the design matrix. An arbitrarily large p or an arbitrarily small l_* increases the probability that the linear effect of a true predictor can be almost totally replaced by some combination of false predictors.

2. If, furthermore, there exists a sufficient small number e_5 and the following condition holds

   (f) $(r - |\boldsymbol{t}|)c_\lambda > |\boldsymbol{t}|a_\lambda + \sigma^2(n - |\boldsymbol{t}| + 2\log(1/e_5) + \sqrt{(n - |\boldsymbol{t}|)\log(1/e_5)})$,

   then in probability greater than $1 - e_5$, the following inequality holds

   $$(r + 1)c_\lambda > |\boldsymbol{t}|(c_\lambda + a_\lambda) + R_{\boldsymbol{t}},$$

   which implies that for any $\boldsymbol{\beta}$ with $|\boldsymbol{\xi}(\boldsymbol{\beta})| > r$, $L_\lambda(\boldsymbol{\beta}) > L_\lambda(\hat{\boldsymbol{\beta}}^o)$ holds. Hence the constraint $|\boldsymbol{\xi}(\boldsymbol{\beta})| \leq r$ is automatically satisfied in minimization of $L_\lambda(\boldsymbol{\beta})$. In this case, (B.4) is equivalent to

   $$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}}\{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p} P_\lambda(\beta_j)\}$$

   without the model size constraint.

To prove Theorem 3.2.1, we let $e_1 = e_2 = e_3 = e_4 = \exp(-K_n)$. Thus, the conditions of Lemma B.2.1 are satisfied when n is sufficiently large, and this concludes the consistency of rLasso for variable selection and parameter estimation.

APPENDIX C

MISCELLANEOUS MATERIAL

C.1 Computation Issue of Bayesian Variable Selection

The computation cost is of serious concern in the Bayesian variable selection. In our work, the prior specification allow β has positive probability take value 0, i.e., the so call "two-group" model. Such prior leads to direct model selection result, but the computation of posterior requires reversible jump or search the discrete model space.

Recently, there has been a development of scalable continuous shrinkage prior based on mixture of normal distribution (see e.g. [26], [59]), which aim at avoiding reversible jumps and increasing the computational effciency. Usually, gibbs sampler is implemented for shrinkage priors, where the full condition of $\boldsymbol{\beta}$ follows multivariate normal distribution. Sampling from p_n dimensional multivariate normal with non-diagnal covariance matrix needs computation complexity of order $O(p_n^3)$. Under the massive data simulation setting where p_n equals half million, the authors experience shows that it cost one day to update only 20 more iterations. In contrast, the computation complexity of proposed posterior is at most of the order of $O(\bar{r}_n^3)$. A serial analysis of this half-million-predictor big data set by the proposed SaM procedure cost approximated 16 hours on a single core of Intel® Xeon® CPU E5-2690(2.90Ghz). The computation time should be significant reduced if implemented in a parallel architecture. Since our approach has an embarrassingly parallel structure, the implementation of parallel computing does require communication among different nodes except collecting the selected variables' index at the end of first stage.

Thus the bandwidth limit of the connection between nodes should not influence the computation time very much.

[3] show that such shrinkage priors lead to posterior consistency in the case of $p_n = o(n)$, and [59] also demonstrate its prediction outperformance compared to frequentist methods. However, to the authors' best knowledge, a general study of estimation consistency under shrinkage prior for the high dimensional regression is not established yet.

A simple toy example is demonstrated here in figure C.1. The dimensionality of predictors $p_n = 450$, where the first 200 predictors are correlated with correlation equal 0.25. The nonzero coefficients are 0.2, 0.3, 0.4, 0.5, 0.6, and random error standard deviation $\sigma = 0.5$. The following horseshoe prior is used: $\beta_i \sim \mathrm{N}(0, \lambda_i^2\tau^2)$, $\lambda_i^2 \sim \mathrm{IB}(0.5, 0.5)$, $\tau^2 \sim \mathrm{IB}(0.5, 0.5)$, $\pi(\sigma^2) = 1/\sigma^2$. Consider the sample size is $n = 500, 400, 300, 200, 100$. We plot the L_2 estimation error of the posterior sample mean of $\boldsymbol{\beta}$, as the iterations go on. It is clear that when $n = 500, 400, 300, 200$, the horseshoe estimator is consistent, but when $n = 100$, the estimate has great bias. In table C.1 we report the L_2 estimation error of horseshoe estimator. This toy example shows that the horseshoe estimator is not stable with repect to the high dimensionality.

n	500	400	300	200	100
L_2 error	.0127	.0385	.0145	.0286	3.144

Table C.1: Failure of Bayesian shrinkage prior for high dimensional data. The plot gives L_2 estimation error of horseshoe estimator

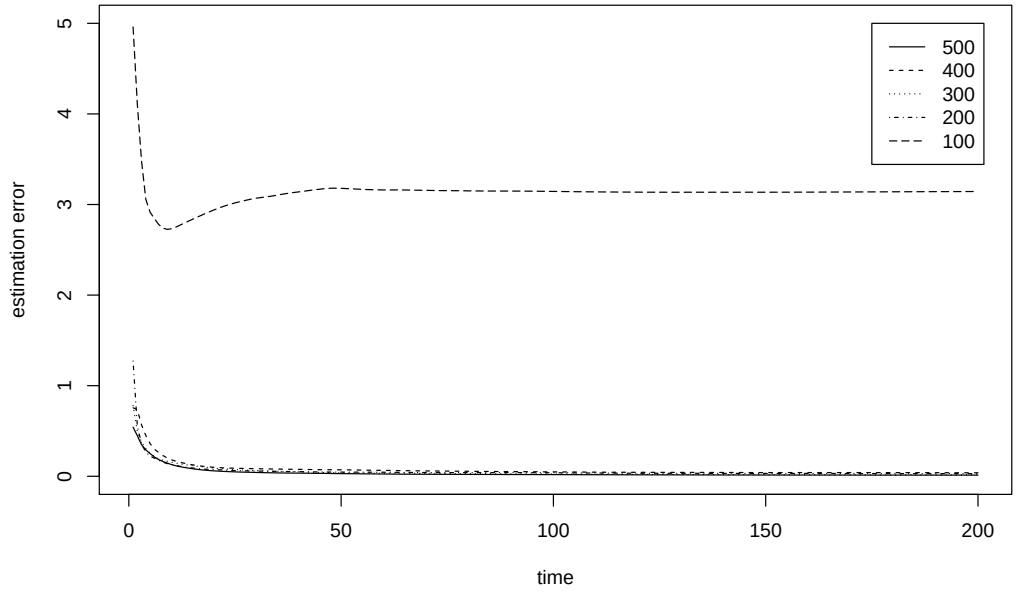

Figure C.1: Failure of Bayesian shrinkage prior for high dimensional data. The plot gives the curve of L_2 error of the parameter estimation of horseshoe prior as Markov chain goes on.

C.2 Full Simulation Results for rLasso

Table C.2 and Table C.3 provide numerical summaries for the results of rLasso simulation, under independent and dependent situation respectively. For each method and each sample size n, we reported the average values of nsr, fsr and squared coefficient estimation error as well as their standard deviations. In addition, considering the skewness of some statistics, we also reported their medians.

n		rLasso	MCP	Lasso	Ela.Net	SIS	ISIS	EBIC
80	nsr(%)	24.88	56.00	14.50	17.25	26.63	20.50	48.88
		(2.60)	(3.20)	(1.89)	(2.18)	(1.64)	(2.01)	(3.35)
		18.75	62.5	12.50	12.50	25.00	12.50	50.00
	fsr(%)	12.44	25.2	78.8	82.02	73.94	81.07	10.22
		(2.12)	(2.78)	(0.64)	(0.73)	(0.82)	(0.48)	(2.25)
		0.00	20.0	80.49	82.93	74.07	79.41	0.000
	error[1]	3.678	5.700	4.042	5.073	5.162	6.414	4.60
		(0.401)	(0.280)	(0.152)	(0.127)	(0.225)	(0.258)	(0.308)
		2.604	6.260	3.979	4.935	5.107	5.996	5.109
100	nsr(%)	2.50	14.13	2.13	3.00	18.38	5.25	8.25
		(0.64)	(2.85)	(0.64)	(0.69)	(1.24)	(0.99)	(1.86)
		0.00	0.00	0.00	0.00	12.50	0.00	0.00
	fsr(%)	2.37	14.49	81.65	84.43	70.96	77.65	3.74
		(0.65)	(1.93)	(0.44)	(0.19)	(0.76)	(0.24)	(0.84)
		0.00	11.11	83.67	84.62	71.43	76.47	0.00
	error	0.612	1.829	2.618	3.783	3.635	4.075	0.97
		(0.075)	(0.290)	(0.099)	(0.094)	(0.188)	(0.153)	(0.155)

Continued on next page

Table C.2 – *Continued from previous page*

n		rLasso	MCP	Lasso	Ela.Net	SIS	ISIS	EBIC
		0.285	0.414	2.491	3.789	3.387	3.675	0.232
120	nsr(%)	0.25	0.13	0.25	0.75	9.63	0.88	0.38
		(0.18)	(0.13)	(0.25)	(0.35)	(0.97)	(0.37)	(0.38)
		0.00	0.00	0.00	0.00	12.50	0.00	0.000
	fsr(%)	1.78	6.52	83.09	86.65	66.79	76.66	2.27
		(0.58)	(1.08)	(0.57)	(0.19)	(1.02)	(0.09)	(0.51)
		0.00	0.00	85.05	87.10	69.23	76.47	0.00
	error	0.290	0.288	1.917	2.920	2.204	3.006	0.247
		(0.030)	(0.030)	(0.066)	(0.078)	(0.140)	(0.079)	(0.034)
		0.196	0.176	1.809	2.794	2.030	2.907	0.172
150	nsr(%)	0.00	0.00	0.00	0.00	5.125	0.00	0.13
		(0.00)	(0.00)	(0.00)	(0.00)	(0.76)	(0.00)	(0.13)
		0.00	0.00	0.00	0.00	0.00	0.00	0.00
	fsr(%)	0.64	3.27	83.39	88.74	64.06	76.43	1.20
		(0.29)	(0.82)	(0.54)	(0.16)	(1.20)	(0.02)	(0.37)
		0.00	0.00	84.61	89.47	66.67	76.47	0.00
	error	0.162	0.186	1.335	2.092	1.308	2.064	0.187
		(0.013)	(0.017)	(0.046)	(0.056)	(0.101)	(0.056)	(0.017)
		0.125	0.141	1.259	2.045	0.924	2.004	0.140
170	nsr(%)	0.00	0.00	0.00	0.00	3.500	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.59)	(0.00)	(0.00)
		0.00	0.00	0.00	0.00	0.00	0.00	0.00

Continued on next page

Table C.2 – *Continued from previous page*

n		rLasso	MCP	Lasso	Ela.Net	SIS	ISIS	EBIC
		1.63	3.53	82.43	88.97	63.71	76.41	2.29
	fsr(%)	(0.51)	(0.77)	(0.68)	(0.27)	(1.23)	(0.02)	(0.49)
		0.00	0.00	83.50	90.30	66.67	76.47	0.00
		0.189	0.167	1.058	1.676	1.062	1.735	0.151
	error	(0.021)	(0.012)	(0.034)	(0.045)	(0.075)	(0.044)	(0.012)
		0.130	0.130	1.036	1.654	0.785	1.696	0.106
		0.00	0.00	0.00	0.00	1.13	0.00	0.00
	nsr(%)	(0.00)	(0.00)	(0.00)	(0.00)	(0.36)	(0.00)	(0.00)
		0.00	0.00	0.00	0.00	0.00	0.00	0.00
		0.83	3.54	82.31	89.33	65.15	78.95	1.51
200	fsr(%)	(0.36)	(0.88)	(0.80)	(0.31)	(1.34)	(0.00)	(0.43)
		0.00	0.00	83.67	90.24	69.23	78.95	0.00
		0.121	0.135	0.842	1.348	0.750	1.611	0.116
	error	(0.012)	(0.014)	(0.023)	(0.031)	(0.053)	(0.031)	(0.010)
		0.096	0.101	0.811	1.301	0.637	1.584	0.084

Table C.2: Full simulation result for rlasso under independent scenario. This table gives numerical result of rLasso, MCP, Lasso Elastic Net, SIS-SCAD, ISIS-SCAD and EBIC in study I. [1]: squared coefficient estimation error, i.e., $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|^2$. In each cell of the table, the top number is the mean, the middle number is the standard deviation of the mean, and bottom number is the median, which are all calculated based on 100 simulated datasets.

n		rLasso	MCP	Lasso	Ela.Net	SIS	ISIS	EBIC
		52.25	61.63	14.00	14.63	55.88	26.13	65.00
	nsr(%)	(2.40)	(2.59)	(1.55)	(1.70)	(1.79)	(1.89)	(2.50)
		62.50	62.50	12.50	12.50	50.00	25.00	75.00
		24.42	23.61	79.64	85.12	52.21	82.40	20.80
100	fsr(%)	(2.36)	(2.76)	(0.75)	(0.33)	(2.41)	(0.45)	(2.72)
		23.61	16.67	81.48	84.91	58.33	81.82	0.00
		7.42	6.458	4.288	5.019	6.082	9.325	6.469
	error[1]	(0.386)	(0.268)	(0.139)	(0.120)	(0.148)	(0.300)	(0.255)
		6.934	6.387	4.212	4.949	6.035	9.144	6.473
		26.58	36.00	4.25	5.75	47.50	11.88	40.63
	nsr(%)	(2.39)	(2.65)	(0.86)	(1.03)	(1.79)	(1.43)	(2.78)
		25.00	31.25	0.00	0.00	50.00	12.50	37.50
		10.72	20.10	81.91	86.11	43.80	79.15	9.06
120	fsr(%)	(1.54)	(2.13)	(0.64)	(0.32)	(2.36)	(0.34)	(1.64)
		0.00	18.33	83.16	86.89	50.00	79.10	0.00
		3.731	4.322	3.296	4.190	5.198	6.785	3.944
	error	(0.294)	(0.300)	(0.100)	(0.104)	(0.143)	(0.216)	(0.256)
		3.411	3.751	3.191	4.098	5.256	6.135	3.870
		6.75	11.75	1.75	1.75	39.88	2.88	12.25
	nsr(%)	(1.26)	(1.94)	(0.47)	(0.47)	(1.76)	(0.66)	(1.95)
		0.00	0.00	0.00	0.00	37.50	0.00	0.00
		6.35	12.78	83.16	87.82	39.33	77.01	5.75
150	fsr(%)	(1.13)	(1.71)	(0.59)	(0.29)	(2.21)	(0.17)	(1.14)

Continued on next page

Table C.3 – *Continued from previous page*

n		rLasso	MCP	Lasso	Ela.Net	SIS	ISIS	EBIC
		0.00	0.00	84.16	89.04	40.00	76.47	0.00
		1.383	1.675	2.515	3.277	4.466	4.604	1.420
	error	(0.172)	(0.195)	(0.088)	(0.087)	(0.140)	(0.119)	(0.183)
		0.551	0.805	2.387	3.202	4.605	4.418	0.697
		2.25	4.25	0.13	0.25	37.50	0.75	5.00
	nsr(%)	(0.60)	(0.96)	(0.13)	(0.18)	(1.68)	(0.30)	(1.32)
		0.00	0.00	0.00	0.00	37.50	0.00	0.00
		4.17	8.08	83.98	88.74	31.43	76.56	2.87
170	fsr(%)	(0.84)	(1.22)	(0.51)	(0.28)	(2.48)	(0.08)	(0.58)
		0.00	0.00	85.05	90.18	30.95	76.47	0.00
		0.776	0.864	2.023	2.783	4.238	3.925	0.702
	error	(0.089)	(0.104)	(0.060)	(0.063)	(0.133)	(0.095)	(0.117)
		0.356	0.475	1.965	2.707	4.322	3.816	0.221
		0.38	0.88	0.00	0.00	31.63	0.00	1.00
	nsr(%)	(0.21)	(0.36)	(0.00)	(0.00)	(1.64)	(0.00)	(0.38)
		0.00	0.00	0.00	0.00	37.50	0.00	0.00
		1.95	3.72	83.67	88.70	28.13	78.30	1.90
200	fsr(%)	(0.55)	(0.77)	(0.58)	(0.35)	(2.40)	(0.03)	(0.50)
		0.00	0.00	84.31	89.68	26.79	78.38	0.00
		0.360	0.378	1.613	2.284	3.639	3.411	0.317
	error	(0.041)	(0.049)	(0.048)	(0.057)	(0.160)	(0.067)	(0.045)
		0.220	0.193	1.614	2.286	3.947	3.413	0.175

Continued on next page

Table C.3 – *Continued from previous page*

n		rLasso	MCP	Lasso	Ela.Net	SIS	ISIS	EBIC

Table C.3: Full simulation result for rlasso under dependent scenario. This table gives numerical result of rLasso, MCP, Lasso Elastic Net, SIS-SCAD, ISIS-SCAD and EBIC in study II. [1]: squared coefficient estimation error, i.e., $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|^2$. In each cell of the table, the top number is the mean, the middle number is the standard deviation of the mean, and bottom number is the median, which are all calculated based on 100 simulated datasets.