Bayesian Variable Selection for Ultrahigh-dimensional Sparse Linear Models

Minerva Mukhopadhyay* and Subhajit Dutta[†]

Abstract. We propose a Bayesian variable selection procedure for ultrahighdimensional linear regression models. The number of regressors involved in regression, p_n , is allowed to grow exponentially with n. Assuming the true model to be sparse, in the sense that only a small number of regressors contribute to this model, we propose a set of priors suitable for this regime. The model selection procedure based on the proposed set of priors is shown to be variable selection consistent when all the 2^{p_n} models are considered. In the ultrahigh-dimensional setting, selection of the true model among all the 2^{p_n} possible ones involves prohibitive computation. To cope with this, we present a two-step model selection algorithm based on screening and Gibbs sampling. The first step of screening discards a large set of unimportant covariates, and retains a smaller set containing all the active covariates with probability tending to one. In the next step, we search for the best model among the covariates obtained in the screening step. This procedure is computationally quite fast, simple and intuitive. We demonstrate competitive performance of the proposed algorithm for a variety of simulated and real data sets when compared with several frequentist, as well as Bayesian methods.

Keywords: Model selection consistency, Screening consistency, Gibbs sampling.

1 Introduction

Variable selection in ultrahigh-dimensional setup is a flourishing area in the contemporary research scenario. It has become more important with increasing availability of data in various fields like genetics, finance, machine learning, etc. Sparsity has frequently been identified as an underlying feature for this kind of data sets. For example, in genome wide association studies (GWAS), "a prototype is measured for a large panel

[©] 0000 arxiv

^{*}Bethune College, Kolkata-700006, India. minervamukherjee@gmail.com

[†] Indian Institute of Technology, Kanpur-208016, India. tijahbus@gmail.com

of individuals, and a large number of single nucleotide polymorphisms (SNPs) throughout the genome are genotyped in all these participants. The goal is to identify SNPs that are statistically associated with the phenotype and ultimately to build statistical models to capture the effect of genetics on the phenotype" (Rosset (2013)). One such data set is the metabolic quantitative trait loci which consists of 10000 SNPs that are close to the regulatory regions (predictor variables) over a total of 50 participants (observations). A previous study by Song and Liang (2015) identified two particular SNPs to be important and significant. We have studied this data set in detail in a later section.

Several methods have been proposed to model high-dimensional data sets in both the frequentist and the Bayesian paradigm. Frequentist solutions to this problem are often through penalized likelihood, among which variants of LASSO like the elastic net of Zou and Hastie (2005), the group LASSO of Yuan and Lin (2006) and the adaptive LASSO of Zou (2006) are worth mentioning. Another important frequentist solution to this problem involves a screening algorithm to first reduce the data dimension, and then use some classical methods on this reduced data. This idea is implemented in sure independence screening (SIS) of Fan and Lv (2008), iterative SIS (ISIS) of Fan and Song (2010), forward selection based screening of Wang (2009), nonparametric independence screening (NIS) of Fan et al. (2011), iterative varying-coefficient screening (IVIS) of Song et al. (2014), etc. Other ways of approaching this problem is by using the smoothly clipped absolute deviation (SCAD) of Fan and Li (2001), the Dantzig selector of Candès and Tao (2007), modified EBIC of Chen and Chen (2012), etc. A detailed and nice review of most of these methods is contained in the paper by Fan and Lv (2010).

In the Bayesian literature, popular methods include the empirical Bayes variable selection of George and Foster (2000), and the spike and slab variable selection of Ishwaran and Rao (2005). Among recent developments, the methods of Bondell and Reich (2012), Liang et al. (2013), Song and Liang (2015) and Castillo et al. (2015) use the idea of penalized credible regions to accomplish variable selection in the ultrahigh-dimensional setting. While Castillo et al. (2015) have proved theoretical results related to the posterior consistency for the regression parameter, Liang et al. (2013) have shown the equivalence of posterior consistency and model selection consistency under appropriate sparsity assumptions. The authors of Narisetty and He (2014) claim to prove the

'strongest selection consistency result' using the spike and slab prior in the Bayesian framework. They introduce shrinking and diffusing priors, and establish strong selection consistency of their approach. In all of the above studies, the authors have considered the case where $\log p_n = o(n)$.

Note that the algorithms for computing the posterior distribution for the spike and slab prior are routine for small values of p_n and n, but the resulting computations are quite intensive for higher dimensions due to the large number of possible models. Several authors have developed MCMC algorithms that can cope with larger numbers of covariates, but truly high-dimensional models are still 'out of reach of fully Bayesian methods at the present time' (see Castillo et al. (2015)).

In this paper, we propose a Bayesian method for model selection, and examine model selection consistency for the same under the assumption of sparsity. In cases where $p_n >> n$, the number of competing models is so large that one first requires a screening algorithm to discard unimportant covariates. We present a two-step model selection procedure based on a screening algorithm and Gibbs sampling. The first step of the algorithm is shown to achieve *screening consistency* in the sense that it discards a large set of unimportant covariates with probability one.

The objective of the present work is three-fold. First, to develop a method which is suitable for ultrahigh-dimensional models. Secondly, to provide a faster and intuitive model selection algorithm. Finally, to keep the method and the algorithm as simple as possible. The proposed set of priors has the advantage of generating closed form expressions of marginals, which makes the method as tractable as a simple penalized likelihood method, such as Bayesian information criterion (BIC). To the best of our knowledge, this is the first work in the area of Bayesian variable selection which can accommodate cases with $\log p_n = O(n)$. The selection algorithm we adopt is simple and intuitive, and it makes the selection procedure quite fast. Further, its good performance is supported through theoretical results.

In Section 2, the prior setup and the model selection algorithm are described in detail. Section 3 contains the theoretical results including model selection consistency of the proposed set of priors, and consistency of the proposed algorithm. In Sections 4 and 5, we validate the performance of the proposed algorithm using simulated and real

data sets, respectively. Proofs of the main results are provided in Section 6, that of the other results and mathematical details are provided in a supplementary file.

2 The Proposed Prior and Model Selection Algorithm

2.1 Setup

Suppose we have n data points, each consisting of p_n regressors $\{x_{1,i}, x_{2,i}, \ldots, x_{p_n,i}\}$ and a response y_i with $i = 1, 2, \ldots n$. The response vector \mathbf{y}_n is modeled as follows

$$\mathbf{y}_n = X_n \boldsymbol{\beta} + \mathbf{e}_n, \tag{2.1}$$

where X_n is the $n \times p_n$ design matrix, $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_{p_n})'$ is the vector of corresponding regression parameters and \mathbf{e}_n is the vector of regression errors. We consider a sparse situation, where only a small number of regressors contributes to the model, while $p_n >> n$. For simplicity, we assume that the design matrix X_n is non-stochastic and $\mathbf{e}_n \sim N(\mathbf{0}, \sigma^2 I_n)$.

The space of all the 2^{p_n} models is denoted by \mathcal{A} , and indexed by α . Here, each α consists of a subset of size $p_n(\alpha)$ $(0 \leq p_n(\alpha) \leq p_n)$ of the set $\{1, 2, \ldots, p_n\}$, indicating which regressors are selected in the model. Under M_{α} , with $\alpha \in \mathcal{A}$, \mathbf{y}_n is modeled as

$$M_{\alpha}: \mathbf{y}_n = X_{\alpha} \boldsymbol{\beta}_{\alpha} + \mathbf{e}_n,$$

where X_{α} is a sub-matrix of X_n consisting of the $p_n(\alpha)$ columns specified by α and β_{α} is the corresponding vector of regression coefficients. When M_{α} is true, we assume that all the elements of β_{α} are non-zero. We consider the problem of selecting the model M_{α} with $\alpha \in \mathcal{A}$, which best explains the data. The true data generating model, denoted by M_{α_c} , is assumed to be an element of \mathcal{A} , and is expressed as

$$M_{\alpha_c}$$
: $\mathbf{y}_n = \boldsymbol{\mu}_n + \mathbf{e}_n = X_{\alpha_c} \boldsymbol{\beta}_{\alpha_c} + \mathbf{e}_n$,

where μ_n is the regression of \mathbf{y}_n given X_n . The dimension of M_{α_c} , denoted by $p(\alpha_c)$, is assumed to be a small number, free of n.

In a Bayesian approach, each model M_{α} is assigned a prior probability $p(M_{\alpha})$, and the corresponding set of parameters $\boldsymbol{\theta}_{\alpha} = (\beta_0, \boldsymbol{\beta}_{\alpha}, \sigma^2)$ involved in M_{α} , is also assigned

a prior distribution $p(\boldsymbol{\theta}_{\alpha}|M_{\alpha})$. Given the set of priors, one computes the posterior probability of each model. The posterior probability of the model M_{α} is given by

$$p(M_{\alpha}|\mathbf{y}_n) = \frac{p(M_{\alpha})m_{\alpha}(\mathbf{y}_n)}{\sum_{\alpha \in \mathcal{A}} p(M_{\alpha})m_{\alpha}(\mathbf{y}_n)},$$

where

$$m_{\alpha}(\mathbf{y}_n) = \int p(\mathbf{y}_n | \boldsymbol{\theta}_{\alpha}, M_{\alpha}) p(\boldsymbol{\theta}_{\alpha} | M_{\alpha}) d\boldsymbol{\theta}_{\alpha}$$

is the marginal density of \mathbf{y}_n , $p(\mathbf{y}_n|\boldsymbol{\theta}_{\alpha}, M_{\alpha})$ is the density of \mathbf{y}_n given the model parameters $\boldsymbol{\theta}_{\alpha}$ and $p(\boldsymbol{\theta}_{\alpha}|M_{\alpha})$ is the prior density of $\boldsymbol{\theta}_{\alpha}$ under M_{α} . We consider the procedure that selects the model in \mathcal{A} with the highest posterior probability.

We denote the rank of the design matrix of model M_{α} by r_{α} , i.e., $r\left(X'_{\alpha}X_{\alpha}\right)=r_{\alpha}$, and also refer r_{α} as the rank of M_{α} . For two numbers a and b, the notations $a \vee b$ and $a \wedge b$ are used to denote $\max\{a,b\}$ and $\min\{a,b\}$, respectively. For α , $\alpha^* \in \mathcal{A}$, the notations $X_{\alpha \vee \alpha^*}$ and $X_{\alpha \wedge \alpha^*}$ are used to denote sub-matrices of X formed by columns corresponding to either X_{α} or X_{α^*} (or both), and columns which are common to both X_{α} and X_{α^*} , respectively. For two square matrices A and B of the same order, $A \leq B$ means that B - A is positive semidefinite.

2.2 Prior Specification

On each model M_{α} with $\alpha \in \mathcal{A}$, we assign the *Bernoulli* prior probability as follows:

$$P(M_{\alpha}) = q_n^{p_n(\alpha)} \left(1 - q_n\right)^{p_n - p_n(\alpha)} \quad \text{with} \quad q_n = 1/p_n.$$

Given a model M_{α} , we consider the conjugate prior on β_{α} as

$$\boldsymbol{\beta}_{\alpha}|\sigma^2, M_{\alpha} \sim N(\mathbf{0}, g_n \sigma^2 I_{p_n(\alpha)}),$$

where g_n is a hyperparameter which depends on n. When σ^2 is unknown, we consider the popular Jeffreys prior $\pi(\sigma^2) \propto 1/\sigma^2$.

The Bernoulli prior probability is widely used as model prior probability because of its property of favoring, or penalizing models of large, or small dimensions. The choice $q_n = 1/p_n$ has previously been considered by Narisetty and He (2014). This prior is

particularly useful for sparse regression models, where it is known in advance that the true model is small-dimensional, and p_n is quite large.

The use of the inverse gamma prior for error variance is fairly conventional in the literature of model selection (see, e.g., Johnson and Rossell (2012), Narisetty and He (2014)). Jeffreys prior is the limit of inverse gamma prior, as both the hyperparameters involved in inverse gamma prior approach zero. The property of invariance under reparametrization makes it suitable as a prior on the scale parameter.

We choose a simple set of priors. Except for the choice of g_n , we completely specify the set of priors. We do not provide any specific choice of g_n , rather indicate the optimal order which is necessary to achieve consistency. The posterior probabilities generated using this set of priors is of a closed form, which makes the resulting method easily applicable.

2.3 Model Selection Algorithm

Our model selection procedure is quite simple as it chooses the model with the highest posterior probability among all competing models. In the next section, we will show that the proposed set of priors is model selection consistent in the sense that the posterior probability of the true model goes to one. However, identifying the model with the highest posterior probability still remains a challenging task for ultrahigh-dimensional data. As $p_n = \exp\{O(n)\}$, it is impossible to evaluate all the 2^{p_n} models in the model space even for small values of n. For example, if n = 5, the model space can be of order $\exp(45)$, which is a huge number. Therefore, we need to develop a screening algorithm, which reduces the model space to a tractable one. In other words, we need to discard a set of 'unimportant' variables at the beginning using some suitable algorithm. After implementation of the algorithm, ideally, we will be left with a smaller set of variables which includes all the active covariates. We describe the algorithm in detail below.

The Two-step Algorithm.

Screening: The first step discards a large set of unimportant covariates. Here, we use the fact that the number of regressors in the true model, $p(\alpha_c)$, is very small and free of n. First, we choose an integer d such that $d/K_0 \leq p(\alpha_c) < d$, where K_0 is a positive

number free of n. We will choose the best model in the class of all models of dimension d. Thus among the 2^{p_n} models, we only compare $\binom{p_n}{d}$ models. Once d is selected, we proceed along the following steps:

- 1. Initialization. Choose any model M_{α_0} of dimension d, and calculate its marginal.
- 2. Evaluation. Consider each of the covariates present in M_{α_0} individually. Let x_1 be a covariate of M_{α_0} . Replace x_1 with each of the covariates which are not present in M_{α_0} one by one, and compute the marginal density. Update M_{α_0} by replacing x_1 with the covariate that yields the highest marginal density, say x^* , if $x^* \neq x_1$. Retain x_1 otherwise. Do the same for the other (d-1) active covariates of M_{α_0} as well.
- 3. Replication. Repeat the previous step N times, where N is a moderately large number.

In the next section, we will show that if N is moderately large, the screening algorithm finally selects a supermodel of the true model with probability tending to one.

Model Selection: Once the screening algorithm selects a model, say M_{α^*} , we discard all the regressors that are not present in M_{α^*} . In the next step, we apply the Gibbs sampling algorithm to select the best model among the 2^d models, which can be formed by the d regressors present in M_{α^*} . The sampling scheme that we use is completely described in Chipman et al. (2001, Section 3.5) in the section on Gibbs Sampling Algorithms under the conjugate setup. Note that the Gibbs sampling algorithm chooses models directly following a Markov chain with ratio of the posterior probabilities as the transition kernel, and the set of regressors obtained at the end of screening step contains all the active covariates with probability tending to one. Therefore, after sufficient iterations, the algorithm must select the model with highest posterior probability, i.e., the true

Remark 2.1. Note that the total number of models among which the algorithm selects the best one is $\binom{p_n}{d} + 2^d$. Thus, if we have some idea about the actual number of active covariates, we can use it to choose d as small as possible. A small choice of d makes the algorithm much faster.

Remark 2.2. In the evaluation step of the screening algorithm, we update the chosen model $d(p_n-d)$ times. If we repeat the evaluation step N times, then $Nd(p_n-d)$ updates take place. Therefore, it is enough to choose a moderately large N.

Note that $Nd(p_n - d)$ is also the computational complexity of the screening step. Even if we consider all the 2^d competing models for comparison in the second stage, the total computational complexity of the proposed algorithm would be $Nd(p_n - d) + 2^d$, which is linear in p_n and much smaller than 2^{p_n} .

3 Consistency of the Proposed Prior

This section is dedicated towards asserting consistency results of the proposed method of model selection. We consider the cases with known, as well as unknown error variances σ^2 separately, stating clearly the assumptions required in each case.

3.1 Results for Known Error Variance

Often one has enough data to estimate the variance σ^2 properly, or being independent of the design matrix, σ^2 is estimated from earlier data sets. In such cases, σ^2 may be assumed to be known. In this subsection, we discuss results for the case with known σ^2 .

Given σ^2 and g_n , the posterior probability of model M_{α} is proportional to

$$P(M_{\alpha}|\mathbf{y}_n) \propto \left(\frac{1}{p_n - 1}\right)^{p_n(\alpha)} \left| I_{p_n(\alpha)} + g_n X_{\alpha}' X_{\alpha} \right|^{-1/2} \exp\left\{ -\frac{R_{\alpha}^{2*}}{2\sigma^2} \right\},\,$$

where $R_{\alpha}^{2*} = \mathbf{y}_n' \left\{ I_n - X_{\alpha} \left(I_{p_n(\alpha)} / g_n + X_{\alpha}' X_{\alpha} \right)^{-1} X_{\alpha}' \right\} \mathbf{y}_n$. Our results for the case with known σ^2 is based on the following set of assumptions.

- (A1) The number of regressors $p_n = \exp\{b_0 n^{1-r}\}$ where $0 \le r < 1$ and b_0 is any number free of n.
- (A2) The true model M_{α_c} is unique. There exists constants τ_{\max}^* and τ_{\min}^* , free of n, such that $n\tau_{\min}^* I_{p(\alpha_c)} \leq X'_{\alpha_c} X_{\alpha_c} \leq n\tau_{\max}^* I_{p(\alpha_c)}$.
- (A3) Let τ_{max} and τ_{min} be the highest and lowest non-zero eigenvalues of $X'_n X_n / n$, then $\tau_{\text{max}} \leq p_n^{|z_n|}$ with $z_n \to 0$, and $\tau_{\text{min}} \geq p_n^{-|w_n|}$ with $w_n \to 0$.

(A4) Consider the constants $K_0 > 6$ and $\Delta_0 = \{\delta n^{1-s}\} \vee \{6\sigma^2 p(\alpha_c) \log p_n\}$ with $\delta > 0$ and $0 < s \le 0.5$. Let $\mathcal{A}_3 = \{\alpha \in \mathcal{A} : M_{\alpha_c} \nsubseteq M_{\alpha}, r_{\alpha} \le K_0 \ p(\alpha_c)\}, \ \boldsymbol{\mu}_n = X_{\alpha_c} \boldsymbol{\beta}_{\alpha_c}$ and $P_n(\alpha)$ be the projection matrix onto the span of X_{α} . We assume that

$$\inf_{\alpha \in \mathcal{A}_3} \boldsymbol{\mu}_n' (I - P_n(\alpha)) \boldsymbol{\mu}_n > \Delta_0.$$

(A5) The hyperparameter g_n is such that $ng_n = p_n^{2+\delta_1}$, for some $5/(K_0 - 1) \le \delta_1 \le 1$ free of n, where K_0 is as stated in assumption (A4) above.

Assumptions (A1) and (A5) describe the setup and our choice of the hyperparameter g_n , respectively. Assumption (A2) states that the true model is unique, and it includes a set of independent regressors. The design matrix corresponding to $X'_{\alpha_c}X_{\alpha_c}$ depends on n, but not on p_n . Therefore, we allow the eigenvalues of the true model to vary only with n. Assumption (A3) is also quite general, as we allow the eigenvalues of X'_nX_n to vary with both n and p_n . This is more reasonable since the dimension of X'_nX_n depends on both n and p_n .

Assumption (A4) is commonly termed as an identifiability condition for model selection. The quantity $\mu'_n(I-P_n(\alpha))\mu_n$ may be interpreted as the distance of the α^{th} model from the true model. For consistent model selection, it is necessary for the true model to keep a distance from other models. Otherwise, the true model may not be identifiable. It has been proved in Moreno et al. (2015, Lemma 3) that $\lim_{n\to\infty} \mu'_n(I-P_n(\alpha))\mu_n/n > 0$ for any non-supermodel of the true model. We have just assumed a uniform lower bound for $\mu'_n(I-P_n(\alpha))\mu_n$ over non-supermodels of low rank, and fixed a threshold value for the case when $\log p_n = b_0 n$ with $b_0 > 0$. When $\log p_n = b_0 n^{1-r}$ with r > 0, the threshold is not even of order n, and therefore, the condition is trivially satisfied (by Moreno et al. (2015, Lemma 3)).

The consistency results are split into two parts. Model selection consistency of the proposed set of priors is shown in Section 3.1.1, and consistency of the model selection algorithm is shown in Section 3.1.2.

3.1.1. Model Selection Consistency

If the true model is among one of the candidate models in the model space, it is natural to check whether a model selection procedure can identify the true model with probability tending to one. This property, known as 'model selection consistency', requires $P(M_{\alpha_c}|\mathbf{y}_n) \xrightarrow{p} 1$, which is equivalent to showing that

$$\sum_{\alpha \in \mathcal{A} \setminus \{\alpha_c\}} \frac{P(M_{\alpha}|\mathbf{y}_n)}{P(M_{\alpha_c}|\mathbf{y}_n)} \xrightarrow{p} 0.$$
 (3.1)

We now state the result on model selection consistency for the case where σ^2 is known.

Theorem 3.1. Consider the model (2.1) with known σ^2 . Under assumptions (A1)-(A5), the method based on the proposed set of priors is model selection consistent.

Remark 3.1. The proof of model selection consistency for known σ^2 (see Section 6.2) only requires $p_n \to \infty$, and does not explicitly require $n \to \infty$. As $\log p_n \leq b_0 n$, an appropriately large p_n and only a moderately large n is sufficient for good performance of this set of priors in practice. From this point of view, the proposed set of priors is suitable for high-dimensional medium sample size settings.

3.1.2. Consistency of the Model Selection Algorithm

In the proposed model selection algorithm at the end of the screening step, one will be left with a model of dimension d. We claim that this step of screening is consistent in the sense that the model chosen at the end of it, say M_{α^*} , may not be unique but it would be a supermodel of the true model, i.e., $M_{\alpha_c} \subseteq M_{\alpha^*}$ with probability tending to one. We now consider the following result.

Theorem 3.2. Let d be an integer such that $d/K_0 \leq p(\alpha_c) < d$, and M_{α_1} and M_{α_2} be a supermodel, and a non-supermodel of M_{α_c} of dimension d, respectively. If σ^2 is known and assumptions (A1)-(A5) hold, then for the proposed set of priors we have

$$\sup_{\alpha_1,\alpha_2} \frac{P(M_{\alpha_2}|\mathbf{y}_n)}{P(M_{\alpha_1}|\mathbf{y}_n)} \xrightarrow{p} 0, \quad or, \ equivalently \quad \sup_{\alpha_1,\alpha_2} \frac{m(M_{\alpha_2}|\mathbf{y}_n)}{m(M_{\alpha_1}|\mathbf{y}_n)} \xrightarrow{p} 0 \quad under \ M_{\alpha_c}.$$

This theorem states that under M_{α_c} , the posterior probability of a supermodel of M_{α_c} of dimension d is much higher than the posterior probability of a non-supermodel of same dimension. As the algorithm selects a model on the basis of the marginal density, which is equivalent to the posterior probability when models of the same dimension are considered, it is expected that a supermodel will be selected after some iterations.

After a supermodel is selected in the first stage, we only consider the d regressors

included in the supermodel and find the best model among the 2^d possible ones formed by these d regressors. As the Gibbs sampling algorithm in the second step chooses models on the basis of posterior probabilities, consistency of this part is immediate from the screening consistency and the model selection consistency of the proposed set of priors.

3.2 Results for Unknown Error Variance

When the error variance σ^2 is unknown, we assign the standard non-informative Jeffreys prior $\pi(\sigma^2) \propto 1/\sigma^2$. In this case, the posterior probability of any model M_{α} is

$$p\left(M_{\alpha}|\mathbf{y}_{n}\right) \propto \left(\frac{1}{p_{n}-1}\right)^{p_{n}(\alpha)} \left|I+g_{n}X_{\alpha}'X_{\alpha}\right|^{-1/2} \left(R_{\alpha}^{2*}\right)^{-n/2}.$$

As σ^2 is unknown, we need to modify assumptions (A1)-(A5) of the previous subsection. The modified assumptions are stated below:

- (B1) For some positive integer M free of n, $rank(X'_nX_n) \leq M$.
- (B2) The number of regressors $p_n = exp\{b_0n^{1-r}\}\$ where $0 \le r < 1$ and b_0 is any number free of n. For r = 0, we need $b_0 < \left[\xi(1-\xi)\{2(1+\xi)(M-p(\alpha_c))\}^{-1}\right] \wedge (4p(\alpha_c))^{-1}$ for some $1/(K_0 1) < \xi \le 0.1$ and $K_0 > 12$.
- (B3) The hyperparameter g_n is such that $ng_n = p_n^{2+\delta_1}, 7\xi/(1-\xi)^2 + 2/(K_0-1) \le \delta_1 \le 2$.
- (B4) Assumption (A4) holds with $\Delta_0 = \{12\sigma^2 p(\alpha_c) \log p_n\} \vee \{\delta n^{1-s}\}\$ for $0 < s \le 0.5$.

Note that assumption (B1) implies that the highest and the lowest non-zero eigenvalues of the design matrix are free of n. Assumption (B2) imposes some additional restrictions on the dimension p_n when it is of the order $\exp\{O(n)\}$. Unlike the case for known σ^2 , here we fail to accommodate any p_n of the order $\exp\{O(n)\}$ (recall assumption (A1)), rather impose a multiplicative constant b_0 such that $\log p_n \leq b_0 n$. Assumption (B3) indicates that we need a slightly larger value of g_n in order to achieve consistency when the parameter σ^2 is unknown. Finally, assumption (B4) is same as assumption (A4) with a partially changed threshold value. Nevertheless, implications of the assumption and its importance remains the same here. We now state the result on model selection consistency for an unknown value of σ^2 .

Theorem 3.3. Consider the model (2.1) with unknown σ^2 . Under assumptions (B1)-(B4), the method based on the proposed set of priors is model selection consistent.

We do not present a separate result for screening consistency of the algorithm stated in Section 2.3 for the case where σ^2 is unknown. A result similar to Theorem 3.3 can be stated here. A proof similar to that Theorem 3.3 (i.e., the case with known σ^2 , see Section 6.2) can also be presented in this respect using assumptions (B1)-(B4) instead of (A1)-(A5).

4 Simulation Study

We validate the performance of the proposed method of model selection using a wide variety simulated data sets. Under different simulation schemes, we present the proportion of times a model selection algorithm selects the true model.

Our method: The model selection algorithm we follow is completely described in Section 2.3. The number of regressors selected at the first stage, d, is taken to be $\lfloor n/4 \rfloor$ in each case. In the screening step, we choose $g_n = p_n^2/n$ and in the second step of model selection, we choose $g_n = d^2$.

Other methods: As we mentioned in the Introduction, there are several methods for variable selection both from the classical, as well as the Bayesian perspectives. We consider some of the more competitive methods for comparison. Among the classical methods, we consider three approaches based on iterative sure independence screening (ISIS), namely, ISIS-LASSO-BIC, ISIS-SCAD-BIC and ISIS-MCP-BIC. Here an initial set of variables are first selected by ISIS, and then a step of penalized regression is carried out using the least absolute shrinkage and selection operator (LASSO), smoothly clipped absolute deviation (SCAD), or minimax concave penalty (MCP, Zhang (2010)) with the regularization parameter tuned by the BIC. Among the Bayesian competitors, we consider methods based on Bayesian credible region (BCR.marg and BCR.joint, Bondell and Reich (2012)) and Bayesian shrinking and diffusing prior (BASAD, Narisetty and He (2014)). We have used R codes for all the methods. For ISIS, we have implemented codes from the R package SIS. The R codes for BCR is obtained from the first author's website, while the first author of Narisetty and He (2014) kindly shared the codes for BASAD with us. There are two versions for BASAD, one is exact while the other is an

approximate one for high-dimensional data. We have implemented the second version for the sake of saving computing time.

Simulation setup. We consider two values for n, namely, 50 and 100. For n = 50, we choose $p_n = 100$ and 500, while for n = 100 we choose $p_n = 500$, 1000 and 2000.

The model $\mathbf{y}_n = \boldsymbol{\mu}_n + \mathbf{e}_n$ is considered as the true model, where $\boldsymbol{\mu}_n = X_{\alpha_c} \boldsymbol{\beta}_{\alpha_c}$. The vector $\boldsymbol{\beta}_{\alpha_c}$ is assumed to be sparse, i.e., there are only $p(\alpha_c)$ components in $\boldsymbol{\beta}_{\alpha_c}$ with $p(\alpha_c) << p_n$ and these $p(\alpha_c)$ components are chosen randomly from the set of indices $\{1, \ldots, p_n\}$. When p_n is less than or equal to 500 we set the number of active regressors $p(\alpha_c) = 5$, while $p(\alpha_c) = 10$ for higher values of p_n . The $p(\alpha_c)$ values of $\boldsymbol{\beta}_{\alpha_c}$ are taken to be equal (say, β), and we fix a common constant value of $\beta = 2$.

Each data vector \mathbf{x}_i of the design matrix $X_n = (\mathbf{x}_1, \dots, \mathbf{x}_n)'$ is assumed to follow the Gaussian distribution with mean $\mathbf{0}$ and covariance Σ_{p_n} for $i = 1, \dots, n$. The covariance structure of $\Sigma_{p_n} = ((\sigma_{ij}))$ for $1 \leq i, j \leq p_n$ is taken to be the following four types:

Case 1. (Identity) $\Sigma_{p_n} = \mathbf{I}$, i.e., there is no correlation among the covariates.

Case 2. (Block Dependence) Σ_{p_n} has a block covariance setting, where the active covariates have common correlation $\rho_1 = 0.25$, the inactive covariates have common correlation $\rho_2 = 0.75$ and each pair of active and inactive covariate has correlation $\rho_3 = 0.50$. This is an interesting co-variance structure as it attributes different correlations depending on whether the covariate is important, or not (also see Narisetty and He (2014)).

Case 3. (Equi-correlation) $\Sigma_{p_n} = 0.5\mathbf{I} + 0.5\mathbf{11}'$, where 1 is the p_n -dimensional vector of ones. This exhibits a strong dependence structure uniformly among the covariates.

Case 4. (Autoregressive) Here, we take $\sigma_{ii} = 1$ for $1 \leq i \leq p_n$, and $\sigma_{ij} = 0.9^{|i-j|}$ for $1 \leq i \neq j \leq p_n$. Clearly, we have a decaying correlation structure depending on the distance |i-j|. With the increase in distance, here the correlation decreases.

Let $\mathbf{e}_n \sim f_n$, where f_n denotes a n-dimensional multivariate distribution. We have considered two choices for f_n , namely, the standard multivariate Gaussian distribution and a heavy-tailed distribution, namely, the multivariate t_2 distribution. Note that the moments of order 2, or higher fail to exist for the t_2 distribution. In the tables below, we have reported the proportion of times each method selected the true model in the 200 random iterations.

Simulation results. In this simulated regime, we have provided some extra information to BASAD and BCR. When we implemented BCR tuned with BIC, as well as BASAD tuned with BIC, we specified the exact number of non-zero components, i.e., the information of $p(\alpha_c)$. Further, we notice that the covariance structure for Case 2 becomes singular for $p_n = 1000$, or higher and we have restricted it for $p_n = 500$, or less.

Table 1: Proportion of times true model is selected by each method for n = 50

Gaussian error							
Methods	Case 1	Case 2	Case 3	Case 4			
$\downarrow p_n \rightarrow$	100 500	100 500	100 500	100 500			
ISIS-SCAD-BIC	0.710 0.325	0.015 0.000	0.715 0.355	0.560 0.150			
ISIS-MCP-BIC	0.525 0.135	0.010 0.000	0.485 0.155	0.230 0.020			
BCR.marg-BIC	0.280 0.010	0.395 0.015	0.255 0.000	0.180 0.000			
BASAD	0.930 0.445	0.865 0.150	0.925 0.430	0.880 0.605			
BASAD-BIC	1.000 0.625	0.995 0.230	1.000 0.620	1.000 0.805			
Proposed	1.000 0.870	0.685 0.100	0.985 0.775	1.000 0.935			
t_2 error							
	Case 1	Case 2	Case 3	Case 4			
Methods $\downarrow p_n \rightarrow$	Case 1 100 500	Case 2 100 500	Case 3 100 500	Case 4 100 500			
$\begin{array}{c} \text{Methods} \downarrow & p_n \to \\ \text{ISIS-SCAD-BIC} \end{array}$							
	100 500	100 500	100 500	100 500			
ISIS-SCAD-BIC	100 500 0.385 0.330	100 500 0.020 0.000	100 500 0.355 0.335	100 500 0.315 0.230			
ISIS-SCAD-BIC ISIS-MCP-BIC	100 500 0.385 0.330 0.325 0.280	100 500 0.020 0.000 0.025 0.000	100 500 0.355 0.335 0.305 0.285	100 500 0.315 0.230 0.225 0.150			
ISIS-SCAD-BIC ISIS-MCP-BIC BCR.marg-BIC	100 500 0.385 0.330 0.325 0.280 0.180 0.005	100 500 0.020 0.000 0.025 0.000 0.300 0.000	100 500 0.355	100 500 0.315 0.230 0.225 0.150 0.180 0.000			

Among the three methods of variable selection based on ISIS, we have reported the results for SCAD and MCP only. LASSO usually over-estimates β_{α_c} , and we have not reported it for our numerical study. For the other two methods, we observe (see Table 1) that SCAD performed uniformly better than MCP. It is also clear from Table 1 that ISIS is affected drastically when the dependence structure varies among the different sets of covariates (Case 2). Moreover, the proportion of times it selects the true model decreased significantly for heavy-tailed errors. This is explained by the fact that ISIS relies on directly computing covariances between the variables.

Generally, the Bayesian methods turn out to be *more robust* than the moment based approaches. Among the Bayesian methods, BASAD clearly performed the best for n = 50 with p = 100; and n = 100 with p = 500. We observed that BASAD-BIC lead to an improved performance over BASAD in some cases, which is unlike what

Narisetty and He (2014) had observed and it is clear that the additional step of BIC in BASAD is quite sensitive to this tuning parameter $p(\alpha_c)$.

However, the performance of BASAD falls drastically for higher values of p_n (see Table 1). Note that BASAD needs to compute the inverse of the covariance matrix for each model, which is computationally prohibitive for such high-dimensional data. To resolve this problem, they use a block covariance structure to simplify some of the matrix computations and this can be one of the main reasons behind the poor performance. For BCR, we observe that the joint version leads to singularity in several iterations for $p_n = 1000$ onwards. Therefore, we have reported results for the more stable marginal version only. The strength of our proposed method is re-instated from this numerical study, especially for higher values of p_n . Clearly, there is a systematic improvement of the proposed method over BASAD when we move from $p_n = 100$ to $p_n = 500$ for n = 50 across several covariance structures, and for both error distributions.

For n = 100, we consider three values of $p_n = 500$, 1000 and 2000. Again, we observe that the proportion of times ISIS based methods select the true model decreased significantly when we consider t_2 errors instead of Gaussian, as well as for Case 2. BASAD

Table 2: Proportion of times true model is selected by each method for n = 100

Gaussian error										
Methods		Case 1		Case 2		Case 3			Case 4	
$\downarrow p_n \rightarrow$	500	1000	2000	500	500	1000	2000	500	1000	2000
ISIS-SCAD-BIC	0.850	0.460	0.320	0.000	0.835	0.430	0.275	0.610	0.145	0.037
ISIS-MCP-BIC	0.670	0.255	0.230	0.000	0.615	0.240	0.156	0.065	0.000	0.000
BCR.marg	0.280	0.000	0.000	0.275	0.245	0.000	0.000	0.125	0.000	0.000
BASAD	0.985	0.240	0.000	0.935	0.975	0.270	0.000	0.975	0.300	0.000
BASAD-BIC	1.000	0.350	0.000	1.000	1.000	0.450	0.000	0.995	0.600	0.000
Proposed	1.000	0.915	0.580	0.890	1.000	0.920	0.665	1.000	0.885	0.600
t_2 error										
Methods		Case 1		Case 2		Case 3			Case 4	
$\downarrow p_n \rightarrow$	500	1000	2000	500	500	1000	2000	500	1000	2000
ISIS-SCAD-BIC	0.435	0.375	0.325	0.000	0.450	0.360	0.304	0.400	0.260	0.130
ISIS-MCP-BIC	0.385	0.325	0.290	0.000	0.400	0.325	0.275	0.255	0.180	0.080
BCR.marg	0.265	0.000	0.000	0.000	0.205	0.000	0.000	0.095	0.000	0.000
BASAD	0.905	0.060	0.000	0.105	0.875	0.115	0.000	0.755	0.212	0.000
BASAD-BIC	0.945	0.180	0.000	0.165	0.920	0.230	0.000	0.840	0.422	0.000
Proposed	0.930	0.695	0.650	0.075	0.775	0.695	0.392	0.860	0.660	0.375

performs best for the first three co-variance structures when $p_n = 500$ irrespective of the distribution of the errors. However, the proposed method outperforms BASAD for Case 4 and we now observe an improvement over BASAD even for $p_n = 500$. For $p_n = 1000$ with n = 100, this proportion again falls drastically for BASAD while our method yields a more stable performance. Interestingly, methods based on ISIS lead to comparable results in some of the cases for such high-dimensional data. We again observe a systematic improvement of the marginal version over the BCR.joint for $p_n = 2000$ (whenever the joint yielded a valid result), but the overall performance of BCR is not very good compared to other Bayesian methods. The strength of our proposed method is clear from Table 2 with higher values of p_n . In particular, when $p_n = 2000$ we observe that only our method leads to a non-zero value for the proportion among the Bayesian methods that we have studied in this paper.

To check the sensitivity of our method to the value of β_{α_c} , we have done a further simulation study. We consider Case 1 ($\Sigma_{p_n} = \mathbf{I}$) with the error distribution as normal for n = 100; and two choices of β_{α_c} . First, a set of decaying values of β_{α_c} in the range $(1, \ldots, 2)'$ and a set of increasing values of β_{α_c} in the range $(2, \ldots, 3)'$. An increment of 0.2 is taken for $p_n = 500$ so that we have $p(\alpha_c) = 5$, and an increment of 0.1 for $p_n = 1000$ and 2000 so that $p(\alpha_c) = 10$. The results are summarized in Table 3 below.

Table 3: Proportion of times true model is selected by each method for n = 100

Methods	$\beta_{\alpha_c} = (2.0, 1.8, \dots, 1)'$			$\beta_{\alpha_c} = (2.0, 2.1, \dots, 3)'$			
$\downarrow p_n \rightarrow$	500	1000	2000	500	1000	2000	
ISIS-SCAD-BIC	0.660	0.400	0.244	0.820	0.465	0.330	
ISIS-MCP-BIC	0.630	0.260	0.000	0.675	0.265	0.190	
BCR.marg	0.135	0.000	0.000	0.275	0.000	0.000	
BASAD	0.985	0.140	0.000	0.980	0.275	0.000	
BASAD-BIC	0.995	0.300	0.000	1.000	0.555	0.000	
Proposed	1.000	0.930	0.750	1.000	0.940	0.760	

The good performance of the proposed method is further re-instated from the numerical results of this table. However, we observe an improvement in BASAD-BIC for the latter choice β_{α_c} when $p_n = 1000$ than the former, which indicates that this methods is more sensitive to the actual value of the β s than the proposed method. We further notice that SCAD leads to a higher proportion than MCP for larger values of p_n .

5 Real data analysis

We have analyzed two data sets in this section. For each of these data sets, more details can be found in the paper by Song and Liang (2015).

5.1 Metabolic quantitative trait loci experiment

The first example is related to a metabolic quantitative trait loci experiment which links single nucletide polymorphisms (SNPs) data to metabolomics data. The *predictors* come from a GWAS study of the candidate genes for alanine amino transferase enzyme elevation in liver along with the mass spectroscopy metabolomics data. A total of 10000 SNPs are pre-selected as candidate predictors, and the number of subjects included in the data set is 50. The genotype of each SNP is coded as 0, 1 and 2 for homozygous rare, heterozygous and homozygous common allele, respectively. A particular metabolite bin that discriminates well between the disease status of the clinical trial's participants is selected as the *response variable*.

The SAM approach of Song and Liang (2015) selected two SNPs, rs17041311 and rs17392161. The first SNP rs7896824 has the same genotype as the SNP rs17041311, while the SNP rs17392161 shares the same genotype with eleven other SNPs rs17390419, rs12328732, rs2164473, rs322664, rs17415876, rs16950829, rs6607364, rs829156, rs829157, rs2946537 and rs9756 across all the 50 subjects. We implement the algorithm for our proposed method starting from d=5 till d=50 (which is the maximum possible value that d may attain). From our analysis, the proposed method identifies all the SNPs (two from the first group, and all the twelve from the second group) from d=25 onwards. We further observe that the proposed method consistently identifies a new set of SNPs consists of rs6704330 and rs12744386; and this is a novel set of SNPs which was not detected in the earlier study.

For the sake of comparison, we implement all the competing methods from our simulations, namely, ISIS-SCAD-BIC, ISIS-MCP-BIC, BCR.marg-BIC, BASAD and BASAD-BIC on this data. We first fix a value of the model size (d), and then a model selection method is used to obtain a subset of the predictor variables. To assess the relative performance of these methods, we compute both the mean and the median square errors based on leave-one-out cross-validation (CV). For all the methods, values

of the mean square errors turn out to be quite high. Therefore, we use the median square errors for comparison. For increasing values of d, Figure 1 below gives us an idea about the overall performance of each of these methods. Clearly, BASAD yields the lowest median square of errors, while the performance for our proposal is the second best. For BASAD-BIC and BCR, surprisingly, we observe an abrupt increase in the value of the error corresponding to d=50.

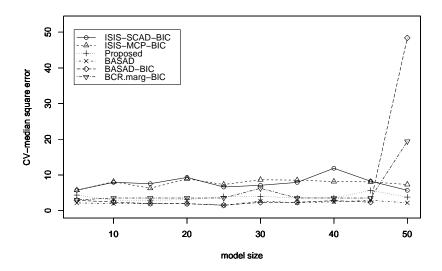


Figure 1: Comparison of the different methods using median square errors

5.2 Polymerase chain reaction

This data is related to a polymerase chain reaction. A total of 60 samples, with 31 female and 29 male mice, are used to monitor the expression levels of 22575 genes. Some physiological phenotypes, including numbers of phosphoenopyruvate carboxykinase, glycerol-3-phosphate acyltransferase and stearoyl-CoA desaturase 1 are measured by quantitative realtime polymerase chain reaction. In this data, the relationship between the gene expression level (perdictor) and phosphoenopyruvate carboxykinase (response) is studied. The gene expression data is standardized before the statistical analysis. To analyze this data, we repeat the same procedure as above.

Both BASAD and BCR could not be implemented for this data due to memory overflow for d = 22575. Figure 2 gives us the overall picture of the performance of the other

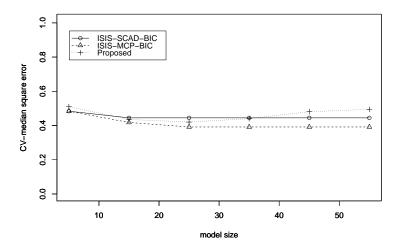


Figure 2: Comparison of the different methods using median square errors

methods, and they all yield quite low median square errors. Clearly, ISIS-MCP leads to the lowest errors, and the proposed method performs marginally better than ISIS-SCAD for d=15 and 25. However, the maximum difference in errors of the proposed method with both the methods based on ISIS is less than 0.11 over all values of d.

6 Appendix

For simplicity of presentation we drop the suffixes of p_n , g_n , X_n and $p_n(\alpha)$.

6.1 Auxiliary Results

In this section, we present auxiliary results which are used in proving the main results.

- **Lemma 6.1.** (a) Let X be a $n \times p$ matrix, such that X'X has non-zero eigenvalues $\phi_1, \phi_2, \dots, \phi_r, r \leq n$. Then $|I + X'X| = (1 + \phi_1)(1 + \phi_2) \dots (1 + \phi_r)$, where |A| is the determinant of the matrix A.
 - (b) Let X_1 be a sub-matrix of rank r of the matrix X constructed by taking a subset of the columns of X. If X'X has the non-zero eigenvalues $\phi_1, \phi_2, \ldots, \phi_m$ $(r \leq m)$, then $(1+\phi_{\min})^r \leq |I+X_1'X_1| \leq (1+\phi_{\max})^r$, where ϕ_{\max} and ϕ_{\min} are respectively the highest and lowest eigen values of X'X.

Lemma 6.2. If W follows a non-central χ^2 distribution with degrees of freedom (df) r and non-centrality parameter (ncp) λ , then

$$P(W > t) \le P\left(\chi_r^2 > \frac{t - \lambda}{2}\right) + P\left(Z > \frac{t - \lambda}{4\sqrt{\lambda}}\right),$$

where χ_r^2 denotes a central χ^2 random variable with df r and $Z \sim N(0,1)$.

Lemma 6.3. Let M_{α} , $\alpha \in \mathcal{A}$, be any model and $M_{\alpha'}$ be a model such that $n\tau'_{\min}I \leq X'_{\alpha'}X_{\alpha'} \leq n\tau'_{\max}I$ for some τ'_{\min} , and τ'_{\max} , free of n. Then under assumption (A3),

$$\frac{|I + gX'_{\alpha}X_{\alpha}|^{-1}}{|I + gX'_{\alpha'}X_{\alpha'}|^{-1}} \le c(ng)^{r_{\alpha'}-r_{\alpha}} \tau'_{\max}^{r_{\alpha'}} \tau'_{\min}^{-(r_{\alpha}+r_{1})} \tau_{\max}^{r_{1}}$$

for sufficiently large p, where $r_1 = r(X_{\alpha \wedge \alpha'})$ and c > 1 is some constant.

Lemma 6.4. If M_{α_c} be the true model, then under the setup (2.1) and assumptions (A2) and (A5), and for any $\epsilon > 0$, the probabilities of the following three events

(a)
$$R_{\alpha_c}^{2*} - R_{\alpha_c}^2 > \epsilon$$
, (b) $R_{\alpha_c}^2 > n(1+\epsilon)\sigma^2$, and (c) $R_{\alpha_c}^2 < n(1-\epsilon)\sigma^2$, are tending to 0 exponentially in n .

Lemma 6.5. Let A_1 be the set of all super models of the true model M_{α_c} , A_1^* be the subset of A_1 containing models of rank at most d, for some $d > p(\alpha_c)$, and $A_2 = \{\alpha : M_{\alpha_c} \nsubseteq M_{\alpha}, r_{\alpha} > K_0 p(\alpha_c)\}$. Then the following statements hold.

(a) For any R > 2, with probability tending to 1

$$\max_{\alpha \in \mathcal{A}_1} (R_{\alpha_c}^2 - R_{\alpha}^2) \le R\sigma^2(r_{\alpha} - p(\alpha_c)) \log p.$$

- (b) For any $\alpha \in \mathcal{A}_1^*$ and any $\epsilon > 0$, the probability that $R_{\alpha}^{2*} R_{\alpha}^2 > \epsilon$, tends to 0 exponentially in n.
- (c) For $R > 2K_0/(K_0 1)$, with probability tending to 1,

$$\max_{\alpha \in A_2} (R_{\alpha_c}^2 - R_{\alpha_c \vee \alpha}^2) \le R\sigma^2(r_\alpha - p(\alpha_c)) \log p.$$

Lemma 6.6. Let $\mathbf{y}_n = \boldsymbol{\mu}_n + \mathbf{e}_n$ with $\mathbf{e}_n \sim N(\mathbf{0}, \sigma^2 I)$ and $\boldsymbol{\mu}'_n \boldsymbol{\mu}_n = O(n)$. For any h_n , such that $h_n = n^k$ for some 0.5 < k < 1, we have $|\boldsymbol{\mu}'_n \mathbf{e}_n| = o_p(h_n)$.

The proofs of Lemma 6.1-6.6 are given in the supplementary file.

6.2 Main Results

Proof of Theorem 3.1. The ratio of posterior probabilities of any model to the true model is given by

$$\frac{P(M_{\alpha}|\mathbf{y}_n)}{P(M_{\alpha_c}|\mathbf{y}_n)} = \left(\frac{1}{p_n - 1}\right)^{p_n(\alpha) - p(\alpha_c)} \exp\left\{-\frac{R_{\alpha}^{2*} - R_{\alpha_c}^{2*}}{2\sigma^2}\right\} \frac{\left|I + g_n X_{\alpha_c}' X_{\alpha_c}\right|^{1/2}}{\left|I + g_n X_{\alpha}' X_{\alpha}\right|^{1/2}}. \quad (6.1)$$

We split A into three subclasses as follows:

- (i) Supermodel of the true model, $A_1 = \{\alpha : M_{\alpha_c} \subset M_{\alpha}\}.$
- (ii) Non-supermodel of large dimension, $A_2 = \{\alpha : M_{\alpha_c} \nsubseteq M_{\alpha}; r_{\alpha} > K_0 \ p(\alpha_c)\}$ where r_{α} is the rank of X_{α} .
- (iii) Non-supermodel of small to moderate rank, $A_3 = \{\alpha : M_{\alpha_c} \nsubseteq M_{\alpha}; r_{\alpha} \le K_0 p(\alpha_c)\}$. We prove (3.1) separately for models in $A = A_i$, for i = 1, 2, 3.

Case I: Super-models ($\alpha \in A_1$) First, we obtain a uniform upper bound for ratio of the posterior probabilities of any model M_{α} and M_{α_c} , given in (6.1). Note that

$$R_{\alpha_c}^{2*} - R_{\alpha}^{2*} \le R_{\alpha_c}^{2*} - R_{\alpha}^2 = R_{\alpha_c}^{2*} - R_{\alpha_c}^2 + R_{\alpha_c}^2 - R_{\alpha}^2,$$

where $R_{\alpha}^{2} = \mathbf{y}_{n}' \left\{ I - X_{\alpha} \left(X_{\alpha}' X_{\alpha} \right)^{-1} X_{\alpha}' \right\} \mathbf{y}_{n} \text{ and } R_{\alpha}^{2*} \geq R_{\alpha}^{2}.$

By part (a) of Lemma 6.4 we have $R_{\alpha_c}^{2*} - R_{\alpha_c}^2 = o_p(1)$. By part (a) of Lemma 6.5, for $\alpha \in \mathcal{A}_1$ and some $R = 2(1 + \epsilon)$, $\epsilon > 0$, we have $\max_{\alpha \in \mathcal{A}_1} \left(R_{\alpha_c}^2 - R_{\alpha}^2 \right) > R\sigma^2(r_{\alpha} - p(\alpha_c)) \log p$. Therefore, for any $\epsilon > 0$

$$\exp\left\{-\frac{1}{2\sigma^2}(R_{\alpha}^{2*}-R_{\alpha_c}^{2*})\right\} \leq p^{R(r_{\alpha}-p(\alpha_c))/2+o_p(1)}.$$

Again, by Lemma 6.3 and assumptions (A2)-(A3) we have

$$\frac{\left|I + g X_{\alpha}' X_{\alpha}\right|^{-1/2}}{\left|I + g X_{\alpha_c}' X_{\alpha_c}\right|^{-1/2}} \le c^* (ng\tau_{\min})^{-(r_{\alpha} - p(\alpha_c))/2} \tau_{\max}^{p(\alpha_c)/2} \le c^* (ng\tau_{\min})^{-(r_{\alpha} - p(\alpha_c))/2} p^{o(1)},$$

where c^* is some appropriate constant. Therefore, summing the ratio of posterior probabilities over $M_{\alpha} \in \mathcal{A}_1$, we have

$$\sum_{\alpha \in \mathcal{A}_1} \frac{p(M_{\alpha}|\mathbf{y}_n)}{p(M_{\alpha_c}|\mathbf{y}_n)} \leq \sum_{\alpha \in \mathcal{A}_1} \frac{c^* p^{(1+\epsilon)(r_{\alpha}-p(\alpha_c))+o_p(1)+o(1)}}{(p-1)^{p_n(\alpha)-p(\alpha_c)} (\tau_{\min} n \ g)^{(r_{\alpha}-p(\alpha_c))/2}}$$

$$\leq \sum_{\alpha \in \mathcal{A}_1} \left(\sqrt{\frac{p^{2+\delta_1/3}}{p^{2+\delta_1}}} \right)^{r_\alpha - p(\alpha_c)} \frac{1}{(p-1)^{p_n(\alpha) - p(\alpha_c)}},$$

for some suitably chosen $\epsilon > 0$. This is due to the fact that we can choose ϵ so that the term $\epsilon + o_p(1) + o(1) < \delta_1/3$, for sufficiently large p. By assumption (A2), the true model is unique and is of full rank, and therefore $r_{\alpha} - p(\alpha_c) \ge 1$. Thus, the above expression is less than

$$p^{-\delta_1/3} \sum_{q=1}^{p-p(\alpha_c)} \binom{p-p(\alpha_c)}{q} \frac{1}{(p-1)^q} \le p^{-\delta_1/3} \left\{ \left(1 + \frac{1}{p-1}\right)^p - 1 \right\},\,$$

and this tends to 0 as $p \to \infty$.

Case II: Non-super models of large dimension ($\alpha \in A_2$) We split $R_{\alpha_c}^{2*} - R_{\alpha}^{2*}$ as before and use the fact that $R_{\alpha \vee \alpha_c}^2 \leq R_{\alpha}^2$. Thus, we have

$$R_{\alpha_c}^{2*} - R_{\alpha}^{2*} \leq R_{\alpha_c}^{2*} - R_{\alpha}^2 \leq R_{\alpha_c}^{2*} - R_{\alpha_c}^2 + R_{\alpha_c}^2 - R_{\alpha\vee\alpha_c}^2.$$

From part (c) of Lemma 6.5, with probability tending to 1, $R_{\alpha_c}^2 - R_{\alpha \vee \alpha_c}^2 \leq R\sigma^2(r_\alpha - p(\alpha_c))\log p$ for R = 2(1+s) with $s > 1/(K_0-1)$. Using Lemma 6.3, along with assumptions (A2)-(A3) as in the previous case, we have

$$\sum_{\alpha \in \mathcal{A}_{2}} \frac{p(M_{\alpha}|\mathbf{y}_{n})}{p(M_{\alpha_{c}}|\mathbf{y}_{n})} \leq \sum_{\alpha \in \mathcal{A}_{2}} \frac{c^{*}p^{(1+s)(r_{\alpha}-p(\alpha_{c}))+o_{p}(1)+o(1)}}{(p-1)^{p_{n}(\alpha)-p(\alpha_{c})} (\tau_{\min}n \ g)^{(r_{\alpha}-p(\alpha_{c}))/2}} \\
\leq \sum_{\alpha \in \mathcal{A}_{2}} c^{*} \left(\frac{p^{1+6/(5(K_{0}-1))}}{\sqrt{ng}}\right)^{r_{\alpha}-p(\alpha_{c})} p^{p(\alpha_{c})} \frac{1}{(p-1)^{p_{n}(\alpha)}},$$

for an appropriately chosen c^* and s so that $s+o_p(1)+o(1) \leq 6/(5(K_0-1))$ for sufficiently large n. As $r_{\alpha}-p(\alpha_c)>(K_0-1)p(\alpha_c)$, the above expression is less than

$$c^* \left(\frac{p^{1+11/(5(K_0-1))}}{\sqrt{ng}} \right)^{(K_0-1)p(\alpha_c)} \sum_{q=K_0t+1}^p \binom{p}{q} \frac{1}{(p-1)^q}.$$
 (6.2)

Also $\delta_1 \geq 5/(K_0-1)$, and so the second term in the above expression is no bigger than $p^{-3p(\alpha_c)/10}$, which converges to 0 as $p \to \infty$. However, the third term is dominated by $\sum_{q=1}^{p} \binom{p}{q} (p-1)^{-q}$, which converges to e as $p \to \infty$. The above facts together imply that (6.2) converges to 0 as $p \to \infty$.

Case III: Non-super models of small to moderate rank ($\alpha \in \mathcal{A}_3$) As in the previous case, we have $R_{\alpha_c}^{2*} - R_{\alpha}^{2*} \leq R_{\alpha_c}^{2*} - R_{\alpha_c}^2 + R_{\alpha_c}^2 - R_{\alpha}^2$.

By part (a) of Result 6.4, we have $R_{\alpha_c}^{2*} - R_{\alpha_c}^2 = o_p(1)$. Next, consider the third part in the right hand side of the above expression.

$$R_{\alpha}^{2} - R_{\alpha_{c}}^{2} = \mathbf{y}_{n}'(P_{n}(\alpha_{c}) - P_{n}(\alpha))\mathbf{y}_{n}$$

$$= \boldsymbol{\mu}_{n}'(P_{n}(\alpha_{c}) - P_{n}(\alpha))\boldsymbol{\mu}_{n} + 2\boldsymbol{\mu}_{n}'(P_{n}(\alpha_{c}) - P_{n}(\alpha))\mathbf{e}_{n} + \mathbf{e}_{n}'(P_{n}(\alpha_{c}) - P_{n}(\alpha))\mathbf{e}_{n}.$$

Note that $\mu'_n(P_n(\alpha_c) - P_n(\alpha))\mu_n = \mu'_n(I - P_n(\alpha))\mu_n > \Delta_0$ by assumption (A4). Again,

$$\mu'_n(P_n(\alpha_c) - P_n(\alpha))\mathbf{e}_n = \mu'_n(P_n(\alpha_c) - P_n(\alpha \vee \alpha_c))\mathbf{e}_n + \mu'_n(P_n(\alpha \vee \alpha_c) - P_n(\alpha))\mathbf{e}_n \ge -2|\mu'_n\mathbf{e}_n|.$$

By Lemma 6.6, $|\boldsymbol{\mu}_n' \mathbf{e}_n| = o_p(h_n)$ for $h_n = n^d$ for some 0.5 < d < 1. Finally, we get

$$\mathbf{e}'_n(P_n(\alpha_c) - P_n(\alpha))\mathbf{e}_n \ge -\mathbf{e}'_n(P_n(\alpha \vee \alpha_c) - P_n(\alpha))\mathbf{e}_n.$$

As $P_n(\alpha \vee \alpha_c) - P_n(\alpha)$ is an idempotent matrix, we have $\mathbf{e}'_n(P_n(\alpha \vee \alpha_c) - P_n(\alpha))\mathbf{e}_n \geq 0$. Note that $\mathbf{e}'_n(P_n(\alpha \vee \alpha_c) - P_n(\alpha))\mathbf{e}_n \leq \mathbf{e}'_nP_n(\alpha_c)\mathbf{e}_n$ for any $\alpha \in \mathcal{A}_3$ (see Section 2.3.2 of Yanai et al. (2011)). Also, $\mathbf{e}'_nP_n(\alpha_c)\mathbf{e}_n = O_p(1)$ since it follows the $\sigma^2\chi^2$ distribution with df $p(\alpha_c)$. Combining all these facts and using Assumption (A4), we have

$$R_{\alpha_c}^{2*} - R_{\alpha}^{2*} \le -\Delta_0(1 + o_p(1)).$$

Further, from Lemma 6.3, the ratio of determinants in the last term of (6.1) is less than $c^* \left(\sqrt{ng\tau_{max}} \right)^{p(\alpha_c)}$ for an appropriately chosen $c^* > 0$. Therefore,

$$\sum_{\alpha \in \mathcal{A}_3} \frac{p(M_{\alpha}|\mathbf{y}_n)}{p(M_{\alpha_c}|\mathbf{y}_n)} \le c^* \left(p\sqrt{ng\tau_{max}}\right)^{p(\alpha_c)} \exp\left\{-\frac{\Delta_0}{2\sigma^2}(1+o_p(1))\right\} \sum_{q=1}^p \binom{p}{q} \frac{1}{(p-1)^q}. (6.3)$$

For sufficiently large p, we have $c^*\left(p\sqrt{ng\tau_{max}}\right)^{p(\alpha_c)} \leq p^{2(1+\delta_1/3)p(\alpha_c)}$. By assumption (A4), $\exp\{-\Delta_0/(2\sigma^2)\} \leq p^{-3p(\alpha_c)}$. Thus the product of first three terms in the right hand side (rhs) of (6.3) converges to zero, whereas the last term converges to e. Using the above facts it is evident that the rhs of (6.3) is less than $p^{-(1-\delta_1/3)p(\alpha_c)}$. As $p \to \infty$, the result follows.

Proof of Theorem 3.2. First note that M_{α_1} and M_{α_2} are of the dimension d, and d is a constant free of n. Therefore, the ranks of both the models r_{α_1} and r_{α_2} , are also

free of n. We now have

$$\sup_{\alpha_1, \alpha_2} \frac{m(M_{\alpha_2} | \mathbf{y}_n)}{m(M_{\alpha_1} | \mathbf{y}_n)} = \sup_{\alpha_1, \alpha_2} \exp\left\{ -\frac{1}{2\sigma^2} (R_{\alpha_2}^{*2} - R_{\alpha_1}^{*2}) \right\} \frac{\left| I + gX_{\alpha_2}' X_{\alpha_2} \right|^{-1/2}}{\left| I + gX_{\alpha_1}' X_{\alpha_1} \right|^{-1/2}}.$$
 (6.4)

By assumptions (A2) and (A3) and Lemma 6.3, we get

$$\sup_{\alpha_1,\alpha_2} \frac{\left|I + gX'_{\alpha_1}X_{\alpha_1}\right|^{1/2}}{\left|I + gX'_{\alpha_2}X_{\alpha_2}\right|^{1/2}} \le \left(\sqrt{ng}\right)^{r_{\alpha_1} - r_{\alpha_2}} \tau_{\max}^{r_1 + p(\alpha_c)} \tau_{\min}^{r_1 - r_{\alpha_2}} (1 + \xi_n) \le \left(\sqrt{ng}\right)^{r_{\alpha_1} - r_{\alpha_2} + o(1)}$$

where $r_1 = rank(X_{\alpha_1 \wedge \alpha_2})$. We also have

$$R_{\alpha_1}^{*2} - R_{\alpha_2}^{*2} \le R_{\alpha_1}^{*2} - R_{\alpha_2}^2 = R_{\alpha_1}^{*2} - R_{\alpha_1}^2 + R_{\alpha_1}^2 - R_{\alpha_2}^2 + R_{\alpha_2}^2 - R_{\alpha_2}^2$$

From part (b) of Lemma 6.5, we get $\sup_{\alpha_1} (R_{\alpha_1}^{*2} - R_{\alpha_1}^2) = o_p(1)$. By the properties of projection matrices, $R_{\alpha_1}^2 - R_{\alpha_c}^2 \le 0$. Next consider $R_{\alpha_2}^2 - R_{\alpha_c}^2$ which is equal to

$$\mathbf{y}_n'(P_n(\alpha_c) - P_n(\alpha_2))\mathbf{y}_n$$

$$= \boldsymbol{\mu}_n'(P_n(\alpha_c) - P_n(\alpha_2))\boldsymbol{\mu}_n + 2\boldsymbol{\mu}_n'(P_n(\alpha_c) - P_n(\alpha_2))\mathbf{e}_n + \mathbf{e}_n'(P_n(\alpha_c) - P_n(\alpha_2))\mathbf{e}_n.$$

We now have

$$\mu'_n(P_n(\alpha_c) - P_n(\alpha_2))\mathbf{e}_n$$

$$= \mu'_n(P_n(\alpha_c) - P_n(\alpha_2 \vee \alpha_c))\mathbf{e}_n + \mu'_n(P_n(\alpha_2 \vee \alpha_c) - P_n(\alpha))\mathbf{e}_n \ge -2|\mu'_n\mathbf{e}_n|.$$

By Lemma 6.6, $|\boldsymbol{\mu}_n' \mathbf{e}_n| = o_p(h_n)$ for $h_n = n^k$ for some 0.5 < k < 1. Finally,

$$\mathbf{e}'_n(P_n(\alpha_c) - P_n(\alpha_2))\mathbf{e}_n \ge -\mathbf{e}'_n(P_n(\alpha_2 \vee \alpha_c) - P_n(\alpha_c))\mathbf{e}_n$$

as $\mathbf{e}'_n(P_n(\alpha_2 \vee \alpha_c) - P_n(\alpha_2))\mathbf{e}_n \geq 0$. Note that $\mathbf{e}'_n(P_n(\alpha_2 \vee \alpha_c) - P_n(\alpha_c))\mathbf{e}_n \leq \mathbf{e}'_nP_n(\alpha_c)\mathbf{e}_n$, and $\mathbf{e}'_nP_n(\alpha_c)\mathbf{e}_n = O_p(1)$.

Again, by assumption (A4), we have $\mu'_n(P_n(\alpha_c) - P_n(\alpha_2))\mu_n \geq \Delta_0$. Combining the above statements and using assumption (A5), from (6.4) we have

$$\sup_{\alpha_1,\alpha_2} \frac{m(M_{\alpha_2}|\mathbf{y}_n)}{m(M_{\alpha_1}|\mathbf{y}_n)} \leq (\sqrt{ng})^{p(\alpha_c)} p^{o(1)} \exp\left\{-\frac{\Delta_0}{2\sigma^2}(1+o_p(1))\right\} \leq p^{-(1-\delta_1/2+o_p(1))p(\alpha_c)}$$
 which converges to 0 as $p \to \infty$.

Supplementary Material

(https://drive.google.com/open?id=0By7-ldtnmyfvUjlpWUZnaGpwNWs). The Supplementary Material contains proofs of all the Lemmas and Theorem 3.3. A pdf copy is available at the link mentioned above.

References

- Bondell, H. D. and Reich, B. J. (2012). "Consistent high-dimensional Bayesian variable selection via penalized credible regions." *J. Amer. Statist. Assoc.*, 107(500): 1610–1624.
- Candès, E. and Tao, T. (2007). "Rejoinder: "The Dantzig selector: statistical estimation when p is much larger than n" [Ann. Statist. **35** (2007), no. 6, 2313–2351; MR2382644]." Ann. Statist., 35(6): 2392–2404.
- Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015). "Bayesian linear regression with sparse priors." *Ann. Statist.*, 43(5): 1986–2018.
- Chen, J. and Chen, Z. (2012). "Extended BIC for small-n-large-P sparse GLM." Statist. Sinica, 22(2): 555–574.
- Chipman, H., George, E. I., and McCulloch, R. E. (2001). "The practical implementation of Bayesian model selection." In *Model selection*, volume 38 of *IMS Lecture Notes Monogr. Ser.*, 65–134. Inst. Math. Statist., Beachwood, OH.
- Fan, J., Feng, Y., and Song, R. (2011). "Nonparametric independence screening in sparse ultra-high-dimensional additive models." *J. Amer. Statist. Assoc.*, 106(494): 544–557.
- Fan, J. and Li, R. (2001). "Variable selection via nonconcave penalized likelihood and its oracle properties." *J. Amer. Statist. Assoc.*, 96(456): 1348–1360.
- Fan, J. and Lv, J. (2008). "Sure independence screening for ultrahigh dimensional feature space." J. R. Stat. Soc. Ser. B Stat. Methodol., 70(5): 849–911.
- (2010). "A selective overview of variable selection in high dimensional feature space." Statist. Sinica, 20(1): 101–148.
- Fan, J. and Song, R. (2010). "Sure independence screening in generalized linear models with NP-dimensionality." *Ann. Statist.*, 38(6): 3567–3604.
- George, E. I. and Foster, D. P. (2000). "Calibration and empirical Bayes variable selection." *Biometrika*, 87(4): 731–747.
- Ishwaran, H. and Rao, J. S. (2005). "Spike and slab variable selection: frequentist and

- Bayesian strategies." Ann. Statist., 33(2): 730–773.
- Johnson, V. E. and Rossell, D. (2012). "Bayesian model selection in high-dimensional settings." J. Amer. Statist. Assoc., 107(498): 649–660.
- Liang, F., Song, Q., and Yu, K. (2013). "Bayesian subset modeling for high-dimensional generalized linear models." J. Amer. Statist. Assoc., 108(502): 589–606.
- Moreno, E., Girón, J., and Casella, G. (2015). "Posterior model consistency in variable selection as the model dimension grows." *Statist. Sci.*, 30(2): 228–241.
- Narisetty, N. N. and He, X. (2014). "Bayesian variable selection with shrinking and diffusing priors." *Ann. Statist.*, 42(2): 789–817.
- Rosset, S. (2013). "Practical Sparse Modeling: an Overview and Two Examples from Genetics." Chapter 3 in Practical Applications of Sparse Modeling, I. Rish et al. (eds.), MIT Press.
- Song, Q. and Liang, F. (2015). "A split-and-merge Bayesian variable selection approach for ultrahigh dimensional regression." J. R. Stat. Soc. Ser. B. Stat. Methodol., 77(5): 947–972.
- Song, R., Yi, F., and Zou, H. (2014). "On varying-coefficient independence screening for high-dimensional varying-coefficient models." *Statist. Sinica*, 24(4): 1735–1752.
- Wang, H. (2009). "Forward regression for ultra-high dimensional variable screening." J. Amer. Statist. Assoc., 104(488): 1512–1524.
- Yanai, H., Takeuchi, K., and Takane, Y. (2011). Projection matrices, generalized inverse matrices, and singular value decomposition. Statistics for Social and Behavioral Sciences. Springer, New York.
- Yuan, M. and Lin, Y. (2006). "Model selection and estimation in regression with grouped variables." J. R. Stat. Soc. Ser. B Stat. Methodol., 68(1): 49–67.
- Zhang, C.-H. (2010). "Nearly unbiased variable selection under minimax concave penalty." *Ann. Statist.*, 38(2): 894–942.
- Zou, H. (2006). "The adaptive lasso and its oracle properties." J. Amer. Statist. Assoc., 101(476): 1418–1429.

Zou, H. and Hastie, T. (2005). "Regularization and variable selection via the elastic net." J. R. Stat. Soc. Ser. B Stat. Methodol., 67(2): 301–320.

Acknowledgments

The authors are sincerely thankful to Prof. Tapas Samanta for his insightful comments, and guidance towards this article.