Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 485–508
http://www.ssca.org.in/journal

Bayesian Variable Selection for Ultrahigh-dimensional Sparse Linear Models

Minerva Mukhopadhyay[1] **and Subhajit Dutta**[2]
[1]*Interdisciplinary Statistical Research Unit,*
Indian Statistical Institute Kolkata, 203 B. T. Road, Kolkata – 700108, WB, India.
[2]*Applied Statistics Unit,*
Indian Statistical Institute Kolkata, 203 B. T. Road, Kolkata – 700108, WB, India.
[1,2]*Department of Mathematics and Statistics,*
Indian Institute of Technology Kanpur, Kanpur - 208016, UP, India.

Received: 12 August 2024; Revised: 25 September 2024; Accepted: 30 September 2024

Abstract

We consider the problem of variable selection for the ultrahigh-dimensional linear regression model, allowing the number of covariates p_n to grow exponentially with n. Assuming the true model to be sparse, we propose a set of priors suitable for this regime. In the ultrahigh-dimensional setting, the selection of the unique true model among all the 2^{p_n} possible ones involves prohibitive computation. To cope with this, a two-stage model selection algorithm is proposed. In the first stage, an efficient screening algorithm is employed to find a *good* d_n-dimensional model, where $d_n \ll n$. In the next stage, an explicit model search algorithm is employed on the space of all submodels of the first-stage-selected model. Theoretical investigations justify the two-stage procedure. It is demonstrated that the first-stage screening is expected to select a supermodel of the true model, consequently, the second-stage algorithm identifies the true model with probability tending to one. This procedure is computationally efficient, simple and intuitive. We validate the competitive performance of the proposed algorithm with a variety of simulated and real data sets, and compare with several frequentist as well as Bayesian methods.

Key words: Model selection consistency; Reversible jump MCMC; Screening consistency.

AMS Subject Classifications: 62H05; 11P70; 47A57.

1. Introduction

Variable selection in ultrahigh-dimensional regression setup has become a flourishing area in the contemporary research, due to increasing availability of data in various fields like genetics, finance, machine learning. Consider, for example, in genome-wide association studies (GWAS), where a phenotype is measured for a panel of individuals and a large number of single nucleotide polymorphisms (SNPs) are genotyped for each individual. The goal is to identify SNPs that are statistically associated with the phenotype. Sparsity has

Corresponding Author: Minerva Mukhopadhyay
Email: minervam@iitk.ac.in

frequently been identified as an underlying feature for such data sets, where among a large number of covariates (SNPs) only a small subset are actually important.

Several variable selection methods have been proposed for high-dimensional data in both the frequentist and the Bayesian paradigms. Two predominant classes of methods in frequentist paradigm are penalized likelihood methods and screening based methods. Penalized likelihood methods includes Least Absolute Shrinkage and Selection Operator (LASSO) and its variants like the elastic net of Zou and Hastie (2005), the group LASSO of Yuan and Lin (2006) and the adaptive LASSO of Zou (2006), *etc.*, while the screening based methods include 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.* For a comprehensive review of frequestist variable selection method, see Bühlmann and van de Geer (2011).

In situations with extreme sparsity LASSO-type estimates are outperformed by testing-based subset selection methods (see, for example (Tibshirani, 1996, Section 11)), and tend to overfit. On the other hand, screening based methods focus on marginal association of covariates with the response, and therefore fail to capture the joint structure of the covariates. As a result these methods suffer under presence of multicollinearity, which is almost inenviable in high-dimensional scenario.

In the Bayesian literature, popular methods include the empirical Bayes variable selection (see George and Foster (2000)), where a mixture of testing and optimization is employed to identify the optimal model, fully testing-based methods like spike and slab variable selection (see Ishwaran and Rao (2005)), and optimization and thresholding-based shrinkage prior methods for variable selection like Bayesian LASSO (see Park and Casella (2008)). 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.

Among notable theoretical developments, Castillo *et al.* (2015) proved results related to the posterior consistency for regression parameters, while Liang *et al.* (2013) have shown the equivalence of posterior consistency and model selection consistency under appropriate sparsity assumptions. Narisetty and He (2014) claim to prove the *'strongest selection consistency result'* using the spike and slab prior under under the $\log p_n = o(n)$ setting.

Although the optimization based methods are fast and easily implementable to high-dimensional framework, strong selection consistency property is usually not investigated for these methods. Strong selection consistency, requiring posterior probability of the true model stochastically converging to one, has been shown in Narisetty and He (2014), however, for implementation they rely on the stochastic search variable selection (SSVS) algorithm which is not scalable in high-dimensional situations.

Neighborhood search based SSVS algorithms for the optimal model search are routine for small values of p_n and n, but the resulting computations are quite intensive for higher dimensions due to a large number of possible models. Several authors have developed methods to cope with the high-dimensionality, *e.g.*, Shin *et al.* (2018) proposed a simplified shotgun stochastic search and screening algorithm that employs a variable screening to reduce neighborhood size in the SSVS algorithm, Li *et al.* (2023) have proposed a highly scalable model-based screening method to explore model space efficiently.

In this paper, we propose a Bayesian method for variable selection and examine its properties both theoretically and numerically, under sparsity assumption. Considering the popular Zellner's g_n-prior (Zellner, 1986) framework, we propose a prior setup suitable for the ultrahigh-dimensional situation. The proposed set of priors has the advantage of generating closed form expressions of the marginals, which makes the resultant method as tractable as the simple information criterion based methods like AIC or BIC.

In a $p_n \gg n$ setting, the size of the model space becomes gigantic and a simple SSVS algorithm can not identify the true model in a finite time. To cope with this situation, we present a two-stage model selection procedure based on an initial screening. The first stage algorithm is intended to select a *good* d_n-dimensional model, where $d_n \ll n$. Under the sparsity assumption, the posterior probability of the class of d_n-dimensional supermodels of the true model uniformly dominates that of of all d_n-dimensional models. Motivated by this result, we first employ a model search algorithm on the space of all d_n-dimensional models. Given an initial model, the algorithm transits to the neighboring d_n-dimensional model with the highest posterior probability. Due to the uniform dominance of the class of supermodels and the less challenging goal of selecting any model in this class, the first-stage algorithm selects a d_n-dimensional supermodel quite efficiently while taking care of joint structure of the covariates, unlike the other screening methods which rely on marginal information.

In the second stage, an SSVS algorithm is employed to search the space of submodels of the first-stage-selected model. Given that a supermodel of the true model is selected at the first stage, the second stage algorithm identifies the true model quite efficiently as $d_n \ll n$. The proposed two-stage algorithm is fast and intuitive. Its good performance is supported by theoretical results under the $\log p_n = O(n)$ settings. To the best of our knowledge, this is the first work on exponential growth of covariates with sample size. The performance of the algorithm is validated extensively with ample simulated and real data sets.

In Section 2, the prior setup and the maximum-a-posteriori (MAP) approach are described. In Section 3, the two-stage algorithm is introduced. Section 4 contains the theoretical results justifying the proposed two-stage algorithm. In Sections 5 and 6, the performance of the proposed algorithm is validated using simulated and real data sets. Section 7 contains concluding remarks. Proofs of all the theoretical results are provided in Section A.

2. The proposed prior setup and the MAP approach

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

$$\mathbf{y}_n = X_n\boldsymbol{\beta} + \mathbf{e}_n, \tag{1}$$

where X_n is the $n \times p_n$ design matrix, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_{p_n})'$ is the vector of regression parameters and $\mathbf{e}_n$ is the vector of random errors. 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 models that can be formed by taking at least one covariate is denoted by $\mathcal{G}$, and indexed by γ. Here, $\gamma \in \mathcal{G}$ is a subset of $\{1, \ldots, p_n\}$ of size $p_n(\gamma)$ $(1 \leq p_n(\gamma) \leq$

p_n), indicating the index set of the covariates corresponding to the model M_γ. Under M_γ, we assume $\mathbf{y}_n = X_\gamma \boldsymbol{\beta}_\gamma + \mathbf{e}_n$, where X_γ is a sub-matrix of X_n consisting of the $p_n(\gamma)$ columns specified by γ and $\boldsymbol{\beta}_\gamma$ is the corresponding vector of regression coefficients. We consider the problem of selecting the sparsest model M_γ with $\gamma \in \mathcal{G}$ that best explains the data.

In a Bayesian approach, each model M_γ is assigned a prior probability and the corresponding set of parameters $\boldsymbol{\theta}_\gamma = (\beta_0, \boldsymbol{\beta}_\gamma, \sigma^2)'$ involved in M_γ, is also assigned a prior distribution. Given prior probability $P(M_\gamma)$ on M_γ and conditional prior density $p(\boldsymbol{\theta}_\gamma | M_\gamma)$ on $\boldsymbol{\theta}_\gamma$ under M_γ, one computes the posterior probability of each model as follows

$$P(M_\gamma | \mathbf{y}_n) = \frac{P(M_\gamma) m_\gamma(\mathbf{y}_n)}{\sum_{\gamma \in \mathcal{G}} P(M_\gamma) m_\gamma(\mathbf{y}_n)}, \qquad \text{where} \qquad m_\gamma(\mathbf{y}_n) = \int p(\mathbf{y}_n | \boldsymbol{\theta}_\gamma, M_\gamma) p(\boldsymbol{\theta}_\gamma | M_\gamma) d\boldsymbol{\theta}_\gamma$$

is the marginal likelihood and $p(\mathbf{y}_n | \boldsymbol{\theta}_\gamma, M_\gamma)$ is the density of $\mathbf{y}_n$ under M_γ. We consider the maximum a-posteriori (MAP) approach which selects the model $\gamma^\star$ in $\mathcal{G}$ with the highest posterior probability as the optimal model.

Throughout this paper, we have considered the following notations and conventions. For two numbers a and b, the notations $a \vee b$ and $a \wedge b$ denote $\max\{a, b\}$ and $\min\{a, b\}$, respectively. For two sequences of real numbers $\{a_n\}$ and $\{b_n\}$, $a_n \lesssim b_n$ indicates either $a_n/b_n \to 0$ or $a_n \leq c b_n$ for all sufficiently large n, and some constant $0 < c < \infty$. Further, if $a_n \gtrsim b_n$ and $b_n \lesssim a_n$, then we write $a_n \sim b_n$. For any square matrix A, $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ are the highest and the lowest non-zero eigenvalues of A. For two square matrices A and B of the same order, $A \leq B$ means that $B - A$ is positive semidefinite. A model M_γ with dimension $p_n(\gamma) < n$ is said to be of full-rank if $\text{rank}(X'_\gamma X_\gamma) = p_n(\gamma)$.

2.1. Prior specification and posterior probability

Each model M_γ with $\gamma \in \mathcal{G}$ is assigned *Bernoulli* prior $P(M_\gamma) = q_n^{p_n(\gamma)} (1 - q_n)^{p_n - p_n(\gamma)}$ with $q_n = 1/p_n$. Given a model M_γ, we consider a conjugate prior on $\boldsymbol{\beta}_\gamma$ as

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

where g_n is a hyperparameter. We impose the popular Jeffreys prior $\pi(\sigma^2) \propto 1/\sigma^2$ on σ^2.

The Bernoulli prior is widely used as a model prior probability because of its property of penalizing the models of large 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, as it assigns $1/p_n$ weight to each covariate. Thus, the prior probability of a model increases p_n times if one covariate is dropped.

Use of the inverse-gamma prior for error variance is fairly conventional in the literature (see, *e.g.*, George and Mcculloch (1993)). The Jeffreys prior is the limit of inverse-gamma, as both the hyperparameters in the inverse-gamma prior approach zero. The property of invariance under reparametrization makes it suitable as a prior on scale parameter.

For the proposed set of priors, the posterior probability of the model M_γ is

$$P\left(M_\gamma | \mathbf{y}_n\right) \;\propto\; \left(\frac{1}{p_n - 1}\right)^{p_n(\gamma)} \left|I + g_n X_\gamma' X_\gamma\right|^{-1/2} \left(R_\gamma^{2\star}\right)^{-n/2}, \tag{2}$$

where $R_\gamma^{2\star} = \mathbf{y}_n' \left\{I_n - X_\gamma \left(I_{p_n(\gamma)}/g_n + X_\gamma' X_\gamma\right)^{-1} X_\gamma'\right\} \mathbf{y}_n$.

Our prior choices are simple. Except the choice of g_n, the set of priors is completely specified. Rather than providing a specific choice of g_n, we indicate the optimal order of g_n through theoretical consistency results. The availability of the analytic form of the posterior probability generated by the proposed prior setup (in (2)) makes it easily implementable.

3. Implementation in ultrahigh-dimensional settings

Our model selection procedure is simple as it chooses the model with the highest posterior probability in the model space $\mathcal{G}$, *i.e.*, the MAP model. However, identifying the MAP model is a challenging task in an ultrahigh-dimensional settings. As $p_n = \exp\{O(n)\}$, it is impossible to evaluate all the $2^{p_n} - 1$ models in $\mathcal{G}$, even for small values of n. For instance, if $n = 5$, the cardinality of $\mathcal{G}$ can be as large as $\exp(45)$. Thus, we need to develop a screening algorithm to discard a large set of *unimportant* covariates initially. Following the implementation of the screening algorithm, ideally, we will be left with a smaller set of covariates which includes all the covariates involved in the MAP model. Then, an exhaustive model search algorithm can be employed in the second stage to find the MAP model. We describe the proposed two-stage algorithm in detail below.

Proposed two-stage algorithm. The proposed two-stage algorithm is based on the *sparsity* assumption, which states that among the large number of available predictors an insignificant fraction of predictors is actually useful. Consequently, the dimension of the MAP model is small. Now, let d_n be a moderately large number, for instance $d_n \sim \log n$. The first step of the two-stage algorithm is devoted towards finding a *good* model of dimension d_n. As the number of useful predictors is small, it is expected that the d_n-dimensional optimal model chosen in first stage includes all the predictors of the MAP model. Towards finding a d_n-dimensional *good* model, a neighborhood-based search algorithm is employed on the space of all d_n-dimensional models. Below, we describe the algorithm.

Stage 1: Screening: The objective of the screening algorithm is to choose a d_n-dimensional model with high posterior probability. Given the choice of d_n, we employ the following steps to achieve this.

1. *Initialization.* Choose a model, say M_{γ_0}, of dimension d_n, where $\boldsymbol{\gamma}_0 \subseteq \{1, \ldots, p\}$ is the index set of the predictors in M_{γ_0}.
2. *Evaluation.* Fix $r \in \boldsymbol{\gamma}_0$. Define

   $$k^\star = \text{argmax}_{l \in \{1,\ldots,p_n\} \setminus \gamma_0} m_{\gamma_0 \cup \{l\} \setminus \{r\}}(\mathbf{y}_n), \quad \text{and} \quad \mathrm{u} = \mathbb{I}\left(m_{\gamma_0 \cup \{k^\star\} \setminus \{j\}}(\mathbf{y}_n) > m_{\gamma_0}(\mathbf{y}_n)\right)$$

   where $\mathbb{I}(A)$ is the indicator of the event A. If $\mathrm{u} = 1$, then replace x_r by $x_{k^\star}$ in $\boldsymbol{\gamma}_0$. If $\mathrm{u} = 0$, then keep $\boldsymbol{\gamma}_0$ unaltered.

   Repeat step 2 unless all the components in $\boldsymbol{\gamma}_0$ are evaluated.
3. *Replication.* Repeat Step 2 $N(\geq 1)$ times.

In Step 2, we replace the covariates of M_{γ_0} with the best possible inactive covariates of M_{γ_0}, provided the posterior probabilities increase by the replacements. To obtain the best result, instead of starting with any d_n dimensional model, one may choose the covariates of the initial model M_{γ_0} by a forward regression method.

Finally, we argue that with a good initial model M_{γ_0} the choice of N in Step 3 is expected to be small. We provide the following three intuitive reasons for that: (i) In the screening stage the objective is to arrive at any d_n-dimensional model which contains the MAP-covariates. This is a much easier task than searching for the MAP model. (ii) Under reasonable assumptions, the posterior probability of the class of d_n-dimensional models containing the useful covariates, say $\mathcal{G}_{1,d}$, uniformly dominates the space of all d_n-dimensional models (see Section 4). As the screening algorithm transits to a higher posterior probability model at each move, the complementary class of $\mathcal{G}_{1,d}$, having combined posterior probability close to zero, is stepped aside by the algorithm soon. (iii) Unlike other forward or marginal screening algorithms, the proposed algorithm compares the d_n-dimensional models only. Thus, in one hand the variable dimensional search problem is reduced to a fixed dimensional one, on the other hand the joint structures of the covariates are taken care of.

Stage 2: Model selection: Suppose that the first-stage *screening algorithm* selects the model $M_{\gamma^\star}$. In the next stage, we aim to find the highest-posterior probability model among the $2^{d_n} - 1$ models formed by the d_n covariates present in $M_{\gamma^\star}$. Towards that, we employ the reversible jump MCMC (RJMCMC) algorithm described in Chipman *et al.* (2001, Section 3.5), which induces a Markov chain $\mathcal{C}$ with the class of all submodels of $M_{\gamma^\star}$ as the state space, say $\mathcal{G}^\star$. The stationary distribution of $\mathcal{C}$ is the posterior probability distribution of the models restricted to $\mathcal{G}^\star$. Thus, if the covariates of the MAP model of $\mathcal{G}$ is present in $\gamma^\star$, then the MAP model lies in $\mathcal{G}^\star$, and the second stage algorithm reaches the MAP model quite easily, as the cardinality of $\mathcal{G}^\star$ is fairly small.

Remark 1: In practice, the choice of d_n can be as small as possible provided it is larger than the cardinality of the MAP model. A smaller choice of d_n results in faster execution of both the algorithms. The complexity of the first stage screening algorithm is at most of order $O(Nd_np_n)$. Even if one considers all the $2^{d_n} - 1$ competing models in $\mathcal{G}^\star$ for comparison in the second stage, the complexity of the second stage algorithm would be at most $O(nd_n^3)$, if $d_n \sim \log n$. Thus the total complexity of the two-stage algorithm is $o(p_n^r)$ for any $r > 1$.

Remark 2: As in the second stage, one could also employ an MCMC algorithm in the first stage. In each iteration, the algorithm would choose a proposal model from the *swap*-neighborhood of the current model and transit to the same according to a Metropolis-Hastings transition function based on the posterior probabilities of the proposal and current models. The algorithm would induce a Markov chain $\mathcal{C}_1$ in the state space $\mathcal{G}_d = \{M_\gamma : p_n(\gamma) = d_n\}$ that would have the posterior probability distribution restricted to $\mathcal{G}_d$ as the stationary distribution. After convergence, it would select model from the high-probability posterior region, *i.e.*, the region of supermodels. However, we avoid taking that path as the proposed screening algorithm is much faster as we will see in the numerical section.

4. Model selection consistency

We consider a frequentist validation approach to theoretically justify the performance of the proposed two-stage algorithm. Towards that, we assume existence of a unique data generating model, termed as the *true model* (M_{γ_c}), in the model space $\mathcal{G}$. Under M_{γ_c}, $\mathbf{y}_n = \boldsymbol{\mu}_n + \mathbf{e}_n = X_{\gamma_c}\boldsymbol{\beta}_{\gamma_c} + \mathbf{e}_n$, where $\boldsymbol{\mu}_n$ is the expectation of $\mathbf{y}_n$ given X_n. The dimension of M_{γ_c}, denoted by $p(\gamma_c)$, is assumed to be small and free of n. The objective of this section is to show that the two-stage algorithm selects the true model with probability tending to one.

Recall that, the first stage screening algorithm explores the class of all d_n-dimensional models $\mathcal{G}_d$, and at each move it transits to a higher posterior probability model. Thus, it is expected that after sufficient number of moves the algorithm selects a high posterior probability model in $\mathcal{G}_d$. The following subsection (Section 4.1) shows that the posterior probability of the class of all d_n-dimensional supermodels of the true model M_{γ_c}, namely, $\mathcal{G}_{1,d}$, uniformly dominates $\mathcal{G}_d$, with probability tending to one. Thus, with probability tending to one, the high posterior probability model chosen in the first stage will be a supermodel of M_{γ_c}.

In the next stage, we search within the class of all sub-models of the selected model in first stage. As d_n ($\sim \log n$) is small, the second stage RJMCMC algorithm converges to the stationary distribution in finite time. In this case, the stationary distribution is the distribution of posterior probabilities restricted to the sub-models of first stage selected model. Section 4.2 shows that, provided a supermodel of M_{γ_c} is selected at first stage, the restricted posterior distribution converges to a degenerate distribution having non-zero probability mass at M_{γ_c} only, with probability tending to one. Thus, selection of true model is guaranteed with probability tending to one.

Assumptions: Below, we list the assumptions under which our theoretical results hold.

(A1) The number of regressors $p_n = \exp\{b_0 n^r\}$ with $0 < r \leq 1$ and $b_0 > 0$ is free of n.

(A2) The true model M_{γ_c} is unique and its dimension, $p(\gamma_c)$, is free of n. Let $\boldsymbol{\mu}_n = X_{\gamma_c}\boldsymbol{\beta}_{\gamma_c}$ be the true mean of $\mathbf{y}_n$, then $\boldsymbol{\mu}_n'\boldsymbol{\mu}_n = O(n)$.

(A3) Let $\tau_{\max}$ and $\tau_{\min}$ be two positive constants, S be any subset of $\{1, \ldots, p_n\}$ of cardinality $|S| \lesssim \log n$ and X_S be the submatrix of X_n with the columns corresponding to S. Then,

$$n^{-1}\tau_{\min} \leq \inf_S \lambda_{\min}\left(n^{-1}X_S'X_S\right) \leq \sup_S \lambda_{\max}\left(n^{-1}X_S'X_S\right) \leq n\tau_{\max}.$$

(A4) Let $\Delta_0 = \{\delta n^{1-s}\} \vee \{4\sigma^2 p(\gamma_c)\log p_n\}$ for some $\delta > 0$ and $0 < s < 1/2 - \xi$ with $0 < \xi < 1/2$, $\mathcal{G}_0 = \{\gamma \in \mathcal{G} : M_{\gamma_c} \nsubseteq M_\gamma,\ p_n(\gamma) \lesssim \log n\}$ and $P_n(\gamma)$ be the projection matrix onto the span of X_γ. Then, for all sufficiently large n, we have

$$\inf_{\gamma \in \mathcal{G}_0} \boldsymbol{\mu}_n'(I - P_n(\gamma))\boldsymbol{\mu}_n > \Delta_0.$$

Assumption (A1) provides the rate of growth of p_n as a function of n, allowing exponential growth of p_n with respect to n. Assumption (A2) provides the sparsity structure of the true model. Assumption (A3) provides a restriction of the eigenstructure of small dimensional models. By (A3), all models of dimension $O(\log n)$ are of full-rank, although the bounds on the eigenvalues are quite permissive. Assumption (A4) is commonly termed as an identifiability condition for model selection. The quantity $\boldsymbol{\mu}_n'(I - P_n(\gamma))\boldsymbol{\mu}_n$ may be interpreted as the Kullback-Leibler (KL) divergence of the distribution of $\mathbf{y}_n$ under the model M_γ and M_{γ_c}. By Moreno *et al.* (2015, Lemma 3), $\lim_{n\to\infty}\{\boldsymbol{\mu}_n'(I - P_n(\gamma))\boldsymbol{\mu}_n\}/n$ is strictly positive for any non-supermodel of M_{γ_c}. (A3) additionally assumes a uniform lower bound for $\boldsymbol{\mu}_n'(I - P_n(\gamma))\boldsymbol{\mu}_n$ over non-supermodels of small dimension, and fixed a threshold value for the case with $\log p_n \sim b_0 n$. When $\log p_n = b_0 n^{1-r}$ with $r > 0$ the condition is satisfied trivially.

4.1. Consistency of the first-stage screening

Let $\mathcal{G}_d$, $\mathcal{G}_{1,d}$ and $\mathcal{G}_{2,d}$ denote the classes of d_n-dimensional models, supermodels and non-supermodels of M_{γ_c}, respectively. Define $P(M_\gamma \mid \mathcal{G}_d, \mathbf{y}_n)$ as the posterior probability distribution of the models restricted to $\mathcal{G}_d$. The following theorem shows that for any model $\gamma \in \mathcal{G}_d$, the posterior probability of $\gamma \in \mathcal{G}_{1,d}$ uniformly dominates that of $\gamma \in \mathcal{G}_{2,d}$, *i.e.*,

$$P(\gamma \in \mathcal{G}_{1,d} \mid \mathcal{G}_d, \mathbf{y}_n) \to 1, \tag{3}$$

with probability tending to one, as $p_n \to \infty$. This implies that the posterior probability distribution $P(M_\gamma \mid \mathcal{G}_d, \mathbf{y}_n)$ restricted to $\mathcal{G}_d$, assigns nearly 0 probability to $\mathcal{G}_{2,d}$.

Theorem 1: Consider the model stated in (1) with p_n satisfying (A1) and the prior setup discussed in Section 2.1. Suppose there exists a true model M_{γ_c} satisfying (A2) which generates $\mathbf{y}_n$, and let $\mathcal{G}_{1,d}$ and $\mathcal{G}_{2,d}$ be the classes of d_n-dimensional supermodels and non-supermodels of M_{γ_c}. Then, under the assumptions (A3) and (A4) and provided $g_n \gtrsim n$, the following statements hold with a probability at least $1 - \exp\{-c_1 n^{\xi}\}$, where ξ is as in assumption (A4) and $c_1 > 0$ is some constant free on n.

A. For some constant $c_2 > 0$ and any $\epsilon > 0$,

$$\sup_{\gamma_1 \in \mathcal{G}_{1,d}, \gamma_2 \in \mathcal{G}_{2,d}} \frac{P(M_{\gamma_2}|\mathcal{G}_d, \mathbf{y}_n)}{P(M_{\gamma_1}|\mathcal{G}_d, \mathbf{y}_n)} \le c_2 n^{d_n} \exp\{-\Delta_0(1-\epsilon)/(2\sigma^2)\}.$$

B. For some constant $c_3 > 0$ and any $\epsilon > 0$,

$$\frac{\sum_{\gamma_2 \in \mathcal{G}_{2,d}} P(M_{\gamma_2}|\mathcal{G}_d, \mathbf{y}_n)}{\sum_{\gamma_1 \in \mathcal{G}_{1,d}} P(M_{\gamma_1}|\mathcal{G}_d, \mathbf{y}_n)} \le c_3 n^{d_n} p_n^{-(1-2\epsilon)p(\gamma_c)}.$$

C. For any $\gamma \in \mathcal{G}_d$, $P(\gamma \in \mathcal{G}_{1,d} \mid \mathcal{G}_d, \mathbf{y}_n) \to 1$ with probability tending to one, as $n \to \infty$.

In stage 1, the screening algorithm searches for a high-posterior probability model in the restricted model space $\mathcal{G}_d$. By part A of Theorem 1, the posterior probability of the class of models in $\mathcal{G}_{1,d}$ uniformly dominates that of $\mathcal{G}_{2,d}$. Thus, the proposed sequence of $O(Nd_np_n)$ moves in the first-stage algorithm, wherein each move selects a higher posterior probability model, is expected to reach a model in $\mathcal{G}_{1,d}$.

4.2. Consistency of the second-stage selection

As argued in the previous sub-section, the model $M_{\gamma^\star}$ selected in the first stage screening is expected to be a d_n-dimensional supermodel of M_{γ_c}. In the second stage, the RJMCMC algorithm employed explores the class of the all submodels of $M_{\gamma^\star}$, say $\mathcal{G}^\star$. After a sufficient number of iterations, the algorithm selects models as per the posterior distribution restricted to $\mathcal{G}^\star$. The next theorem shows that, if $M_{\gamma^\star}$ is any supermodel of M_{γ_c}, then the posterior distribution restricted to $\mathcal{G}^\star$ limits to a degenerate distribution having non zero probability mass at M_{γ_c}, with probability tending to one. Therefore, provided $M_{\gamma^\star}$ is any supermodel of M_{γ_c}, the second stage algorithm selects M_{γ_c} with probability tending to one.

Theorem 2: Consider the model stated in (1) with p_n satisfying (A1), and the prior setup discussed in Section 2.1 with $g_n \sim p_n^{\delta}$ with some $0 < \delta < 2$. Suppose there exists a true model M_{γ_c} satisfying (A2), which generates $\mathbf{y}_n$. Let $M_{\gamma^\star}$ be a d_n-dimensional supermodel of M_{γ_c}, $\mathcal{G}^\star = \{M_\gamma : \gamma \subseteq \gamma^\star\}$ be the class of all sub-models of $M_{\gamma^\star}$, and $P\left(M_\gamma \mid \mathcal{G}^\star, \mathbf{y}_n\right)$ be the posterior probability of models restricted to $\mathcal{G}^\star$. Then, under assumptions (A3)-(A4), with a probability at least $1 - cp_n^{-c_0} - \exp\{-cn^{\xi}\}$, where $c_0 < \delta/2$ and $c > 0$ are two constants, $\xi > 0$ is as in (A4), and $\delta > 0$ is as stated in the choice of g_n, we have

$$\inf_{\gamma^\star \in \mathcal{G}_{1,d}} P\left(M_{\gamma_c} \mid \mathcal{G}^\star, \mathbf{y}_n\right) \geq \left[1 + cn^{p(\gamma_c)+1}\left(\frac{p_n^{2\epsilon}}{g_n}\right)^{1/2} + c\left(\frac{n\sqrt{g_n}}{p_n^{1-2\epsilon}}\right)^{p(\gamma_c)}\right]^{-1}$$

for any $\epsilon < \delta/2$. Consequently, $\inf_{\gamma^\star \in \mathcal{G}_{1,d}} P\left(M_{\gamma_c} \mid \mathbf{y}_n, \mathcal{G}^\star\right) \to 1$, with probability tending to 1.

Theorem 2 states that, provided the first stage algorithm selects any supermodel of M_{γ_c}, the second stage algorithm selects the true model with probability tending to one.

4.3. Consistency of the two-stage procedure

Finally, we argue that the two stage procedure selects the true model with probability tending to one. Towards that, define $P_2(\cdot|\mathbf{y}_n)$ as the probability distribution of the models after the second stage. Let $M_{\gamma^\star}$ be the d_n-dimensional model selected in the first stage, and $\mathcal{G}^\star = \{M_\gamma : \gamma \subseteq \gamma^\star\}$ be the class of all sub-models of $M_{\gamma^\star}$. Then,

$$P_2(M_\gamma|\mathbf{y}_n) = \textstyle\sum_{\gamma^\star \in \mathcal{G}_d} P(M_\gamma|\mathcal{G}^\star, \mathbf{y}_n)P(M_{\gamma^\star}|\mathcal{G}_d, \mathbf{y}_n),$$

if in the first stage a model is selected randomly as per the posterior distribution restricted to $\mathcal{G}_d$, and in the second stage a model is selected randomly as per the posterior distribution restricted to $\mathcal{G}^\star$. The next theorem shows, with probability tending to one, $P_2(M_{\gamma_c}|\mathbf{y}_n) \to 1$.

Theorem 3: Consider the model stated in (1) with p_n satisfying (A1), and the prior setup discussed in Section 2.1 with $g_n \sim p_n^{\delta}$ with some $0 < \delta < 2$. Suppose there exists a true model M_{γ_c} satisfying (A2), which generates $\mathbf{y}_n$. Further, suppose that a two stage procedure is employed to identify the true model, wherein the first stage selects a d_n-dimensional model $M_{\gamma^\star}$ randomly as per the posterior distribution (2) restricted to $\mathcal{G}_d$ (class of d-dimensional models), and the second stage selects a model randomly from the posterior distribution (2) restricted to the sub-models of $M_{\gamma^\star}$ (*i.e.*, $\mathcal{G}^\star$). Let $P_2(\cdot \mid \mathbf{y}_n)$ be the probability distribution of the models selected at the end of the two-stage procedure then under assumptions (A3)-(A4), $P_2(M_{\gamma_c}|\mathbf{y}_n) \to 1$ as $n \to \infty$, with probability tending to one.

Remark 3: The choice of the only hyperparameter g_n in the prior setup is not specified. However, from the above theoretical developments, we obtain an optimal range of g_n value required for consistency of the two-stage procedure. Theorem 1 holds for any g_n satisfying $g_n \gtrsim n$, while Theorems 2 and 3 requires $g_n \sim p_n^{\delta}$ with $0 < \delta < 2$. These provide a vast range of plausible choices of g_n. For practical purposes some sensitivity analysis would be useful.

5. Simulation study

We now study the performance of the proposed two-stage variable selection procedure using a wide variety of simulated data sets. Under different simulation schemes, we present the proportion of times a variable selection algorithm selects the true model.

Our method: Our model selection algorithm is completely described in Section 3, except for the choices of g_n, d_n and N. The choice of d_n is taken to be $[n/4]$ in each case. In the first stage, we choose $g_n = np_n$ and in the second stage, we choose $g_n = d_n^2$. Note that, the theoretical condition on g_n in Theorems 2 and 3 come from the consideration of the two-stages together. However, practically, the task of the second stage is find the MAP model among the $2^{d_n} - 1$ models formed by d_n covariates. Therefore, informed by Fernández *et al.* (2001), the benchmark prior $g_n = \max\{n, d_n^2\}$ is considered in the second stage. Finally, in the first-stage $N = 10$ iterations are considered, and in the second-stage, the RJMCMC algorithm is iterated 6000 times, with a burning period of 3000 iterations. The post-burning most visited model is considered as the optimal model.

Other methods: Among the frequentist variable selection methods, we consider three approaches based on iterative sure independence screening (ISIS). An initial set of variables is first selected by ISIS, and then a penalized regression step 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 using the BIC. These three methods are termed as ISIS-LASSO-BIC, ISIS-SCAD-BIC and ISIS-MCP-BIC. Among the Bayesian competitors, we consider two methods based on Bayesian credible region (BCR joint and BCR marginal, 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 are obtained from the first author's website, while the first author of Narisetty and He (2014) kindly shared the codes for BASAD with us. Further, we have implemented the approximate version of BASAD to reduce the 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 $\boldsymbol{\mu}_n = X_{\gamma_c}\boldsymbol{\beta}_{\gamma_c}$. The vector $\boldsymbol{\beta}_{\gamma_c}$ is assumed to be sparse, *i.e.*, $p(\gamma_c) \ll p_n$, and these $p(\gamma_c)$ components are chosen randomly from the set of all covariates. When $p_n \leq 500$, we set $p(\gamma_c) = 5$, while $p(\gamma_c) = 10$ is set for higher values of p_n. All the $p(\gamma_c)$ values of $\boldsymbol{\beta}_{\gamma_c}$ are taken to be equal to 2.

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

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

Case 2. (*Block dependence*) $\boldsymbol{\Sigma}_{p_n}$ 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*) $\boldsymbol{\Sigma}_{p_n} = 0.5\mathbf{I} + 0.5\mathbf{1}\mathbf{1}'$, where $\mathbf{1}$ is the p_n-dimensional vector of ones. This exhibits a strong dependence structure uniformly among the covariates.

Case 4. (*Auto-regressive*) Here, we take $\sigma_{ii} = 1$ for $1 \leq i \leq p_n$, and $\sigma_{ij} = 0.9^{|i-j|}$ for

$1 \leq i \neq j \leq p_n$. With the increase in distance, the correlation decreases here.

Although theoretically we consider only Gaussian errors, in simulation studies, we consider two errors distributions, namely, the Gaussian and the heavy-tailed t distribution with 2 degrees of freedom. In the tables below, we report the proportion of times each method selects the true model in 100 random iterations. Additionally, we report the proportion of times our first-stage screening algorithm chooses a supermodel of the true model.

Simulation results. Tables 1 and 2 contain the results corresponding to $n = 50$ and $n = 100$, respectively. We notice that the covariance structure in Case 2 becomes singular for $p_n \geq 1000$, and therefore, we have restricted Case 2 to $p_n \leq 500$.

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	
↓ p_n →	100	500	100	500	100	500	100	500
ISIS-SCAD-BIC	0.65	0.42	0.05	0.00	0.46	0.19	0.66	0.38
ISIS-MCP-BIC	0.44	0.23	0.02	0.00	0.12	0.04	0.50	0.24
BCR	0.26	0.00	0.45	0.00	0.15	0.00	0.22	0.00
BASAD	0.93	0.50	**0.82**	0.07	0.82	0.55	0.92	0.49
Proposed	**0.99**	**0.84**	0.72	**0.09**	**0.96**	**0.80**	**1.00**	**0.87**
Proposed (Step 1)	1.00	0.85	0.77	0.09	0.96	0.81	1.00	0.87
t_2 error								
	Case 1		Case 2		Case 3		Case 4	
Methods ↓ p_n →	100	500	100	500	100	500	100	500
ISIS-SCAD-BIC	0.33	0.34	0.02	0.00	0.28	0.21	0.33	0.29
ISIS-MCP-BIC	0.26	0.26	0.02	0.00	0.20	0.17	0.27	0.26
BCR	0.15	0.01	0.29	0.00	0.12	0.00	0.20	0.00
BASAD	**0.69**	0.30	**0.55**	**0.09**	0.61	0.38	0.69	0.37
Proposed	**0.69**	**0.60**	0.54	0.08	**0.66**	**0.53**	**0.72**	**0.59**
Proposed (Step 1)	0.83	0.67	0.65	0.08	0.77	0.56	0.84	0.65

Among the three frequentist methods based on ISIS, we have reported the results for SCAD and MCP only, as ISIS-LASSO-BIC is outperformed by these two methods. For the other two methods, SCAD has shown uniformly better performance than MCP (see Table 1). For BCR, we observe that the joint version leads to singularity in several iterations in the simulation settings. Therefore, we have reported results for the more stable marginal version only. It is also clear from Tables 1 and 2 that ISIS is affected drastically when the dependence structure varies among the different sets of covariates. For example, for $n = 100$, ISIS-SCAD-BIC leads to the best performance under independence (Case 1) when $p_n = 2000$. However, it fails to identify the true model in a single instance under block-diagonal covariance structure (Case 2). This is due to the fact that ISIS relies on marginal information, and ignores the joint structure of the covariates.

Generally, the Bayesian methods turn out to be *more robust* than frequentist approaches. Among the Bayesian methods, BASAD and the proposed method clearly outperform BCR for all the cases. However, the performance of BASAD falls drastically for higher values of p_n. For example, when $p_n = 2000$, BASAD fails completely, irrespective of the underlying covariance structure. Note that BASAD needs to compute the inverse of

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		
$p_n \to$	500	1000	2000	500	500	1000	2000	500	1000	2000
ISIS-SCAD-BIC	0.85	0.39	0.28	0.00	0.64	0.18	0.02	0.84	0.43	0.25
ISIS-MCP-BIC	0.66	0.25	0.16	0.00	0.11	0.01	0.00	0.62	0.24	0.11
BCR	0.38	0.00	0.00	0.39	0.14	0.00	0.00	0.24	0.00	0.00
BASAD	0.93	0.19	0.00	0.92	0.93	0.36	0.00	0.98	0.27	0.00
Proposed	**0.98**	**0.95**	**0.66**	**0.97**	**1.00**	**0.92**	**0.31**	**1.00**	**0.92**	**0.27**
Proposed (Step 1)	1.00	0.96	0.66	0.97	1.00	0.92	0.31	1.00	0.93	0.57
t_2 error										
Methods ↓	Case 1			Case 2	Case 3			Case 4		
$p_n \to$	500	1000	2000	500	500	1000	2000	500	1000	2000
ISIS-SCAD-BIC	0.44	0.41	**0.32**	0.00	0.39	0.29	0.30	0.45	0.36	0.30
ISIS-MCP-BIC	0.38	0.39	0.29	0.00	0.24	0.23	0.28	0.40	0.33	0.27
BCR	0.26	0.00	0.00	0.23	0.09	0.00	0.00	0.21	0.00	0.00
BASAD	0.91	0.06	0.00	0.75	0.78	0.19	0.00	**0.88**	0.12	0.00
Proposed	**0.93**	**0.70**	0.21	**0.84**	**0.85**	**0.60**	**0.39**	0.78	**0.70**	**0.39**
Proposed (Step 1)	0.96	0.70	0.48	0.87	0.95	0.60	0.39	0.78	0.71	0.40

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 might be one of the reasons behind its poor performance. The strength of our proposed method is re-iterated from the simulation study, especially for higher values of p_n. Notably, there is a systematic improvement of the proposed method over BASAD when we move from $p_n = 100$ to $p_n \geq 500$, especially under cases 1, 3 and 4, for both the error distributions.

The performance of the first-stage screening algorithm is noteworthy. Except for the high-dimension-low-sample size situation with high correlation, *i.e.*, for $n = 50$, $p_n = 500$ in Case 2, this algorithm selects the true model for a high-proportion of times in all other cases.

To check the sensitivity of our method to the value of $\boldsymbol{\beta}_{\gamma_c}$, we perform a further simulation study. We consider Case 1 ($\boldsymbol{\Sigma}_{p_n} = \mathbf{I}$) with the Gaussian error distribution for $n = 100$; and two choices of $\boldsymbol{\beta}_{\gamma_c}$. First, a set of equi-spaced values of $\boldsymbol{\beta}_{\gamma_c}$ in the range $[1, 2]$ and next in the range $[2, 3]$. An increment of 0.2 is taken for $p_n = 500$ so that we have $p(\gamma_c) = 6$, and an increment of 0.1 is taken for $p_n = 1000$ and 2000 so that $p(\gamma_c) = 11$. The results are summarized in Table 3 below.

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

Methods ↓	$\boldsymbol{\beta}_{\gamma_c} = (1.0, 1.2, \ldots, 2)'$			$\boldsymbol{\beta}_{\gamma_c} = (2.0, 2.1, \ldots, 3)'$		
$p_n \to$	500	1000	2000	500	1000	2000
ISIS-SCAD-BIC	0.66	0.40	0.24	0.82	0.47	0.33
ISIS-MCP-BIC	0.63	0.26	0.00	0.68	0.27	0.19
BCR	0.14	0.00	0.00	0.24	0.00	0.00
BASAD	**0.99**	0.14	0.00	0.98	0.28	0.00
Proposed	**1.00**	**0.93**	**0.87**	**1.00**	**0.94**	**0.76**
Proposed (Step 1)	1.00	0.93	0.87	1.00	0.94	0.76

Good performance of the proposed method is further re-iterated from the numerical results of Table 3. Also, it is observed that the method is not much sensitive to the level of signal strength, as long as the minimal signal strength is not negligible.

6. Real data analysis

6.1. Metabolic quantitative trait loci experiment

The first example is related to a metabolic quantitative trait loci experiment which links single nucleotide polymorphisms (SNPs) data to metabolomics data. The *predictors* come from a GWAS study of the candidate genes for alanine amino-transferase enzyme elevation in the 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 has the same genotype as the SNP rs7896824, while the second SNP shares the same genotype with eleven other SNPs. We implement our proposed method by starting with $d_n = 5$ till $d_n = 50$ (which is the maximum possible value of d_n). From our analysis, the proposed method identifies all the SNPs (two from the first group, and all the twelve from the second group) from $d_n = 25$ onwards. We further observe that the proposed method consistently identifies a new set of SNPs consists of rs6704330 and rs12744386. This is a novel set of SNPs which *were not detected* in the earlier study, and further investigation may establish their association with the metabolite under study.

For the sake of comparison, we implement all the competing methods from our simulations in Section 5. We first fix a value of the model size (d_n), and then a model selection method is used to obtain a d_n-dimensional 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 (LOOCV). 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_n, 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.

6.2. Polymerase chain reaction

This data is related to a polymerase chain reaction. A total of $n = 60$ samples, with 31 female and 29 male mice, are used to monitor the expression levels of $p_n = 22575$ genes. Some physiological phenotypes, including numbers of phosphoenolpyruvate carboxykinase, glycerol-3-phosphate acyltransferase, and stearoyl-CoA desaturase 1 are measured by quantitative real-time polymerase chain reaction. The relationship between the gene expression level (*perdictor*) and phosphoenolpyruvate carboxykinase (*response*) is of interest in this data. The gene expression data is standardized before the statistical analysis. To analyze this data, we repeat the same procedure as in Section 6.1 above.

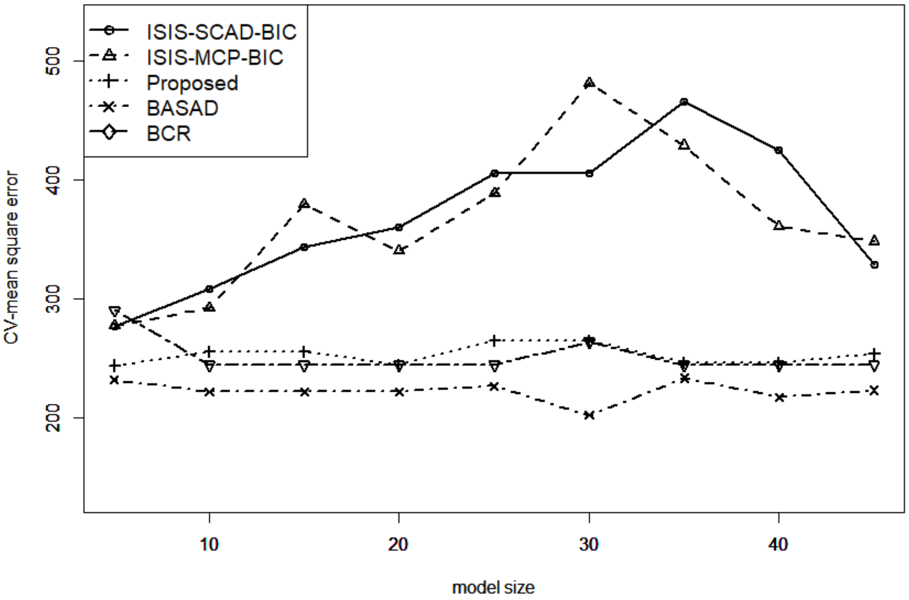

Figure 1: Comparison of the different methods using median square errors

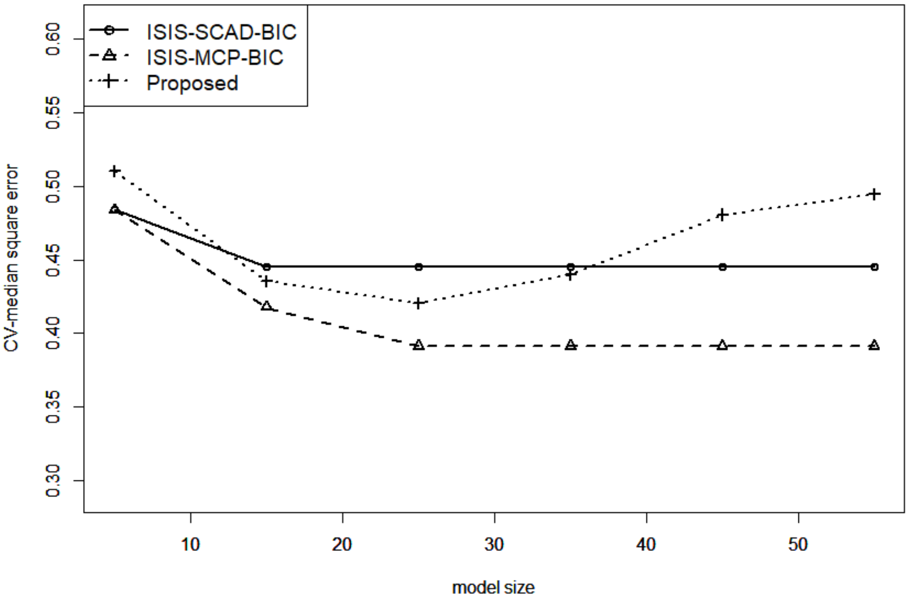

Figure 2: Comparison of the different methods using median square errors

Both BASAD and BCR could not be implemented for this data due to memory overflow for this data. Figure 2 gives us the overall picture of the performance of the other methods, and they all yield quite low median square errors. Clearly, ISIS-MCP leads to the lowest overall errors, and the proposed method performs marginally better than ISIS-SCAD for $d_n = 15$ to 35. However, the maximum difference in errors of the proposed method with both methods based on ISIS is less than 0.11 over all values of d_n.

7. Concluding remarks

This paper addresses the variable selection problem in ultrahigh-dimensional linear regression settings. A new methodology for variable selection based on Zellner's g-prior is developed, taking into account the key features of the ultrahigh-dimensional regression settings, such as sparsity and multicollinearity, and adapting it accordingly. Variable selection in ultrahigh dimensions poses significant challenges due to the exponential growth of the model space with the number of covariates. Despite its various advantages, the predominant Bayesian variable selection procedure, the maximum a-posteriori (MAP) approach, becomes impractical in this context due to the vast model space. To address this problem, we propose a two-stepped model selection procedure that incorporates an initial screening.

While the idea of screening out unimportant covariates in the initial stage is not new, existing screening algorithms typically rely on marginal utilities and overlook the joint structure of the covariates. Our proposed screening algorithm takes the joint structure of the covariates into account, demonstrating greater efficiency and robustness across various correlation structures, as evidenced by our numerical results. In the second stage, we conduct a thorough model search within the class of submodels of the first-stage-selected model. Notably, we establish the strong selection consistency property of our two-stage algorithm theoretically under exponential growth of p_n with n. To our knowledge, this is the first selection consistency result addressing the exponential growth of p_n with n.

We conclude this section with some future directions. The effectiveness of our proposed two-stage procedure is heavily dependent on the sparsity assumption of the optimal model. While sparsity is commonly observed in high-dimensional regression, it is essential to expedite the search for the MAP model in denser cases as well. Relevantly, the choice of d_n is a critical factor in our method. A smaller d_n can enhance the speed and efficiency of both algorithms but may also lead to exclusion of important covariates. Thus, it is necessary to develop a mechanism for determining the optimal choice of d_n based on the data at hand.

Acknowledgements

We are thankful to the reviewer and the Editor for their constructive comments and helpful suggestions. The first author has been partially supported by the grant MTR/2023/001170, while the second author has been partially supported by the grant MTR/2023/000884.

Reproducibility

Codes for the proposed two-stage algorithm are available in the following link: https://github.com/mmukhopadhyay/Bayesian-Variable-Selection-for-Ultrahigh-dimensional-Sparse-Linear-Models.git.

References

Bondell, H. D. and Reich, B. J. (2012). Consistent high-dimensional Bayesian variable selection via penalized credible regions. *Journal of the American Statistical Association*, **107**, 1610–1624.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*. Springer Series in Statistics. Springer, Heidelberg. Methods, theory and applications.

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

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-Monograph Series*, pages 65–134. Institute of Mathematical Statistics, Beachwood, OH.

Fan, J., Feng, Y., and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, **106**, 544–557.

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

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics*, **38**, 3567–3604.

Fernández, C., Ley, E., and Steel, M. F. J. (2001). Benchmark priors for Bayesian model averaging. *Journal of Econometrics*, **100**, 381–427.

George, E. I. and Foster, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika*, **87**, 731–747.

George, E. I. and Mcculloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, **88**, 881–889.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. *The Annals of Statistics*, **33**, 730–773.

Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, **28**, 1302–1338.

Li, D., Dutta, S., and Roy, V. (2023). Model based screening embedded Bayesian variable selection for ultra-high dimensional settings. *Journal of Computational and Graphical Statistics*, **32**, 61–73.

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

Moreno, E., Girón, J., and Casella, G. (2015). Posterior model consistency in variable selection as the model dimension grows. *Statistical Science. A Review Journal of the Institute of Mathematical Statistics*, **30**, 228–241.

Narisetty, N. N. and He, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, **42**, 789–817.

Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, **103**, 681–686.

Shin, M., Bhattacharya, A., and Johnson, V. E. (2018). Scalable Bayesian variable selection using nonlocal prior densities in ultrahigh-dimensional settings. *Statistica Sinica*, **28**, 1053–1078.

Song, Q. and Liang, F. (2015). A split-and-merge Bayesian variable selection approach for ultrahigh dimensional regression. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, **77**, 947–972.

Song, R., Yi, F., and Zou, H. (2014). On varying-coefficient independence screening for high-dimensional varying-coefficient models. *Statistica Sinica*, **24**, 1735–1752.

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

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, **104**, 1512–1524.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, **68**, 49–67.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti (P. K. Goel and A. Zellner, eds.)*, pages 233–243.

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

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

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

A. ANNEXURE

This section contains the proof of all the theorems. In all the proofs, the notation c is used as a generic symbol for constants. In many situations, the existence of a constant, rather than the value, is important. In such cases, the constant is denoted by c. Thus, all constants denoted by c are not necessarily the same.

A.1. Auxiliary results

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

Lemma 1: Let M_γ, $\gamma \in \mathcal{G}$ and $M_{\gamma'}$ be two models with dimensions $p_n(\gamma)$ and $p_n(\gamma')$, where $p_n(\gamma), p_n(\gamma') \lesssim \log n$. Further, suppose $\tau_{\min} \le \lambda_{\min}(A) \le \lambda_{\max}(A) \le n^2\tau_{\max}$ for some $\tau_{\min} > 0$ and $\tau_{\max} > 0$ (free of n) for both matrices $A = X_\gamma' X_\gamma$ and $A = X_{\gamma'}' X_{\gamma'}$. Then,

$$\frac{\left|I + g_n X_\gamma' X_\gamma\right|^{-1}}{\left|I + g_n X_{\gamma'}' X_{\gamma'}\right|^{-1}} = \frac{\left|I + g_n X_{\gamma'}' X_{\gamma'}\right|}{\left|I + g_n X_\gamma' X_\gamma\right|} \le (1+\epsilon)\left(\frac{\tau_{\max}}{\tau_{\min} \wedge 1}\right)^{p_n(\gamma)\vee p_n(\gamma')} n^{2p_n(\gamma')} g_n^{p_n(\gamma')-p_n(\gamma)},$$

for any $\epsilon > 0$, when $g_n \gtrsim n$.

Proof: The j-th largest eigenvalue of any square matrix of the form $I + A$ are $1 + \lambda_j(A)$, where $\lambda_j(A)$ is the j-th largest eigenvalue of A. Further, both $X_\gamma' X_\gamma$ and $X_{\gamma'}' X_{\gamma'}$ are non-negative definite. Therefore, the highest eigenvalue of $I + g_n X_{\gamma'}' X_{\gamma'}$ is $1 + g_n n^2 \tau_{\max}$ and the lowest eigenvalue of $I + g_n X_\gamma' X_\gamma$ is $1 + \tau_{\min}$. By the trivial bound $\lambda_{\min}^d(A) \le |A| \le \lambda_{\max}^d(A)$, where d is the dimension of A, we get

$$\begin{aligned}
\frac{\left|I + g_n X_{\gamma'}' X_{\gamma'}\right|}{\left|I + g_n X_\gamma' X_\gamma\right|} &\le \frac{(1 + g_n n^2 \tau_{\max})^{p_n(\gamma')}}{(1 + g_n \tau_{\min})^{p_n(\gamma)}} \\
&= n^{2p_n(\gamma')} g_n^{p_n(\gamma')-p_n(\gamma)} \tau_{\max}^{p_n(\gamma')} \tau_{\min}^{-p_n(\gamma)} \frac{\{1 + 1/(g_n n^2 \tau_{\max})\}^{p_n(\gamma')}}{\{1 + 1/(g_n \tau_{\min})\}^{p_n(\gamma)}} \\
&\le (1+\epsilon)\left(\frac{\tau_{\max}}{\tau_{\min} \wedge 1}\right)^{p_n(\gamma)\vee p_n(\gamma')} n^{2p_n(\gamma')} g_n^{p_n(\gamma')-p_n(\gamma)},
\end{aligned}$$

for any $\epsilon > 0$ whenever $g_n \gtrsim n$. The last inequality is due to the fact that both terms $(1 + g_n n^2 \tau_{\max})^{p_n(\gamma')}$ and $(1 + g_n \tau_{\min})^{p_n(\gamma)}$ converges to one as $n \to \infty$ if $g_n \gtrsim n$. □

Lemma 2: Let M_γ be a full-rank model, $R_\gamma^{2\star} = \mathbf{y}_n' \left\{ I_n - X_\gamma \left(I_{p_n(\gamma)}/g_n + X_\gamma' X_\gamma \right)^{-1} X_\gamma' \right\} \mathbf{y}_n$, and $R_\gamma^2 = \mathbf{y}_n' \left\{ I_n - P_n(\gamma) \right\} \mathbf{y}_n$, where $P_n(\gamma) = X_\gamma (X_\gamma' X_\gamma)^{-1} X_\gamma'$ is the projection matrix on the column space of X_γ. Then, under the assumptions (A2)-(A3), the following statements hold.

(a) $R_\gamma^{2\star} \ge R_\gamma^2$, and for any model M_γ satisfying (A3), $\sup_\gamma R_\gamma^{2\star} - R_\gamma^2 \le cn/(1 + g_n \tau_{\min})$ for some appropriate constant $c > 0$ with probability at least $1 - \exp\{-n\}$,

(b) For any $\epsilon > 0$, there exists an appropriate constant $c > 0$ such that $R_{\gamma_c}^2 > n(1+\epsilon)\sigma^2$, and $R_{\gamma_c}^2 < n(1-\epsilon)\sigma^2$, with probability at least $1 - \exp\{-cn\}$.

Proof: Part(a). Observe that $I_{p_n(\gamma)}/g_n + X_\gamma' X_\gamma \geq X_\gamma' X_\gamma$, and so, $I_n - X_\gamma \left(I_{p_n(\gamma)}/g_n + X_\gamma' X_\gamma\right)^{-1} X_\gamma' \geq I_n - P_n(\gamma)$, which proves $R_\gamma^2 \leq R_\gamma^{2\star}$.

To see the other side, observe that under (A3), and uniformly over any model M_γ

$$\begin{aligned} X_\gamma \left(I_{p_n(\gamma)}/g_n + X_\gamma' X_\gamma\right)^{-1} X_\gamma' &= X_\gamma \left(X_\gamma' X_\gamma\right)^{-1/2} \left[I_{p_n(\gamma)} + \left(X_\gamma' X_\gamma\right)^{-1}/g_n\right]^{-1} \left(X_\gamma' X_\gamma\right)^{-1/2} X_\gamma' \\ &\geq \left\{1 + 1/(g_n \tau_{\min})\right\}^{-1} P_n(\gamma) \end{aligned}$$

as $\lambda_{\max}\left(I_{p_n(\gamma)} + \left(X_\gamma' X_\gamma\right)^{-1}/g_n\right) \leq 1 + 1/(g_n \tau_{\min})$. Therefore,

$$\begin{aligned} \sup_{\gamma: p_n(\gamma) \lesssim \log n} R_\gamma^{2\star} - R_\gamma^2 &\leq \sup_{\gamma: p_n(\gamma) \lesssim \log n} \mathbf{y}_n' \left[I_n - \left\{1 + 1/(g_n \tau_{\min})\right\}^{-1} P_n(\gamma) - I_n + P_n(\gamma)\right] \mathbf{y}_n \\ &= \sup_{\gamma: p_n(\gamma) \lesssim \log n} \frac{1}{1 + g_n \tau_{\min}} \mathbf{y}_n' P_n(\gamma) \mathbf{y}_n \leq \frac{1}{1 + g_n \tau_{\min}} \mathbf{y}_n' \mathbf{y}_n. \end{aligned}$$

Now, $\mathbf{y}_n' \mathbf{y}_n \leq 2\|\boldsymbol{\mu}_n\|^2 + 2\|\mathbf{e}_n\|^2$. By assumption (A2), $\|\boldsymbol{\mu}_n\|^2 = O(n)$ and as $\|\mathbf{e}_n\|^2 \sim \sigma^2 \chi_n^2$, from Laurent and Massart (2000), we have $\|\mathbf{e}_n\|^2 \leq 6n\sigma^2$ with probability at least $1 - \exp\{-n\}$. Therefore, with probability at least $1 - \exp\{-n\}$, $R_\gamma^{2\star} - R_\gamma^2 \leq cn/(1 + g_n \tau_{\min})$ for some appropriate constant $c > 0$.

Part(b). The random variable $\mathbf{e}_n'(I - P_n(\gamma_c))\mathbf{e}_n/\sigma^2$ follows a χ^2 distribution with $(n - p(\gamma_c))$ degrees of freedom. By (Laurent and Massart, 2000, Lemma 1), we have

$$\begin{aligned} P(R_{\gamma_c}^2 > n(1+\epsilon)\sigma^2) &= P(\mathbf{y}_n'(I - P_n(\gamma_c))\mathbf{y}_n > n(1+\epsilon)\sigma^2) \\ &= P\left(\mathbf{e}_n'(I - P_n(\gamma_c))\mathbf{e}_n > n(1+\epsilon)\sigma^2\right) \\ &\leq \exp\left\{-c\frac{(n\epsilon + p(\gamma_c))^2}{(n - p(\gamma_c))}\right\} \leq \exp\{-cn\}, \end{aligned}$$

for some $c > 0$. Thus, the first part of the result follows. The proof of the second part follows similarly from (Laurent and Massart, 2000, Lemma 1). □

Lemma 3: 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 $0.5 < k < 1$ and $\epsilon > 0$, there exists a constant $c > 0$ such that $n^{-k}|\boldsymbol{\mu}_n' \mathbf{e}_n| < \epsilon$ with probability at least $1 - \exp\{-cn^{2k-1}\}$

Proof: The random variable $\boldsymbol{\mu}_n' \mathbf{e}_n$ is distributed as a centered normal distribution with variance $\sigma^2 \|\boldsymbol{\mu}_n\|^2$. Therefore, we get

$$P\left(|\boldsymbol{\mu}_n' \mathbf{e}_n| \geq \epsilon n^k\right) \leq \exp\{-cn^{2k}/\|\boldsymbol{\mu}_n\|^2\} \tag{4}$$

for an appropriate constant $c > 0$ depending on ϵ. By assumption (A2), $\|\boldsymbol{\mu}_n\|^2 = O(n)$. Therefore, the quantity on the right-hand side of the above expression is bounded above by $\exp\{-cn^{2k-1}\}$ for some $c > 0$. Thus, the result follows. □

A.2. Main results

A.2.1. Proof of Theorem 1

Proof: **[Part A.]** By (2), the ratio of posterior probabilities is

$$\begin{aligned}\sup_{\gamma_1\in\mathcal{G}_{1,d},\gamma_2\in\mathcal{G}_{2,d}}\frac{P(M_{\gamma_2}|\mathcal{G}_d,\mathbf{y}_n)}{P(M_{\gamma_1}|\mathcal{G}_d,\mathbf{y}_n)} &= \sup_{\gamma_1\in\mathcal{G}_{1,d},\gamma_2\in\mathcal{G}_{2,d}}\frac{P(M_{\gamma_2}|\mathbf{y}_n)}{P(M_{\gamma_1}|\mathbf{y}_n)}\\ &= \sup_{\gamma_1\in\mathcal{G}_{1,d},\gamma_2\in\mathcal{G}_{2,d}}\frac{\left|I+g_nX'_{\gamma_1}X_{\gamma_1}\right|^{1/2}}{\left|I+g_nX'_{\gamma_2}X_{\gamma_2}\right|^{1/2}}\left(\frac{R^{\star 2}_{\gamma_1}}{R^{\star 2}_{\gamma_2}}\right)^{n/2}. \qquad (5)\end{aligned}$$

By assumption (A3) and Lemma 1

$$\sup_{\gamma_1\in\mathcal{G}_{1,d},\gamma_2\in\mathcal{G}_{2,d}}\frac{\left|I+g_nX'_{\gamma_1}X_{\gamma_1}\right|^{1/2}}{\left|I+g_nX'_{\gamma_2}X_{\gamma_2}\right|^{1/2}}\le 2\left(\frac{\tau_{\max}}{\tau_{\min}\wedge 1}\right)^{d_n/2}n^{d_n}. \qquad (6)$$

Next, we write the last part in the RHS of (5) as follows:

$$\begin{aligned}\sup_{\gamma_1\in\mathcal{G}_{1,d},\gamma_2\in\mathcal{G}_{2,d}}\left(\frac{R^{\star 2}_{\gamma_1}}{R^{\star 2}_{\gamma_2}}\right)^{n/2} &\le \sup_{\gamma_1\in\mathcal{G}_{1,d}}\left(\frac{R^{\star 2}_{\gamma_1}}{R^{2}_{\gamma_1}}\right)^{n/2}\sup_{\gamma_1\in\mathcal{G}_{1,d}}\left(\frac{R^{2}_{\gamma_1}}{R^{2}_{\gamma_c}}\right)^{n/2}\\ &\qquad\sup_{\gamma_2\in\mathcal{G}_{2,d}}\left(\frac{R^{2}_{\gamma_c}}{R^{2}_{\gamma_2}}\right)^{n/2}\sup_{\gamma_2\in\mathcal{G}_{2,d}}\left(\frac{R^{2}_{\gamma_2}}{R^{\star 2}_{\gamma_2}}\right)^{n/2}. \qquad (7)\end{aligned}$$

We consider each term of the RHS of the above expression consecutively. By Lemma 2

$$\begin{aligned}\sup_{\gamma_1\in\mathcal{G}_{1,d}}\left(\frac{R^{\star 2}_{\gamma_1}}{R^{2}_{\gamma_1}}\right)^{n/2} &= \sup_{\gamma_1\in\mathcal{G}_{1,d}}\left(1+\frac{R^{\star 2}_{\gamma_1}-R^{2}_{\gamma_1}}{R^{2}_{\gamma_1}}\right)^{n/2}\le\sup_{\gamma_1\in\mathcal{G}_{1,d}}\left(1+\frac{R^{\star 2}_{\gamma_1}-R^{2}_{\gamma_1}}{R^{2}_{\gamma_c}}\right)^{n/2}\\ &\le \left(1+\frac{c}{1+g_n\tau_{\min}}\right)^{n/2}, \qquad (8)\end{aligned}$$

with probability at least $1-\exp\{-n\}$ for some $c>0$. Consider the second term of (7)

$$\sup_{\gamma_1\in\mathcal{G}_{1,d}}\left(\frac{R^{2}_{\gamma_1}}{R^{2}_{\gamma_c}}\right)^{n/2}=\sup_{\gamma_1\in\mathcal{G}_{1,d}}\left(1-\frac{R^{2}_{\gamma_c}-R_{\gamma_1}}{R_{\gamma_c}}\right)^{n/2}\le 1,$$

by the fact that $R^2_{\gamma_c}-R_{\gamma_1}=\mathbf{y}'_n\left(P_n(\gamma_1)-P_n(\gamma_c)\right)\mathbf{y}_n\ge 0$ as $\gamma_c\subseteq\gamma_1$ and consequently, $P_n(\gamma_1)-P_n(\gamma_c)$ is non-negative definite matrix.

Next, consider the third expression of (7). The ratio

$$\inf_{\gamma_2\in\mathcal{G}_{2,d}}\left(\frac{R^{2}_{\gamma_2}}{R^{2}_{\gamma_c}}\right)^{n/2} = \inf_{\gamma_2\in\mathcal{G}_{2,d}}\left(1+\frac{R^{2}_{\gamma_2}-R^{2}_{\gamma_c}}{R^{2}_{\gamma_c}}\right)^{n/2}. \qquad (9)$$

Now, by assumption (A4)

$$\begin{aligned}
R^2_{\gamma_2} - R^2_{\gamma_c} &= \boldsymbol{\mu}_n' \left\{I - P_n(\gamma_2)\right\} \boldsymbol{\mu}_n + \mathbf{e}_n' \left\{P_n(\gamma_c) - P_n(\gamma_2)\right\} \mathbf{e}_n + 2\boldsymbol{\mu}_n' \left\{I - P_n(\gamma_2)\right\} \mathbf{e}_n \\
&\geq \boldsymbol{\mu}_n' \left\{I - P_n(\gamma_2)\right\} \boldsymbol{\mu}_n - 2\boldsymbol{\mu}_n' P_n(\gamma_2)\mathbf{e}_n \\
&\geq \Delta_0 - 2\left|\boldsymbol{\mu}_n'\mathbf{e}_n\right|,
\end{aligned}$$

uniformly over $\mathcal{G}_{2,d}$ as $\mathcal{G}_{2,d} \subseteq \mathcal{G}_0$, with probability one. By the choice of Δ_0 in (A4) and Lemma 3, we have $|\boldsymbol{\mu}_n'\mathbf{e}_n| = o\left(\Delta_0\right)$ with probability at least $1 - \exp\{-cn^{\xi}\}$ for $\xi > 0$ as in (A4) and some $c > 0$. Thus, from (9), by the above derivations,

$$\inf_{\gamma_2 \in \mathcal{G}_{2,d}} \left(1 + \frac{R^2_{\gamma_2} - R^2_{\gamma_c}}{R^2_{\gamma_c}}\right)^{n/2} \geq \left(1 + \frac{\Delta_0\left\{1 + o(1)\right\}}{n\sigma^2(1+\epsilon)}\right)^{n/2} \gtrsim \exp\left\{\Delta_0(1-\epsilon)/(2\sigma^2)\right\},$$

for any $\epsilon > 0$. Finally, it can be verified by examining the definitions of R^2_{γ} and $R^{2\star}_{\gamma_c}$ that the last part of RHS of (7) is bounded above by 1. Thus, combining all the above facts we get, for any $\epsilon > 0$, and with probability at least $1 - \exp\{-cn^{\xi}\}$ for some $c > 0$,

$$\sup_{\gamma_1 \in \mathcal{G}_{1,d}, \gamma_2 \in \mathcal{G}_{2,d}} \left(\frac{R^{\star 2}_{\gamma_1}}{R^{\star 2}_{\gamma_2}}\right)^{n/2} \leq \left(1 + \frac{c}{1 + g_n \tau_{\min}}\right)^{n/2} \exp\{-\Delta_0(1-\epsilon)/(2\sigma^2)\}$$

and

$$\begin{aligned}
&\sup_{\gamma_1 \in \mathcal{G}_{1,d}, \gamma_2 \in \mathcal{G}_{2,d}} \frac{P(M_{\gamma_2}|\mathcal{G}_d, \mathbf{y}_n)}{P(M_{\gamma_1}|\mathcal{G}_d, \mathbf{y}_n)} \\
&\leq 2\left(\frac{\tau_{\max}}{\tau_{\min} \wedge 1}\right)^{d_n/2} n^{d_n} \left(1 + \frac{c}{1 + g_n\tau_{\min}}\right)^{n/2} \exp\{-\Delta_0(1-\epsilon)/(2\sigma^2)\} \\
&\leq cn^{d_n} \exp\{-\Delta_0(1-\epsilon)/(2\sigma^2)\} \to 0,
\end{aligned}$$

for an appropriate constant $c > 0$. This completes the proof of part A.

[Part B.] Observe that, by choice of δ_0 in (A4)

$$\begin{aligned}
\frac{\sum_{\gamma_2 \in \mathcal{G}_{2,d}} P(M_{\gamma_2}|\mathcal{G}_d, \mathbf{y}_n)}{\sum_{\gamma_1 \in \mathcal{G}_{1,d}} P(M_{\gamma_1}|\mathcal{G}_d, \mathbf{y}_n)} &\leq \sup_{\gamma_1 \in \mathcal{G}_{1,d}, \gamma_2 \in \mathcal{G}_{2,d}} \frac{P(M_{\gamma_2}|\mathbf{y}_n)}{P(M_{\gamma_1}|\mathbf{y}_n)} \frac{|\mathcal{G}_{2,d}|}{|\mathcal{G}_{1,d}|} \\
&\leq cn^{d_n} \exp\{-\Delta_0(1-\epsilon)/(2\sigma^2)\} \frac{\binom{p_n}{d}}{\binom{p_n - p(\gamma_c)}{d - p(\gamma_c)}} \\
&\leq cn^{d_n} p_n^{p(\gamma_c)} \exp\{-2(1-\epsilon)p(\gamma_c)\log p_n\} \\
&\leq cn^{d_n} p_n^{-(1-2\epsilon)p(\gamma_c)}
\end{aligned}$$

with probability at least $1 - \exp\{-cn^{\xi}\}$ for some $c > 0$, and for any $\epsilon > 0$.

[Part C.] Observe that $P(\gamma \in \mathcal{G}_d \mid \mathcal{G}_d, \mathbf{y}_n) = 1$. Therefore,

$$\begin{aligned}
1 &= P(\gamma \in \mathcal{G}_{1,d} \mid \mathcal{G}_d, \mathbf{y}_n) + P(\gamma \in \mathcal{G}_{2,d} \mid \mathcal{G}_d, \mathbf{y}_n) \\
&= P(\gamma \in \mathcal{G}_{1,d} \mid \mathcal{G}_d, \mathbf{y}_n)\left\{1 + \frac{P(\gamma \in \mathcal{G}_{2,d} \mid \mathcal{G}_d, \mathbf{y}_n)}{P(\gamma \in \mathcal{G}_{1,d} \mid \mathcal{G}_d, \mathbf{y}_n)}\right\} \\
&= P(\gamma \in \mathcal{G}_{1,d} \mid \mathcal{G}_d, \mathbf{y}_n)\left\{1 + \frac{\sum_{\gamma_2 \in \mathcal{G}_{2,d}} P(M_{\gamma_2}|\mathcal{G}_d, \mathbf{y}_n)}{\sum_{\gamma_1 \in \mathcal{G}_{1,d}} P(M_{\gamma_1}|\mathcal{G}_d, \mathbf{y}_n)}\right\} \\
&\leq P(\gamma \in \mathcal{G}_{1,d} \mid \mathcal{G}_d, \mathbf{y}_n)\left\{1 + cn^{d_n} p_n^{-(1-2\epsilon)p(\gamma_c)}\right\},
\end{aligned}$$

with probability at least $1-\exp\{-cn^{\xi}\}$ from part B. Observe that from (A1) and the choice of $d_n \sim \log n$, the sequence $n^{d_n} p_n^{-(1-2\epsilon)p(\gamma_c)} \to 0$, as $p_n \to \infty$. Further, as $n \to \infty$, the probability $1-\exp\{-cn^{\xi}\}$ converges to one. This completes the proof. □

A.3. Proof of Theorem 2

Proof: Recall that, $P\left(M_{\gamma_c} \mid \mathbf{y}_n, \mathcal{G}^{\star}\right)$ is the posterior probability of the model M_{γ_c}, restricted to the class $\mathcal{G}^{\star}$. We will first provide an uniform probabilistic upper bound to $P\left(M_{\gamma_c} \mid \mathbf{y}_n, \mathcal{G}^{\star}\right)$ for any fixed $\gamma^{\star}$ such that $\gamma_c \in \gamma^{\star}$. Observe that

$$P\left(M_{\gamma_c} \mid \mathcal{G}^{\star}, \mathbf{y}_n\right) = \left\{1 + \sum_{\gamma \subseteq \gamma^{\star}, \gamma \neq \gamma_c} \frac{P(M_{\gamma} \mid \mathbf{y}_n)}{P(M_{\gamma_c} \mid \mathbf{y}_n)}\right\}^{-1}.$$

The ratio of posterior probabilities of any model to the true model is given by

$$\frac{P(M_{\gamma}|\mathbf{y}_n)}{P(M_{\gamma_c}|\mathbf{y}_n)} = \left(\frac{1}{p_n-1}\right)^{p_n(\gamma)-p(\gamma_c)} \left(\frac{R^{2\star}_{\gamma_c}}{R^{2\star}_{\gamma}}\right)^{n/2} \frac{\left|I + g_n X'_{\gamma_c} X_{\gamma_c}\right|^{1/2}}{\left|I + g_n X'_{\gamma} X_{\gamma}\right|^{1/2}}. \tag{10}$$

We split $\mathcal{G}$ into two subclasses as follows:

(i) *Supermodel of the true model,* $\mathcal{G}^{\star}_1 = \{\gamma : M_{\gamma_c} \subset M_{\gamma}\} \cap \mathcal{G}^{\star}$.

(ii) *Non-supermodels,* $\mathcal{G}^{\star}_2 = \{\gamma : M_{\gamma_c} \not\subseteq M_{\gamma}\} \cap \mathcal{G}^{\star}$.

Case I: Super-models ($\gamma \in \mathcal{G}^{\star}_1$) First, we obtain a uniform upper bound for the ratio of the posterior probabilities of any model M_{γ} and M_{γ_c}, given in (10). Note that

$$\frac{R^{2\star}_{\gamma}}{R^{2\star}_{\gamma_c}} = \frac{R^{2\star}_{\gamma}}{R^{2}_{\gamma}} \frac{R^{2}_{\gamma}}{R^{2}_{\gamma_c}} \frac{R^{2}_{\gamma_c}}{R^{2\star}_{\gamma_c}} \geq \left(1 - \frac{\epsilon}{n(1+\epsilon)}\right) \frac{R^{2}_{\gamma}}{R^{2}_{\gamma_c}} \tag{11}$$

by Lemma 3 and $R^{2\star}_{\gamma} \geq R^{2}_{\gamma}$, and with probability at least $1-\exp\{-cn\}$ for some $c > 0$.

Next, consider that for any $\epsilon > 0$ and $R = 2(1+\epsilon)$, we have

$$\begin{aligned} &P\left[\sup_{\gamma^{\star} \in \mathcal{G}_{1,d}} \sup_{\gamma_c \subseteq \gamma \subseteq \gamma^{\star}} \left(R^2_{\gamma_c} - R^2_{\gamma}\right) < R\sigma^2\{p_n(\gamma) - p(\gamma_c)\} \log p_n\right] \\ &= P\left[\sup_{\{\gamma : \gamma_c \subseteq \gamma, |\gamma| \leq d_n\}} \left(R^2_{\gamma_c} - R^2_{\gamma}\right) < R\sigma^2\{p_n(\gamma) - p(\gamma_c)\} \log p_n\right]. \end{aligned} \tag{12}$$

The last equality holds due to the equality of the sets

$$\{\gamma : \gamma_c \subseteq \gamma, |\gamma| \leq d_n\} = \{\gamma : \gamma_c \subseteq \gamma \subseteq \gamma^{\star}, \gamma^{\star} \in \mathcal{G}_{1,d}\}.$$

Next, observe that the right-hand side (RHS) of (12) is bounded above by

$$\begin{aligned}
&\sum_{\{\gamma:\gamma_c\subseteq\gamma,|\gamma|\leq d_n\}} P\left[\left(R^2_{\gamma_c}-R^2_{\gamma}\right)<R\sigma^2\{p_n(\gamma)-p(\gamma_c)\}\log p_n\right] \\
&\leq \sum_{p_n(\gamma)-p(\gamma_c)=1}^{d_n-p(\gamma_c)} \binom{p_n-p(\gamma_c)}{p_n(\gamma)-p(\gamma_c)} \exp\left\{-R\{p_n(\gamma)-p(\gamma_c)\}\log p_n/2\right\} \\
&\leq \sum_{p_n(\gamma)-p(\gamma_c)=1}^{d_n-p(\gamma_c)} (p_n-p(\gamma_c))^{p_n(\gamma)-p(\gamma_c)} p_n^{-R\{p_n(\gamma)-p(\gamma_c)\}/2} \\
&\leq \quad (d_n-p(\gamma_c))p_n^{-\epsilon}\to 0, \qquad (13)
\end{aligned}$$

where $\epsilon>0$ be any constant. Therefore, with probability at least $1-cp_n^{-\epsilon}$ for any $\epsilon>0$ and an appropriate $c>0$, the following holds uniformly over $\{\gamma:\gamma_c\subseteq\gamma\subseteq\gamma^\star,\gamma^\star\in\mathcal{G}_{1,d}\}$

$$\left(\frac{R^{2\star}_{\gamma_c}}{R^{2\star}_{\gamma}}\right)^{n/2}\leq(1+\epsilon)\left(1-\frac{R(p_n(\gamma)-p(\gamma_c))\log p_n}{n(1-\epsilon)}\right)^{-n/2}\lesssim(1+\epsilon)p_n^{(1+\epsilon)(p_n(\gamma)-p(\gamma_c))}.$$

Again, by Lemma 1 and assumptions (A2)-(A3) we have

$$\frac{\left|I+g_nX'_\gamma X_\gamma\right|^{-1/2}}{\left|I+g_nX'_{\gamma_c}X_{\gamma_c}\right|^{-1/2}}\leq cg_n^{-(p_n(\gamma)-p(\gamma_c))/2}n^{p(\gamma_c)},$$

where $c>0$ is some appropriate constant. Therefore, summing the ratio of posterior probabilities over $M_\gamma\in\mathcal{G}_1^\star$, we have

$$\begin{aligned}
\sum_{\gamma\in\mathcal{G}_1^\star}\frac{p(M_\gamma|\mathbf{y}_n)}{p(M_{\gamma_c}|\mathbf{y}_n)} &\leq n^{p(\gamma_c)}\sum_{\gamma\in\mathcal{G}_1^\star}\frac{cp_n^{(1+\epsilon)(p_n(\gamma)-p(\gamma_c))}}{\{\sqrt{g_n}(p_n-1)\}^{p_n(\gamma)-p(\gamma_c)}} \\
&\leq \sum_{p_n(\gamma)-p(\gamma_c)=1}^{d_n-p(\gamma_c)}\binom{d_n-p(\gamma_c)}{p_n(\gamma)-p(\gamma_c)}n^{p(\gamma_c)}c\left(\frac{p_n^{2\epsilon}}{g_n}\right)^{(p_n(\gamma)-p(\gamma_c))/2} \\
&\leq c2^{d_n-p(\gamma_c)}n^{p(\gamma_c)}\left(\frac{p_n^{2\epsilon}}{g_n}\right)^{1/2} \\
&\leq c2^{d_n-p(\gamma_c)}n^{p(\gamma_c)}\left(\frac{p_n^{2\epsilon}}{g_n}\right)^{1/2}\leq cn^{p(\gamma_c)+1}\left(\frac{p_n^{2\epsilon}}{g_n}\right)^{1/2}
\end{aligned}$$

for any $\epsilon>0$ and a suitable choice of $c>0$. When we choose $\epsilon<\delta/3$, we get that the above expression converges to 0, as $p_n\to\infty$.

Case II: Non-super models ($\gamma\in\mathcal{G}_2^\star$) We split $R^{2\star}_\gamma/R^{2\star}_{\gamma_c}$ as before in (11). Observe that

$$\begin{aligned}
R^2_\gamma-R^2_{\gamma_c} &= \mathbf{y}'_n(P_n(\gamma_c)-P_n(\gamma))\mathbf{y}_n \\
&= \boldsymbol{\mu}'_n(P_n(\gamma_c)-P_n(\gamma))\boldsymbol{\mu}_n+2\boldsymbol{\mu}'_n(P_n(\gamma_c)-P_n(\gamma))\mathbf{e}_n+\mathbf{e}'_n(P_n(\gamma_c)-P_n(\gamma))\mathbf{e}_n \\
&\geq \boldsymbol{\mu}'_n(P_n(\gamma_c)-P_n(\gamma))\boldsymbol{\mu}_n-2|\boldsymbol{\mu}'_n\mathbf{e}_n|.
\end{aligned}$$

Note that $\boldsymbol{\mu}_n'(P_n(\gamma_c) - P_n(\gamma))\boldsymbol{\mu}_n = \boldsymbol{\mu}_n'(I - P_n(\gamma))\boldsymbol{\mu}_n > \Delta_0$ uniformly over the class of all small dimensional non-supermodels by assumption (A4). Further, by Lemma 3, we get $|\boldsymbol{\mu}_n'\mathbf{e}_n| = o\left(\Delta_0\right)$ with probability at least $1-\exp\{-cn^{\xi}\}$ for $\xi > 0$ as in (A4) and some $c > 0$. Combining all these facts and using (A4), we have with probability at least $1 - \exp\{-cn^{\xi}\}$

$$\sup_{\gamma^\star \in \mathcal{G}_{1,d}} \sup_{\gamma \in \mathcal{G}_2^\star} \left(\frac{R_\gamma^{2\star}}{R_{\gamma_c}^{2\star}}\right)^{-n/2} \leq (1+\epsilon)\left(1 + (1-\epsilon)\frac{\Delta_0}{n\sigma^2}\right)^{-n/2} \lesssim (1+\epsilon)\exp\left\{-(1-\epsilon)\Delta_0/2\sigma^2\right\}.$$

Further, from Lemma 1, the ratio of determinants in the last term of (10) is less than $c\left(n\sqrt{g_n\tau_{\max}}\right)^{p(\gamma_c)}$ for an appropriately chosen $c > 0$. Therefore,

$$\begin{aligned}\sup_{\gamma^\star \in \mathcal{G}_{1,d}} \sum_{\gamma \in \mathcal{G}_2^\star} \frac{p(M_\gamma|\mathbf{y}_n)}{p(M_{\gamma_c}|\mathbf{y}_n)} &\leq c\left(np_n\sqrt{g_n\tau_{\max}}\right)^{p(\gamma_c)}\exp\left\{-(1-\epsilon)\frac{\Delta_0}{2\sigma^2}\right\}\sum_{q=1}^{d_n}\binom{d_n}{q}\frac{1}{(p_n-1)^q} \\ &\leq c\left(\frac{n\sqrt{g_n}}{p_n^{1-2\epsilon}}\right)^{p(\gamma_c)}, \end{aligned} \tag{14}$$

for any $\epsilon > 0$, with probability at least $1-\exp\{-cn^{\xi}\}$ for $\xi > 0$ as in (A4), and uniformly over $\gamma^\star \in \mathcal{G}_{1,d}$. Combining the above facts, we get, with probability at least $1-cp_n^{-c_0}-\exp\{-cn^{\xi}\}$, where $c_0 \ll \delta/2$ and $\xi > 0$ is as in (A4),

$$\inf_{\gamma^\star \in \mathcal{G}_{1,d}} P\left(M_{\gamma_c} \mid \mathbf{y}_n, \mathcal{G}^\star\right) \geq \left[1 + cn^{p(\gamma_c)+1}\left(\frac{p_n^{2\epsilon}}{g_n}\right)^{1/2} + c\left(\frac{n\sqrt{g_n}}{p_n^{1-2\epsilon}}\right)^{p(\gamma_c)}\right]^{-1}$$

for some $\epsilon \ll \delta/2$, where δ is as in the choice of g_n. For the choice of g_n taken in Theorem 2, the above expression converges to 1 as $p_n \to \infty$. □

A.4. Proof of Theorem 3

Proof: Observe that

$$\begin{aligned} P_2(M_{\gamma_c}|\mathbf{y}_n) &= \sum_{\gamma^\star \in \mathcal{G}_d} P_2(M_{\gamma_c}, M_{\gamma^\star}|\mathbf{y}_n) = \sum_{\gamma^\star \in \mathcal{G}_d} P(M_{\gamma_c}|\mathbf{y}_n, \mathcal{G}^\star)P_1(M_{\gamma^\star}|\mathbf{y}_n) \\ &= \sum_{\gamma^\star \in \mathcal{G}_{1,d}} P(M_{\gamma_c}|\mathbf{y}_n, \mathcal{G}^\star)P_1(M_{\gamma^\star}|\mathbf{y}_n) + \sum_{\gamma^\star \in \mathcal{G}_{2,d}} P(M_{\gamma_c}|\mathbf{y}_n, \mathcal{G}^\star)P_1(M_{\gamma^\star}|\mathbf{y}_n), \end{aligned}$$

where $M_{\gamma^\star}$ is the model chosen in the first stage. Observe that $P(M_{\gamma_c}|\mathbf{y}_n, \mathcal{G}^\star) = 0$ if $\gamma^\star \in \mathcal{G}_{2,d}$, *i.e.*, if the model chosen in the first stage is a non-supermodel. Therefore,

$$\begin{aligned} P_2(M_{\gamma_c}|\mathbf{y}_n) &= \sum_{\gamma^\star \in \mathcal{G}_{1,d}} P(M_{\gamma_c}|\mathbf{y}_n, \mathcal{G}^\star)P_1(M_{\gamma^\star}|\mathbf{y}_n) \\ &\geq \inf_{\gamma^\star \in A_{1,d}} P\left(M_{\gamma_c} \mid \mathbf{y}_n, \mathcal{G}^\star\right)\left[1 + \frac{\sum_{\gamma^\star \in \mathcal{G}_{2,d}} P_1(M_{\gamma^\star}|\mathbf{y}_n)}{\sum_{\gamma^\star \in \mathcal{G}_{1,d}} P_1(M_{\gamma^\star}|\mathbf{y}_n)}\right]^{-1} \\ &\geq \left[1 + cn^{p(\gamma_c)+1}\left(\frac{p_n^{2\epsilon}}{g_n}\right)^{1/2} + c\left(\frac{n\sqrt{g_n}}{p_n^{1-2\epsilon}}\right)^{p(\gamma_c)}\right]^{-1}\left[1 + c_3 n^{d_n/2} p_n^{-(1-2\epsilon)p(\gamma_c)}\right]^{-1}, \end{aligned}$$

with a probability at least $1 - cp_n^{-c_0} - 2\exp\{-cn^{\xi}\}$, where $c_0 \ll \delta/2$ and $\xi > 0$ is as in (A4). Thus, $P_2(M_{\gamma_c}|\mathbf{y}_n) \to 1$ with probability tending to 1. □