Electronic Journal of Statistics
Vol. 13 (2019) 284–309
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1529

Fast Bayesian variable selection for high dimensional linear models: Marginal solo spike and slab priors

Su Chen

Department of Statistics and Data Sciences
University of Texas at Austin
e-mail: s.chen@utexas.edu

and

Stephen G. Walker

Department of Mathematics
University of Texas at Austin
e-mail: s.g.walker@math.utexas.edu

Abstract: This paper presents a method for fast Bayesian variable selection in the normal linear regression model with high dimensional data. A novel approach is adopted in which an explicit posterior probability for including a covariate is obtained. The method is sequential but not order dependent, one deals with each covariate one by one, and a spike and slab prior is only assigned to the coefficient under investigation. We adopt the well-known spike and slab Gaussian priors with a sample size dependent variance, which achieves strong selection consistency for marginal posterior probabilities even when the number of covariates grows almost exponentially with sample size. Numerical illustrations are presented where it is shown that the new approach provides essentially equivalent results to the standard spike and slab priors, i.e. the same marginal posterior probabilities of the coefficients being nonzero, which are estimated via Gibbs sampling. Hence, we obtain the same results via the direct calculation of p probabilities, compared to a stochastic search over a space of 2^p elements. Our procedure only requires p probabilities to be calculated, which can be done exactly, hence parallel computation when p is large is feasible.

Keywords and phrases: Bayesian variable selection, spike and slab priors, high dimensional linear model, strong selection consistency, parallel computation.

Received July 2018.

Contents

1. Introduction

Variable selection for the linear model is currently a topic of immense interest. In this paper, we consider the Gaussian linear regression model under high dimensional setting. In particular,

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \sigma\epsilon, \tag{1}$$

where $\boldsymbol{Y}$ is a $n \times 1$ vector of response variables, $\boldsymbol{X}$ is a $n \times p$ matrix of predictor variables, β is a $p \times 1$ vector of coefficients, σ^2 an unknown variance term and ϵ is a $n \times 1$ vector of i.i.d. standard normal random errors. Variable selection in the high-dimensional setup, $p >> n$, is a flourishing area, driven primarily by challenging applications in various fields like genetics, finance, machine learning, etc, with increasing availability of data. Sparsity has frequently been identified as an underlying feature for these kind of data sets. For example, in genetic studies, where the response variable corresponds to a particular observable trait, the number of subjects n may be of order 10^3, while the number of genetic features p can be of order 10^5. Despite the large number of features, usually only a few have a genuine association with the trait. Therefore, it is reasonable to assume that the true β, written as β^*, is sparse, i.e., has a fixed finite number of non-zero elements, even when p is growing. In other words, let S^* denote the set of indices for the active covariates in the true model. The sparsity assumption says that even if the total number of covariates p may grow with n, $|S^*|$ is fixed. Thus, zeros may be added to the coefficient vector β as n increases, but no nonzero components. So we assume after exceeding some dimension $p \geq |S^*|$, after which as n increases, only zeros can be added to the vector β. However, we do not need the index in the set S^* to be fixed, as long as $|S^*|$ is fixed. This is a typical assumption for the variable selection literature for diverging number of covariates regarding selection consistency, such as in [15] and [18].

There has been a substantial body of literature on variable selection for high-dimensional data in both the frequentist and the Bayesian paradigm, given the practical importance of the problem. Frequentist solutions are often available based on maximizing a penalized likelihood, with a penalty on the model complexity. This includes the well-known least absolute shrinkage and selection operator (LASSO), [22], the smoothly clipped absolute deviation SCAD, [9], the adaptive LASSO, [23], and the Dantzig selector, [4], among others. Another important frequentist solution involves a screening algorithm to first reduce the data dimension, and then to use some classical methods to perform variable selection on the reduced data. This idea is implemented in sure independence screening (SIS), [10], nonparametric independence screening (NIS), [8], iterative

varying-coefficient screening (IVIS), [21], to name a few. A detailed review of the selective methods above and other frequentist methods is provided by [10].

From the Bayesian perspective, popular methods for variable selection include [17], who introduced the spike and slab prior for specifically seeking a posterior probability of a model; stochastic search variable selection (SSVS), [12], and [13]; empirical Bayes variable selection, [11]; spike-and-slab variable selection, [14]. A comprehensive review is provided in [7] and [19]. Among more recent developments we mention the method of [2] which uses the idea of penalized credible regions; the non-local prior, [15]; Bayesian shrinking and diffusing (BASAD) prior [18]; and Spike and Slab Lasso [20]. Other notable works in Bayesian variable selection include [3], which involves multivariate distributions and [5], who use Bayes factors and model selection ideas.

Here we discuss the spike and slab prior, which is the most popular prior for Bayesian variable selection. A binary latent vector Z, of same dimension as the regression coefficient β, is usually introduced to indicate whether each coefficient β_j is "in or out". A prior distribution on the binary vector Z is assumed to be the prior distribution on model space. The prior for β_j given $Z_j = 0$ is usually a point mass at 0 or a normal distribution concentrated around 0, called the spike prior; and the prior for β_j given $Z_j = 1$ is a flat or diffused distribution, called the slab prior. The posterior distribution of the latent vector Z is used to identify the model with the highest posterior probability, thus the selection criteria. Various forms of spike and slab priors have been proposed together with different form of prior on the model space. To the best of our knowledge, all existing Bayesian variable selection methods rely on MCMC to do posterior inference; thus one drawback of all these methods is the feasibility of MCMC chain to fully explore the model space of dimension 2^p, particularly when p is moderate to large. The method proposed in this paper is born out of this concern; we do not need MCMC, indeed all the results are analytical.

To make the introduction concrete we first introduce the model and associated notation. We can rewrite (1) as

$$\boldsymbol{Y} = x_j\beta_j + \boldsymbol{X}_{[-\boldsymbol{j}]}\beta_{[-j]} + \sigma\epsilon$$

for $j = 1, \ldots, p$. Here the $\boldsymbol{x_j}$ is the jth column of the design matrix $\boldsymbol{X}$, and $\boldsymbol{X}_{[-\boldsymbol{j}]}$ is the design matrix $\boldsymbol{X}$ without the jth column. The reason why we isolate $\boldsymbol{x_j}$ and β_j will become clear later. We also introduce the latent binary variable using the same notation Z_j, to indicate whether the particular covariate β_j subject to selection is truly active or not. Without loss of generality, we assume that $\boldsymbol{Y}$ and each column of $\boldsymbol{X}$ are standardized with mean 0 and standard deviation 1, therefore no intercept term in the regression model.

With suitable normal priors for $(\beta_j, \beta_{[-j]})$ and σ^2, it is possible to obtain a closed form expression for the posterior distribution of $(\beta_j, \beta_{[-j]})$. However, for variable selection involving the spike and slab prior, this is no longer possible and necessarily MCMC methods need to be implemented, which is far from trivial. Using a particular model involving the spike and slab prior for β_j only, and a conjugate Gaussian prior for the other $\beta_{[-j]}$, we are able to write

down an explicit expression for $P(Z_j = 0|\text{data})$. Our procedure only requires p probabilities to be calculated, which can be done with exact expressions and in parallel, thus computation is super fast. We also prove that the marginal posterior probability for each covariate to be included in the model is asymptotically consistent if we carefully choose the prior parameters to depend on the sample size n and number of covariates p, with some regularity conditions on the design matrix. In addition, strong selection consistency holds in the sense that the posterior probability of the true model converges to one even when the number of covariates p grows with n at nearly exponential rate. In the paper, we have compared our approach to BASAD [18], which is the traditional spike-and-slab prior with sample size dependent prior variance, which utilizes a marginal posterior probability selection procedure to achieve strong selection consistency. We illustrate that our method provides results that are essentially equivalent to BASAD in two ways: numerically through simulation studies, and also theoretically through proving the same strong selection consistency with similar conditions.

The remaining sections of the paper are as follows. Section 2 describes the model, the motivation and some conditions on the prior parameters. Section 3 provides the posterior inference and describes our methodology for variable selection based on the proposed model. Section 4 presents the main results on the convergence of the posterior distribution of Z and the strong selection consistency. In Section 5 we discuss some computational aspects of the proposed method and present simulation studies. We also show performance of our method compared to some popular existing methods in variable selection. In Section 6, some possible strategies for tuning hyper parameters for the proposed method are discussed for practical purposes. The paper concludes with a discussion in Section 7.

2. The model

From here we use p_n to denote the number of covariates to indicate that it can grow with n. We assume that β is sparse in the sense that only a fixed and finite number of components are nonzero. Our goal is to identify the nonzero coefficients. The working model is as follows: for $j = 1, \ldots, p_n$,

$$\begin{aligned}
&\mathbf{Y}|(\mathbf{X}, \beta, \sigma^2) \sim \mathbf{N}(\beta_j x_j + X_{[-j]}\beta_{-j}, \sigma^2 I),\\
&\beta_j|Z_j, \sigma^2 \sim \begin{cases} \mathbf{N}(0, \sigma^2\tau_{0n}^2) & \text{if} \quad Z_j = 0\\ \mathbf{N}(0, \sigma^2\tau_{1n}^2) & \text{if} \quad Z_j = 1 \end{cases}\\
&\mathbb{P}(Z_j = 1) = 1 - \mathbb{P}(Z_j = 0) = q_{jn}\\
&\beta_{-j}|\sigma^2 \sim \mathbf{N}(\mathbf{0}, \sigma^2\tau_n^2 I), \quad \sigma^2 \sim \mathbf{IG}(a, b)
\end{aligned} \tag{2}$$

where $\mathbf{N}$ denotes normal and $\mathbf{IG}$ inverse gamma distribution. Here we assume $\boldsymbol{Y}$ and each column of $\boldsymbol{X}$ are standardized such that

$$\sum_{i=1}^{n} x_{ij} = 0, \quad \sum_{i=1}^{n} x_{ij}^2 = n, \quad \sum_{i=1}^{n} y_i = 0, \quad \text{and} \quad \sum_{i=1}^{n} y_i^2 = n. \tag{3}$$

For notational purposes, we denote the correlation coefficient between jth and kth columns of the design matrix $\mathbf{X}$ as ρ_{jk}, where

$$\rho_{jk} = \frac{x_j^T x_k}{\|x_j\|\|x_k\|} = \frac{x_j^T x_k}{n}. \tag{4}$$

We use the following notation for the support of true coefficients in the model:

$$S^* \subset \{1, 2, \ldots, p_n\}, \quad \text{and} \quad S^* = \{j : \beta_j^* \neq 0\}. \tag{5}$$

Strictly speaking we are working with p_n models, one for each β_j. We have labeled the prior variances for the spike and slab to include the factors τ_{0n}^2 and τ_{1n}^2, respectively, reflecting the fact that they depend on n. For all the other covariates not under investigation, we assign a conjugate Gaussian prior where τ_n^2 is a hyper-parameter. The idea is that we are selecting one covariate to be under investigation while giving an appropriate shrinkage prior to all the other covariates.

The motive for looking at each β_j separately through the marginal posterior distribution is as follows. Our argument is that using MCMC to infer the dependence structure in the full posterior using spike and slab priors for each β_j is inadequate, and will not guarantee to reveal the correct information about shared sparsity. For out of the 2^p possible models for the Markov chain to visit, it will actually only be a very small percentage which are visited. It would be very difficult to pick off an accurate dependence structure from such a chain. In some sense, for deciding accurately on the fate of each β_j, we argue it is best to treat β_{-j} as a nuisance parameter and to integrate them out of the model. Moreover, even if a MCMC has been employed, many variable selection procedures then do concentrate on the marginal posteriors. Though note that these marginal posteriors will not be exactly the same as ours; but should be close. The advantage of our marginal posteriors is that they are available explicitly and without the use of MCMC.

3. Posterior inference

Under the Bayesian model specified in the previous section, the probability density function of $\boldsymbol{Y}$ given $\boldsymbol{X}$, β_j, β_{-j} and σ^2, is

$$\begin{aligned} &p(Y|\beta_j, \beta_{-j}, \sigma^2) \\ =&(2\pi)^{-n/2}\sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}(Y - x_j\beta_j - X_{[-j]}\beta_{-j})^T(Y - x_j\beta_j - X_{[-j]}\beta_{-j})\right] \end{aligned} \tag{6}$$

If we multiply (6) by $p(\beta_{[-j]}|\sigma^2)$ which is $\mathbf{N}(\mathbf{0}, \sigma^2\tau_n^2 I)$, and integrate over $\beta_{[-j]}$, we obtain

$$\begin{aligned} &p(Y|\beta_j, \sigma^2) \\ \propto&\sigma^{-n}|X_{[-j]}^T X_{[-j]} + \tau^{-2}I|^{-1/2} \exp\left[-\frac{1}{2\sigma^2}(Y - x_j\beta_j)^T(I - \tilde{H}_j)(Y - x_j\beta_j)\right] \end{aligned} \tag{7}$$

where

$$\tilde{H}_j = X_{[-j]} \left(X_{[-j]}^T X_{[-j]} + \tau_n^{-2} I \right)^{-1} X_{[-j]}^T$$

is a hat-matrix with a regularization term. Note in (7) it is straightforward to find the mode for β_j, which is

$$\widehat{\beta_j} = \frac{x_j^T (I - \tilde{H}_j) Y}{x_j^T (I - \tilde{H}_j) x_j}. \tag{8}$$

What is worth pointing out is that this mode for β_j, which we obtained after integrating out all the other β's but does not include the prior for β_j, is almost identical to the jth element in the Ordinary Least Square estimator for β in the vanilla version of the linear model, with $p < n$. In fact, one can recover the jth element of $\hat{\beta}^{OLS} = (X^T X)^{-1} X^T Y$ by

$$\widehat{\beta_j}^{OLS} = \frac{x_j^T (I - H_j) Y}{x_j^T (I - H_j) x_j}, \tag{9}$$

where the $\tilde{H}_j$ is replaced by $H_j = X_{[-j]} (X_{[-j]}^T X_{[-j]})^{-1} X_{[-j]}^T$. This puts on evidence that there is little loss of information by integrating out the β_{-j}. In fact, the posterior estimate of β_j under our model is in the usual form of a weighted average of (8) and the prior mean, which we choose to center at 0. Thus it shares all the good asymptotic properties of (8), as is discussed in later sections.

Now multiply (7) by $p(\beta_j|\sigma^2)$ which is our spike and slab prior in (2), as a mixture of two normal distributions,

$$p(Y|\beta_j, \sigma^2) p(\beta_j|\sigma^2) \propto (1 - q_{jn})\, \omega_{0j}\, \mathbf{N}(\beta_j|\mu_{0j}, \xi_{0j}^2) + q_{jn}\, \omega_{1j}\, \mathbf{N}(\beta_j|\mu_{1j}, \xi_{1j}^2). \tag{10}$$

Given the conjugacy of inverse gamma prior of σ^2, naturally we can multiply (10) by $\mathrm{IG}(\sigma^2|a, b)$ and integrate out σ^2, so we obtain a mixture of two non-standardized Student-t distributions,

$$p(\beta_j|Y) \propto (1 - q_{jn}) F_{0j} \mathbf{t}_0(\beta_j|\nu, \mu_{0j}, \psi_{0j}) + q_{jn} F_{1j} \mathbf{t}_1(\beta_j|\nu, \mu_{1j}, \psi_{1j}) \tag{11}$$

where $\nu = n + 2a$. Here, for $k \in \{0, 1\}$,

$$\mu_{kj} = \frac{x_j^T (I - \tilde{H}_j) Y}{x_j^T (I - \tilde{H}_j) x_j + \tau_{kn}^{-2}} \tag{12}$$

$$\xi_{kj}^2 = \frac{\sigma^2}{x_j^T (I - \tilde{H}_j) x_j + \tau_{kn}^{-2}} \tag{13}$$

$$\omega_{kj} = \sqrt{\frac{\tau_{kn}^{-2}}{x_j^T (I - \tilde{H}_j) x_j + \tau_{kn}^{-2}}} \exp\left\{ \frac{1}{2\sigma^2} \left(\frac{(x_j^T (I - \tilde{H}_j) Y)^2}{x_j^T (I - \tilde{H}_j) x_j + \tau_{kn}^{-2}} \right) \right\} \tag{14}$$

$$\psi_{kj} = \frac{b + \frac{1}{2}Y^T(I - \tilde{H}_j)Y - \frac{1}{2}\frac{(x_j^T(I-\tilde{H}_j)Y)^2}{x_j^T(I-\tilde{H}_j)x_j + \tau_{kn}^{-2}}}{(n+2a)(x_j^T(I-\tilde{H}_j)x_j + \tau_{kn}^{-2})} \tag{15}$$

$$F_{kj} = \sqrt{\frac{\tau_{kn}^{-2}}{x_j^T(I-\tilde{H}_j)x_j + \tau_{kn}^{-2}}} \times \left(b + \frac{Y^T(I-\tilde{H}_j)Y}{2} - \frac{(x_j^T(I-\tilde{H}_j)Y)^2}{2\left(x_j^T(I-\tilde{H}_j)x_j + \tau_{kn}^{-2}\right)} \right)^{-(\frac{n}{2}+a)}. \tag{16}$$

Here we label the two t distributions as $\mathbf{t}_0$ and $\mathbf{t}_1$, being the"spike" t-distribution and the "slab" t-distribution, respectively.

A natural variable selection procedure would be to use (11); if the weight in front of the "slab" t distribution exceeds some pre-specified threshold, we select this β_j, and can use the mean μ_{1j} as the posterior estimate of β_j; otherwise, we do not select this β_j and estimate it to be 0. Later we will show that such a procedure achieves strong selection consistency, meaning the posterior probability of the selected model being the true model converges to 1.

We first explore the asymptotic behavior of the mean μ_{1j} for the"slab" t-distribution. In fact, understanding the behavior of $x_j^T(I - \tilde{H}_j)Y$ is crucial as this expression also appears in F_{kj}, which is the key term to determine the posterior probability of a particular covariate being active or not. Now

$$\mu_{1j} \sim \mathbf{N}\left(\frac{x_j^T(I-\tilde{H}_j)X\beta^*}{x_j^T(I-\tilde{H}_j)x_j + \tau_{1n}^{-2}}, \quad \sigma^2 \, \frac{x_j^T(I-\tilde{H}_j)^2 x_j}{(x_j^T(I-\tilde{H}_j)x_j + \tau_{1n}^{-2})^2} \right). \tag{17}$$

There are two sources of bias for μ_{1j} as an estimator of β_j^*; one is introduced by the prior we put on β_{-j}, specifically through $\tilde{H}_j$; and the other one is introduced by the slab prior we put on β_j. This is easily seen if we compare μ_{kj} with $\hat{\beta}_j^{OLS}$, which is an unbiased estimator. Specifically,

$$\begin{aligned} &\frac{x_j^T(I-\tilde{H}_j)X\beta^*}{x_j^T(I-\tilde{H}_j)x_j + \tau_{1n}^{-2}} \\ =&\frac{x_j^T(I-\tilde{H}_j)x_j\beta_j^*}{x_j^T(I-\tilde{H}_j)x_j + \tau_{1n}^{-2}} + \frac{x_j^T(I-\tilde{H}_j)X_{[-j]}\beta_{-j}^*}{x_j^T(I-\tilde{H}_j)x_j + \tau_{1n}^{-2}} \\ =&\beta_j^* - \frac{\beta_j^*}{1 + x_j^T(I-\tilde{H}_j)x_j\tau_{1n}^2} + \frac{x_j^T(I-\tilde{H}_j)X_{[-j]}\beta_{-j}^*}{x_j^T(I-\tilde{H}_j)x_j + \tau_{1n}^{-2}}. \end{aligned} \tag{18}$$

The first bias term is easy to control with τ_{1n}, the prior variance of the slab normal, which is chosen to be large. In fact, all we need here is to have $x_j^T(I - \tilde{H}_j)x_j\tau_{1n}^2 \to \infty$ as $n \to \infty$, which is easy to achieve as $x_j^T(I - \tilde{H}_j)x_j$ will be of order n. If, in addition, we let $\tau_{1n}^2 \to \infty$, this will speed up the convergence rate of the posterior estimate of β_j. We will show in Section 4 that τ_{0n}^2 together with

q_{jn} controls the size of the model, whereas τ_{1n}^2 controls the bias of the nonzero β, and letting τ_{0n}^2 and τ_{1n}^2 to grow apart at a certain rate will be essential for achieving strong selection consistency and optimal rates of convergence. The following lemma sheds light on the behavior of the second bias term in (18). All the proofs are included in Appendix.

Lemma 1. *Let X_n be the design matrix defined in (2) with standardized columns and dimension $n \times p_n$, where $p_n > n$. Assume $\Sigma = \lim_{n\to\infty} \Sigma_n = \lim_{n\to\infty} X_n^T X_n/n$ is well defined and has all eigenvalues bounded away from 0. Then all the eigenvalues of $(I - \tilde{H}_n)$ are bounded away from 0 when $\sup_n n\tau_n^2 < \infty$.*

Lemma 2. *If $\sup_n n\tau_n^2 < \infty$, then $n^{-1}\, x^T(I - \tilde{H}_j)x = O(1)$ for any n-dimensional vector x with $x'x = n$.*

Lemma 3. *For any $j \in S^{*c}$ and $k \in S^*$, if*

$$\sup_{j\in S^{*c}} \max_{k\in S^*} \left| \frac{\rho_{jk}}{1 + \sum_{l\neq j, l\neq k} \rho_{kl}} \right| = O\left(\sqrt{n}\tau_n^2\right)$$

then $x_j^T(I - \tilde{H}_j)x_k = O(\sqrt{n})$.

Lemma 3 is essential to our result for pairwise consistency, when we consider a true inactive covariate. The key term we need is to bound $x_j^T(I-\tilde{H}_n)x_k$ where x_j is an inactive covariate and x_k is an active covariate.

Next, we have the following lemma regarding the behavior of μ_{0j} and μ_{1j}, the means of the"spike" and "slab" t-distributions, respectively. Not surprisingly, we would expect the "spike" t-distribution to converge to a point mass centered at 0, and the "slab" t-distribution to converge to a point mass centered at the true β^* value. This can be achieved by adding conditions on the spike and slab normal variances to be sample size dependent.

Lemma 4. *Given the same assumption of Lemmas 1 - 3, and also $n\tau_{0n}^2 \to 0$, $n\tau_{1n}^2 \to \infty$ as $n \to \infty$, then $\mu_{0j} \xrightarrow{\textbf{q.m.}} 0$, and $\mu_{1j} \xrightarrow{\textbf{q.m.}} \beta_j^*$ where β_j^* is the true parameter value for the jth covariate in (2), which implies $\mu_{0j} \xrightarrow{\boldsymbol{P}} 0$, and $\mu_{1j} \xrightarrow{\boldsymbol{P}} \beta_j^*$. Here $X_n \xrightarrow{\textbf{q.m.}} X$ (or $X_n \xrightarrow{L^2} X$) is defined by $\lim_{n\to\infty} \mathbb{E}[(X_n - X)^2] = 0$.*

Now we want to examine the posterior probability of a particular covariate β_j being active or not, i.e. whether we should select this covariate into the model. From (11), we have

$$\mathbb{P}(Z_j = 1|Y) = 1 - \mathbb{P}(Z_j = 0|Y) = \frac{q_{jn}F_{1j}}{q_{jn}F_{1j} + (1 - q_{jn})F_{0j}} \tag{19}$$

where the expressions of F_{1j} and F_{0j} are given in (16). To achieve strong selection consistency, one necessary condition is that $\mathbb{P}(Z_j = 0|Y) \to \mathbb{I}_{\{\beta_j^*=0\}}$, as $n \to \infty$. Therefore, we need to look at the asymptotic behavior of the key term F_{1j}/F_{0j}. From now on we assume $a = b = 0$ for simplicity since the same

asymptotic results follow easily for any $a, b > 0$. We also assume $q_{jn} = q_n$, that is the prior probability for any particular covariate to be an active one is the same. Rewrite F_{1j}/F_{0j} as

$$\frac{F_{1j}}{F_{0j}} = \sqrt{\frac{x_j^T(I-\tilde{H}_j)x_j\tau_{0n}^2+1}{x_j^T(I-\tilde{H}_j)x_j\tau_{1n}^2+1}}\left(\frac{Y^T(I-\tilde{H}_j)Y-\phi_{0j}^2}{Y^T(I-\tilde{H}_j)Y-\phi_{1j}^2}\right)^{\frac{n}{2}} \tag{20}$$

where, for $k \in \{0,1\}$,

$$\begin{aligned}\phi_{kj} &= \sqrt{x_j^T(I-\tilde{H}_j)x_j+\tau_{kn}^{-2}}\cdot\mu_{kj} \\ &\sim \mathbf{N}\left(\frac{x_j^T(I-\tilde{H}_j)X\beta^*}{\sqrt{x_j^T(I-\tilde{H}_j)x_j+\tau_{kn}^{-2}}},\quad \sigma^2\,\frac{x_j^T(I-\tilde{H}_j)^2x_j}{x_j^T(I-\tilde{H}_j)x_j+\tau_{kn}^{-2}}\right)\end{aligned} \tag{21}$$

The next lemma explores the asymptotic behavior of ϕ_{0j}. We show that with slightly stronger conditions on τ_{0n}^2, ϕ_{0j} inherits similar properties as μ_{0j}, as stated in Lemma 4. This is also essential to our result for strong selection consistency.

Lemma 5. *Given the same assumption of Lemma 4, and in addition $n\tau_{0n} \to 0$ as $n \to \infty$, then $\phi_{0j} \xrightarrow{q.m.} 0$.*

With all the fundamental lemmas in place, in the next section we present the main results of the paper.

4. Main results

In this section, we consider the model given by (2) and assume throughout the paper that $p_n > n \to \infty$. The same theoretical results can be obtained for the $p_n \le n$ case with relaxed conditions. However, our focus here would be the high dimensional case. We first state the conditions needed for the main results:

Condition 1. *Dimension of p_n: $p_n = e^{n\delta_n}$ for some $\delta_n \to 0$ as $n \to \infty$; that is, $\log(p_n)/n \to 0$.*

Condition 2. *Prior parameters:*

$$\sup_{j\in S^{*c}}\max_{k\in S^*}\left|\frac{\rho_{jk}}{1+\sum_{l\neq j,l\neq k}\rho_{kl}}\right| = O\left(\sqrt{n}\tau_n^2\right) \text{ and } \tau_n^2 = O\left(\frac{1}{n}\right),$$

with $n\tau_{0n} \to 0$, $n\tau_{1n}^2 \to \infty$, $\log(\tau_{1n}/\tau_{0n})/n \to 0$ and $q_n \sim p_n^{-1}$

Condition 3. *Regularity of the design: The maximum nonzero eigenvalues of the Gram matrix X^TX/n are bounded away from infinity.*

We will discuss these conditions after Theorem 1 which states the pairwise consistency.

Theorem 1. *Assume Conditions 1 - 3 hold. From model (2), there exists an increasing sequence* d_n *with* $\lim_{n\to\infty} d_n = d > 0$*, depending on the data, such that,*

$$\begin{aligned} \text{If } \beta_j^* &= 0, \quad \mathbb{P}(Z_j = 1|Y) = O_P\left(q_n \frac{\tau_{0n}}{\tau_{1n}}\right) && (22)\\ \text{If } \beta_j^* &\neq 0, \quad \mathbb{P}(Z_j = 0|Y) = O_P\left(\frac{1}{q_n}\frac{\tau_{1n}}{\tau_{0n}} e^{-nd_n}\right). && (23) \end{aligned}$$

Therefore, $\mathbb{P}(\mathbb{P}(Z_j = \mathbb{I}_{\{\beta_j^*=0\}}|Y) > \epsilon) \to 0$ *for any* $\epsilon > 0$.

The proof is provided in the Appendix. The following arguments give some heuristics for the pairwise consistency in Theorem 1. It states that the posterior probability of misspecifying a particular covariate (that is either including a inactive covariate or not including an active one) goes to 0 as we gather more data, given the conditions above. The speed of convergence for pairwise consistency depends on two things: the ratio of the spike variance (τ_{0n}) over the slab variance (τ_{1n}), and the choice of q_n, which is the prior probability of including a particular covariate.

To better understand the asymptotic behaviors of the key terms and the role each condition plays, we take a simple example of the $p = 2$ case as an illustration. Suppose $y = \beta_1 x_1 + \beta_2 x_2 + \sigma\epsilon$, and we consider β_1. We assume the data is centered and scaled such that $x_1'x_1 = x_2'x_2 = y^T y = n$, $x_1'x_2 = x_2'x_1 = n\rho$. Simple calculations gives the following quantities,

$$\begin{aligned} &\tilde{H}_1 = \frac{\tau_n^2}{n\tau_n^2+1} x_2 x_2^T \\ &x_1^T(I-\tilde{H}_1)x_1 = n\left(\frac{n\tau_n^2(1-\rho^2)+1}{n\tau_n^2+1}\right) \\ &x_1^T(I-\tilde{H}_1)y = n\left(\frac{n\tau_n^2(1-\rho^2)+1}{n\tau_n^2+1}\beta_1 + \frac{\rho}{n\tau_n^2+1}\beta_2 + \frac{1}{n}\left(x_1 - \frac{n\tau_n^2\rho}{n\tau_n^2+1}x_2\right)^T \epsilon\right) \\ &y^T(I-\tilde{H}_1)y = n\left(1 - \frac{n\tau_n^2}{n\tau_n^2+1}\left((\beta_1\rho+\beta_2)^2 + \left(\frac{x_2^T\epsilon}{n}\right)^2\right)\right). \end{aligned} \tag{24}$$

Assuming $n\tau_{0n}^2 \to 0$, and $n\tau_{1n}^2 \to \infty$ then

$$\begin{aligned} &\text{(i) if } \beta_1 = 0, \\ &\mathbb{P}(Z_1 = 1|Y) \sim q_n \frac{\tau_{0n}}{\tau_{1n}}\left(1 + \frac{\rho^2\beta_2^2}{(1+n\tau_n^2(1-\rho^2))(1+n\tau_n^2(1-\beta_2^2))}\right)^{\frac{n}{2}} \\ &\text{(ii) if } \beta_1 \neq 0, \\ &\mathbb{P}(Z_1 = 0|Y) \sim \frac{1}{q_n}\frac{\tau_{1n}}{\tau_{0n}}\left(1 - \frac{(\beta_1 + n\tau_n^2(1-\rho^2)\beta_1 + \rho\beta_2)^2}{(1+n\tau_n^2(1-\rho^2))(1+n\tau_n^2(1-(\rho\beta_1+\beta_2)^2))}\right)^{\frac{n}{2}}. \end{aligned} \tag{25}$$

Here we look for conditions to ensure both probabilities above converge to 0. The second probability has an exponential decay term no matter how $n\tau_n^2$ behaves,

therefore the only condition required is $\log(\tau_{1n}/\tau_{0n})/n \to 0$. Now for the first probability to also converge to 0, if $n\tau_n^2$ is bounded, we need $|\rho| = O\left(\frac{1}{\sqrt{n}}\right)$; and if $n\tau_n^2 \to \infty$, then we need $|\rho| = O\left(\sqrt{n}\tau_n^2\right)$. We can see that these conditions are a special case of the general conditions where p_n diverges, because here the term $\left|\frac{\rho_{jk}}{1+\sum_{l\neq j, l\neq k}\rho_{kl}}\right|$ collapses to $|\rho|$ in the $p = 2$ case. This over simplified example provides some insights about how each part controls the posterior probability and affects pairwise consistency when selecting one particular covariate.

In the general p_n case, if $\beta_j^* = 0$, then both ϕ_{1j} and ϕ_{0j} defined in (21) are stochastically bounded and centered at 0 (see Theorem 1). The dominant term of $\zeta_n = \left\{(Y^T(I-\tilde{H}_j)Y - \phi_{0j}^2)/(Y^T(I-\tilde{H}_j)Y - \phi_{1j}^2)\right\}^{n/2}$ in (20) is $Y^T(I-\tilde{H}_j)Y$ which grows at the order of n, thus ζ_n converges to a constant. If $\beta_j^* \neq 0$, then ϕ_{0j} is still stochastically bounded and centered at 0, but the mean of ϕ_{1j} is growing at the order of $\sqrt{n}$. Therefore, $(Y^T(I-\tilde{H}_j)Y - \phi_{1j}^2)/(Y^T(I-\tilde{H}_j)Y - \phi_{0j}^2)$ is strictly less than 1, and hence ζ_n decays exponentially in n.

Here we discuss the necessity of the conditions in order to support (22) and (23). Condition 1 restricts the number of covariates to be no greater than exponential n, and Condition 2 provides the shrinking and diffusing rates for the spike and slab prior. Particularly, we require that the variance of spike prior goes to 0 faster than n^{-2}, and the variance of slab prior to be larger in order than n^{-1}. In other words, we need the spike variance and slab variance to be growing apart at a speed of n. However, we also require that this speed not be faster than exponential n. For the variance of the Gaussian prior we put on all the other covariates not subject to selection, we require it to be bounded both above and below. This makes sense intuitively, as we get more data, we do want to put more shrinkage on all the other covariates, given the assumption of sparsity; but neither do we want to shrink them too aggressively, in a way that the shrinkage "cancels" out any information from all the other covariates. It is not surprising that the condition on τ_n^2 depends on p_n, but implicitly through all the correlations between the columns of the design matrix.

We also require that $p_n q_n$ is bounded, which is a natural assumption given the sparsity of β, such that the true β only has a fixed number of nonzero elements. Condition 3 restricts the maximum and minimum non zero eigenvalues of the Gram matrix to be bounded away from infinity and zero as the dimension of the matrix grows as n grows; a standard assumption for the design matrix.

Note that Condition 2 is not only a condition on prior parameters, but also a condition on identifiability, because it implies

$$\sup_{j\in S^{*c}} \max_{k\in S^*} \left|\frac{\rho_{jk}}{1+\sum_{l\neq j, l\neq k}\rho_{kl}}\right| = O\left(\frac{1}{\sqrt{n}}\right). \tag{26}$$

We did not list (26) as a separate condition for the sake of being concise. This condition explicitly ensures identifiability as it restricts the magnitude of the correlation between active and inactive covariates in the true model. Identifiability is an assumption embedded in all variable selection methods for

diverging numbers of covariates. The scale of $1/\sqrt{n}$ is no more than assuming asymptotic independence between the active covariates and inactive ones. The key point is that if asymptotically the active and inactive covariates are highly correlated then strong selection consistency cannot be achieved as p grows much faster than n. We see that this condition is equivalent to Condition 4.4 in [18], which is stated to be a"mild regularity condition that allows us to identify the true model". We confirm this point here; when $p_n >> n$, (26) is trivial because there can be at most n columns of $\mathbf{X}$ being independent, and the remaining $p_n - n$ columns will be correlated, which implies $\left|\sum_{l\neq j, l\neq k}\rho_{kl}\right| = O(p_n - n)$. Thus, as long as $p_n - n \geq \sqrt{n}$, (26) is valid. Note that the restriction is only imposed on the correlation between the active and inactive covariates, and there is no restriction on the correlation structure within the active covariates or inactive covariates.

When $p_n \leq n$, Theorem 1 still holds with simpler conditions. The upper bound of τ_n^2 in Condition 2 is no longer required as it is only needed for the hat matrix to be well defined when $p_n > n$, and the lower bound can be relaxed to $\tau_n^2 = O(\frac{1}{\sqrt{n}})$ because $\sup_{j\in S^{*c}}\max_{k\in S^*}\left|\frac{\rho_{jk}}{1+\sum_{l\neq j, l\neq k}\rho_{kl}}\right| \approx O(1)$ when p_n is small. As a consequence, identifiability is unnecessary as well.

Let T denote a p_n dimensional binary vector that represents the true model, i.e. each element in T being 0 represents an inactive covariate, and each element being 1 represents an active covariate. With such notation, we have the following theorem which ensures the strong selection consistency for our model selection procedure. We include β_j as an active covariate if the marginal posterior probability exceeds some specified cut off value, that is, $\mathbb{P}(Z_j = 1|Y) > c$ for some to be specified $0 < c < 1$.

Theorem 2. *Assume Conditions 1 - 3 hold. From model (2), we have* $\mathbb{P}(Z = T|Y) \xrightarrow{\mathbb{P}} 1$ *as* $n \to \infty$*, that is, the posterior probability of the true model goes to* 1 *as the sample size increases to* ∞*. In particular, for any* $0 < \epsilon < 1$*,* $\mathbb{P}[\mathbb{P}(Z_j = T_j|Y) > \epsilon$ *for all* $j = 1, \ldots, p_n] \geq 1 - O(\tau_{0n}/\tau_{1n}) \to 1$ *as* $n \to \infty$*.*

Theorem 2 ensures that the variable selection procedure based on the marginal posterior probabilities finds the right model with probability going to 1. When the number of potential covariates p_n is growing with n, convergence of marginal posterior probabilities does not assure consistency for overall model selection. Recall that in Condition 2 we require $q_n \sim p_n^{-1}$, which is essential to derive Theorem 2 from Theorem 1. This condition ensures the probability of incorrectly including an inactive covariate in the model will stay sufficiently small as p_n grows. Due to the sparsity assumption, the inactive covariates is of order p_n, to be canceled by the q_n term in (22). On the other hand, since the number of active covariates in the true model is of constant order, (23) going to 0 is enough to make sure the probability of missing an active covariate will also stay sufficiently small. Interestingly, (23) also depends on $q_n \sim p_n^{-1}$ in a way that we have $\log p_n/n \to 0$ so that the term q_n in the denominator will not cancel out the exponential term and the whole thing still goes to 0.

5. Simulation study

In this section, we validate the performance of the proposed method under several experimental settings, and compare with some existing competitive variable selection methods from both the classical, as well as the Bayesian paradigms. In particular, we compare our method with the least absolute shrinkage and selection operator (LASSO), [22], and Bayesian shrinking and diffusing prior (BASAD), [18]. We have used R code for all the methods, either from existing R packages or code kindly shared by the authors. In all simulation results, we will refer to our method as SoloSS; for Solo Spike and Slab priors.

The proposed method has four turning parameters. In all the empirical work we use

$$\tau_{0n}^2 = n^{-1}, \quad \tau_{1n}^2 = n, \quad \tau_n^2 = \begin{cases} \frac{1}{\sqrt{n}}, & \text{if } n \geq p_n \\ \frac{1}{n}, & \text{if } n < p_n \end{cases}, \quad \text{and} \quad q_n = 0.05.$$

These prior choices provide good performance. Our specific choice for the sparsity level is set to be 5%, which is the oracle choice of q_n for the $p_n = 100$ case. We discuss in Section 6 how to tune this parameter and a possible method to quickly find the oracle choice. Note here we are using the same choice of q_n for the $p_n = 1000$ case to demonstrate the performance of our method when q_n is not optimized. The variance of the Gaussian prior for all the other covariates not under selection is chosen to be different for the $p \leq n$ and $p > n$ case, as motivated by Theorem 1. All the tuning parameters for BASAD are chosen by the same criteria described in [18], and results were summarized from $5,000$ iterations of MCMC, with a burn-in of $1,000$. For the LASSO, the penalty parameter is optimized by cross-validation using the built-in function in the R package "glmnet".

We consider two values of p_n, namely 100 and 1000, with a fixed value of $n = 100$; similar scenarios in Narisetty et al. [18] and [15]. For both $p_n = 100$ and $p_n = 1000$, we set the number of active covariates to be 5 to reflect different sparsity levels, and the true value of the active coefficients are taken to be 2. In each case, we show results averaged over 100 data sets generated with different random seeds. Each data vector x_j of the design matrix $X' = (x_1, \cdots, x_n)$ is assumed to follow the Gaussian distribution with mean 0 and covariance matrix Σ_{p_n}, for $i = 1, \ldots, n$. We consider four types of covariance structure of $\Sigma_{p_n} = (\sigma_{i,j})$ for $1 \leq i, j \leq p_n$ described as following, and present simulation results for each case.

- Case 1. Identity: $\Sigma_{p_n} = I$ no correlation among the covariates.

TABLE 1
Simulation result for Case 1, $p_n = 100$

Methods	PP_0	PP_1	FDR	TPR	run time
LASSO	NA	NA	0.392	1.000	**0.002**
BASAD	0.014	**1.000**	0.008	1.000	23.806
SoloSS	**0.002**	0.999	**0.002**	1.000	0.006

TABLE 2
Simulation result for Case 1, $p_n = 1000$

Methods	PP_0	PP_1	FDR	TPR	run time
LASSO	NA	NA	0.529	1.000	**0.003**
BASAD	**0.000**	**0.873**	**0.002**	**0.920**	1382.163
SoloSS	0.004	0.848	0.057	0.872	0.880

- Case 2. Equal correlation: Σ_{p_n} has diagonal elements 1 and off diagonal elements 0.25. This exhibits a moderate dependence structure uniformly among the covariates.

TABLE 3
Simulation result for Case 2, $p_n = 100$

Methods	PP_0	PP_1	FDR	TPR	run time
LASSO	NA	NA	0.525	1.000	**0.002**
BASAD	0.015	**1.000**	0.011	1.000	24.363
SoloSS	**0.002**	0.999	**0.000**	1.000	0.006

TABLE 4
Simulation result for Case 2, $p_n = 1000$

Methods	PP_0	PP_1	FDR	TPR	run time
LASSO	NA	NA	0.741	1.000	**0.003**
BASAD	**0.000**	**0.895**	**0.003**	**0.948**	1031.487
SoloSS	0.005	0.863	0.047	0.886	0.751

- Case 3. Block dependence: Σ_{p_n} has block covariance setting where the true active covariates have common correlation $\rho_1 = 0.25$, and the true inactive covariates have common correlation $\rho_2 = 0.75$ and each pair of active and inactive covariate are assume to be independent. This interesting covariance structure is adopted from [18], where it attributes different correlations depending on whether the covariate is active or not.

TABLE 5
Simulation result for Case 3, $p_n = 100$

Methods	PP_0	PP_1	FDR	TPR	run time
LASSO	NA	NA	0.075	1.000	**0.002**
BASAD	0.022	1.000	0.012	1.000	25.064
SoloSS	**0.002**	1.000	**0.000**	1.000	0.006

TABLE 6
Simulation result for Case 3, $p_n = 1000$

Methods	PP_0	PP_1	FDR	TPR	run time
LASSO	NA	NA	0.100	1.000	**0.003**
BASAD	**0.001**	0.993	0.016	1.000	1102.925
SoloSS	0.003	**0.998**	**0.000**	1.000	0.845

- Case 4. Autoregressive: Σ_{p_n} is defined by $\sigma_{ij} = 0.5^{|i-j|}$ for $1 \leq i \leq j \leq p_n$. In this case, we have a decaying correlation structure depending on the distance $|i - j|$. As the distance increases, the correlation decreases.

TABLE 7
Simulation result for Case 4, $p_n = 100$

Methods	PP_0	PP_1	FDR	TPR	run time
LASSO	NA	NA	0.159	1.000	**0.002**
BASAD	0.014	1.000	0.015	1.000	25.596
SoloSS	**0.002**	1.000	**0.000**	1.000	0.007

TABLE 8
Simulation result for Case 4, $p_n = 1000$

Methods	PP_0	PP_1	FDR	TPR	run time
LASSO	NA	NA	0.162	1.000	**0.004**
BASAD	**0.000**	0.886	**0.007**	0.894	1098.953
SoloSS	0.004	**0.996**	0.018	**0.998**	0.648

The summary of our results are presented in Tables 1–8. In these tables, both BASAD and SoloSSS present the median probability model chosen by thresholding posterior probability of including a covariate at 0.5; this is shown to be optimal in some sense by [1]. The columns of the tables show the average marginal posterior probability assigned to inactive and active covariates, $PP0$ and $PP1$ respectively, false discovery rate (FDR), true positive rate (TPR), and run time on the same machine. Based on our simulation experiment, we highlight the following findings:

(i) Looking at all scenarios, LASSO often have higher true positive rate at the cost of overfitting and false discoveries, especially under the more sparse model setting for $p_n = 1000$. The false discovery rate can be as high as 70%, while our method remains less than 5%.
(ii) Our proposed method is performing better than BASAD in more than half of the scenarios, especially for Cases 3 and 4 where there is a moderate level of correlation among covariates.
(iii) In all scenarios where BASAD shows better results, our method shows only slightly worse result, where the difference is almost ignorable. We argue that given our method only takes a tiny fraction of time to run compared to BASAD, the performance of our proposed method is truly remarkable.
(iv) The reported runtime for our method is recorded as a fully parallelization procedure. It seems that LASSO is still the fastest one among the comparison, however the runtime for LASSO does NOT include the time to do cross-validation for selecting the optimal tuning parameter, while the simulation results presented for LASSO is indeed using the optimal tuning parameter selected by cross-validation.
(iv) Note that BASAD needs to compute the inverse of the covariance matrix for each iteration of MCMC, which is computationally prohibitive for ultrahigh-dimensional data. In fact, any other MCMC based Bayesian method available suffers from this computation bottleneck. Our method only requires to compute the inverse of the covariance matrix once while paralleling p_n same calculation, which is the key to ensure scalability to

big data problems.

Next, we present another simulation study which we raise p_n to be 5000, and we still consider the true model with 5 active covariates but we choose true β to be 0.6. All the other aspects in the set-up remains unchanged as the simulation shown before. This represent a very difficult scenario with high dimension, high sparsity, and low signal to noise ratio. In this case, we are comparing to LASSO, and Spike-and-Slab LASSO (SSLASSO) [20], which replaced the (BASAD) [18] method and acts as a practical compromise, because the traditional Spike-and-Slab priors that require MCMC to do posterior inference is not realistic in such high dimension. In both SSLASSO and BASAD, they showed simulation with $p_n = 1000$, which they considered the high dimension cases.

TABLE 9
Simulation result for Case 1, $p_n = 5000$, low signal

Methods	PP_0	PP_1	FDR	TPR	run time
LASSO	NA	NA	0.477	**0.714**	**0.009**
SSLASSO	NA	NA	**0.015**	0.528	0.683
SoloSS	0.002	0.474	0.201	0.476	10.593

TABLE 10
Simulation result for Case 2, $p_n = 5000$, low signal

Methods	PP_0	PP_1	FDR	TPR	run time
LASSO	NA	NA	0.821	**0.882**	**0.009**
SSLASSO	NA	NA	**0.163**	0.482	0.936
SoloSS	0.002	0.439	0.261	0.412	10.040

TABLE 11
Simulation result for Case 3, $p_n = 5000$, low signal

Methods	PP_0	PP_1	FDR	TPR	run time
LASSO	NA	NA	0.177	**1.000**	**0.01**
SSLASSO	NA	NA	0.928	0.03	2.475
SoloSS	0.002	**0.972**	**0.089**	0.989	9.088

TABLE 12
Simulation result for Case 4, $p_n = 5000$, low signal

Methods	PP_0	PP_1	FDR	TPR	run time
LASSO	NA	NA	0.211	**0.988**	**0.008**
SSLASSO	NA	NA	**0**	0.516	0.782
SoloSS	0.002	**0.978**	0.114	0.982	9.674

The summary of our results are presented in Tables 9–12. We have some interesting findings here:

(i) As expected, under such difficult set-up, the performance of all methods suffered greatly compared to that presented before. In general, no method has dominated others in all 4 cases.

(ii) We have not tuned our prior parameters in this scenario, so we were using the same choices as before. However, the SSLASSO R package is using cross-validation as a default choice to find optimal tuning parameters. Therefore, it is not surprising that their performance is better in some cases.

(iii) LASSO still suffers from very high FDR, especially in Cases 1 and 2, where SSLASSO performs better. Our method performs slightly worse than SSLASSO, while showing similar pattern in terms of striking a balance between FDR and TPR.

(iv) One surprising result is for Case 3 where the design matrix has block correlation structure, SSLASSO pretty much failed, and our method performs the best. The reason for this is that the EM algorithm that SSLASSO employs to explore the posterior mode will be easily stuck on a local mode when the posterior shows a clear sign of multi-modal. The extra long runtime of SSLASSO compared to other cases is also confirming that EM algorithm is having difficulty converging.

(iv) For Cases 3 and 4, our performance is similar to LASSO, but slightly better. The superior performance of our method in Case 3 is supported by the asymptotic theorem where one of the conditions to achieve strong selection consistency is to restrict the highest correlation between any active and inactive covariates while keeping the correlation between others not too small, as referred to the condition for identifiability in (26).

(v) If we include the time to run cross-validation, our runtime will be in similar magnitude as LASSO, but our method provide explicit probability of including a particular variable while neither LASSO and SSLASSO has the ability to do so.

To conclude, this simulation study shows under high dimension where most of other Bayesian variable selection methods no longer apply, our method has the tendency to mimic the better-perform method under different cases of design matrix, and can perform competitive or even better, which is a very desirable property. We are aware that SSLASSO is also very scalable to high dimension, but it can fail under certain scenarios where our method performs well, which makes these two methods complimentary to each other in some sense.

6. Strategies for tuning parameters

In this section, we discuss some possible strategies for tuning each of the prior parameters. Sensitivity analysis suggests results are quite robust to the choice of τ_{0n} and τ_{1n} as long as they are set to be far apart. Our default choices for them would be $\tau_{0n}^2 = 1/n$ and $\tau_{1n}^2 = n$, which are supported by the conditions to achieve strong selection consistency.

A default choice for τ_n^2 is to choose the variance for all the other $\beta_{[-j]}$ not subject to selection to be the same as β_j, the one put under the spike and slab prior. One can simply derive this from the conditional variance formula

$\text{Var}(\beta_j) = \mathbb{E}[\text{Var}(\beta_j \mid Z_j] + \text{Var}(\mathbb{E}[\beta_j \mid Z_j]) = q_n\tau_{1n}^2 + (1-q_n)\tau_{0n}^2$. One issue here is such a choice depends on other tuning parameters q_n, τ_{0n} and τ_{1n}, but in practice it is natural to replace q_n with $\frac{1}{p_n}$ and use the default choices for the spike and slab variances as above. Thus we suggest a default choice for τ_n^2 when $p_n >> n$ to be $n/p_n + (p_n - 1)/np_n$. Such a choice yields $\tau_n^2 \approx 1/n$, which is consistent with the asymptotic condition on τ_n^2. However, for the cases where $p_n \leq n$ and both are in relatively low dimensions, this default choice will rely heavily on the ratio of n and p_n, and thus may not be a reasonable choice. Here we recommend another way to tune τ_n^2 for small n and p_n, which is to use cross validation to choose the best ridge regression penalty parameter, and set τ_n^2 accordingly. This is based on the fact that putting a Gaussian prior on all the other $\beta_{[-j]}$s is equivalent to ridge regression, and even though our approach will integrate out all the other $\beta_{[-j]}$s when making selection decision on the particular β_j, using the optimal shrinkage penalty parameter for all the other $\beta_{[-j]}$s should not be a bad idea. The following tables show the same simulation study results using such tuning strategies for τ_n, and all the other tuning parameters remain the same. We highlight some measures improved by employing such tuning strategies, but notice that both have reasonable performance compared to our choices in Section 5.

TABLE 13
Simulation results for $\tau_n^2 = \frac{n}{p_n} + \frac{p_n - 1}{np_n}$

$p_n = 100$	PP_0	PP_1	FDR	TPR
Case 1	0.023	1.000	0.159	1.000
Case 2	0.023	1.000	0.132	1.000
Case 3	0.015	1.000	0.064	1.000
Case 4	0.014	1.000	0.062	1.000
$p_n = 1000$	PP_0	PP_1	FDR	TPR
Case 1	0.014	0.926	0.199	**0.940**
Case 2	0.014	0.933	0.172	**0.958**
Case 3	0.009	1.000	0.012	1.000
Case 4	0.012	0.998	0.120	0.998

TABLE 14
Simulation results for τ_n^2 *chosen by cross validation using R package glmnet*

$p_n = 100$	PP_0	PP_1	FDR	TPR
Case 1	0.002	0.988	**0.000**	0.996
Case 2	0.002	0.959	0.000	0.974
Case 3	0.002	1.000	0.000	1.000
Case 4	0.002	0.998	0.000	1.000
$p_n = 1000$	PP_0	PP_1	FDR	TPR
Case 1	0.007	0.872	0.110	**0.896**
Case 2	0.006	0.889	0.072	**0.918**
Case 3	0.004	1.000	0.002	1.000
Case 4	0.005	0.997	**0.030**	0.998

The most sensitive prior parameter is q_n, which is not surprising, as we are doing variable selection based on marginal posteriors for one covariate at a time;

thus the prior probability of including a covariate plays a more important role than other turning parameters. Here we discuss two possible tuning strategies for q_n:

- finding the oracle choice by repeatedly updating q_n from the data:
  1. start with $q_n = 1/p_n$, run the variable selection procedure.
  2. if number of active covariates selected from above is $\hat{s}$, update $q_n = \hat{s}/p_n$, and run the variable selection procedure again.
  3. stop until q_n does not update anymore, which would be the estimate of oracle choice for q_n.
  4. repeat by setting initial $q_n = 1/2$ and use same steps above to estimate the oracle choice of q_n. If the data is informative enough, it should be the same as in step 3. If not, set oracle estimate to be somewhere in between.
- setting q_n adaptively based on ridge regression estimate of β:
  - sort $\left|\hat{\beta}^{ridge}\right|$ by descending order, and choose $q_{jn} = 0.5\left|\frac{\hat{\beta_j}^{ridge}}{\max(\hat{\beta}^{ridge})}\right|$ for top $\rho\%$ of the β_j and use default value $1/p$ for the rest.
  - this strategy outperforms others when we have prior knowledge of at least $\rho\%$ of active covariates in the true model. It will be able to capture some small signal with such adaptive q_{jn}.

For simulation results showing in Section 5, we were using the oracle choice of $q_n = 0.05$ for the $p_n = 100$ scenario, and this was achieved by first test running our method by setting $q_n = 1/p_n$, and then a second test run of setting $q_n = 1/2$, and both selected 5 active covariates for all 4 data generating cases. Similar tuning approach can be adopted for larger p_n scenarios, but we chose to use the same choice of $q_n = 0.05$ to show robustness. Therefore it was safe to conclude our method produces satisfactory results under high dimensional setting even when the tuning parameter q_n was misspecified, and it will indeed detect sparseness in the data.

7. Discussion

In this paper, we have presented a fast Bayesian variable selection method for the sparse high-dimensional regression problem using a novel spike and slab prior. The method is sequential, which deals with each covariate one at a time, and an explicit posterior probability for including a covariate is obtained, without the computational burden of MCMC. This allows natural parallelization of computing p covariates at the same time for ultrahigh dimensional data, and avoids the daunting task of exploring the enormous model space with dimension of 2^p. Under mild regularity conditions on the design matrix, our approach achieves strong selection consistency in the sense that the posterior probability of the true model converges to one. Simulation studies show that the finite

sample performance under a variety of settings are equivalent with MCMC, yet using only a fraction of the computation time. To our knowledge, this is the only available MCMC free Bayesian method for variable selection under the high dimensional setting.

For high dimensional data, the strong selection consistency of Bayesian methods has only been established very recently. So [16] have shown the equivalence of posterior consistency and model selection consistency under appropriate sparsity assumptions, and [6] have proved theoretical results related to the posterior consistency for the regression parameters. To our knowledge, [18] is the only paper that has established strong selection consistency while allowing the number of covariates to grow at nearly exponential with sample size. However, we have achieved the same strong selection consistency result under similar conditions as in [18], but with a simpler theoretical proof and no computational burden of having to use MCMC. Finally, we believe our approach can be extended to more general models beyond linear regression.

Appendix A: Appendix

Proof of Lemma 1. For ease of notation write $\tilde{H} = X(X^TX + \tau_n^{-2}I)^{-1}X^T$ with the subscript n removed from X. Let $X/\sqrt{n} = SVD$ be the singular value decomposition, so S is a $n \times n$ unitary matrix, V is a $n \times p$ rectangular diagonal matrix with elements $(\lambda_1, \ldots, \lambda_n)$, and D is a $p \times p$ unitary matrix. Then $\Sigma_n = X^TX/n = D^TV^TVD$ where V^TV is a $p \times p$ diagonal matrix with the diagonal elements being the eigenvalues of Σ_n, which are $(\lambda_1^2, \ldots, \lambda_n^2)$. For the $p_n > n$ case, the rank of X^TX is n, the λ_j^2 are bounded away from 0, and write $\epsilon_n = 1/(n\tau_n^2)$. Then

$$\begin{aligned}
I - \tilde{H} &= I - X(X^TX + \tau_n^{-2}I)^{-1}X^T \\
&= I - SVD(D^TV^TVD + \epsilon_n D^TD)^{-1}D^TV^TS^T \\
&= I - SV(V^TV + \epsilon_n I)^{-1}V^TS^T \\
&= I - SV\left[\text{diag}\left(\tfrac{1}{\lambda_1^2+\epsilon_n}, \cdots, \tfrac{1}{\lambda_n^2+\epsilon_n}, \tfrac{1}{\epsilon_n}, \cdots, \tfrac{1}{\epsilon_n}\right)\right]V^TS^T \\
&= SS^T - S\left[\text{diag}\left(\tfrac{\lambda_1^2}{\lambda_1^2+\epsilon_n}, \cdots, \tfrac{\lambda_n^2}{\lambda_n^2+\epsilon_n}\right)\right]S^T \\
&= S\left[\text{diag}\left(\tfrac{\epsilon_n}{\lambda_1^2+\epsilon_n}, \cdots, \tfrac{\epsilon_n}{\lambda_n^2+\epsilon_n}\right)\right]S^T.
\end{aligned}$$

Hence the eigenvalues of $I - \tilde{H}$ are $(\epsilon_n/(\lambda_j^2 + \epsilon_n))_{j=1}^n$ which are clearly upper bounded by 1 and are bounded away from 0 when $\sup_n n\tau_n^2 < \infty$.

Proof of Lemma 2. This follows immediately from Lemma 1, since

$$\lambda_{\min}x^Tx \leq x^TAx \leq \lambda_{\max}x^Tx$$

where A is a symmetric matrix, with $\lambda_{\min}$ and $\lambda_{\max}$ being its minimum and maximum eigenvalues.

Proof of Lemma 3. Now $x_j^T(I-\tilde{H}_j)x_k$ is the kth element of the vector $x_j^T(I-\tilde{H}_j)X_{-j}$, and so we first look at $(I-\tilde{H}_j)X_{-j}$ and for simplicity of notation, we suppress the subscript j in $\tilde{H}_j$ and X_{-j}. Now

$$\begin{aligned}(I-\tilde{H})X &= X - X(X^TX+\tau_n^{-2}I)^{-1}X^TX \\ &= X - X(X^TX+\tau_n^{-2}I)^{-1}(X^TX+\tau_n^{-2}I-\tau_n^{-2}I) \\ &= X(X^TX+\tau_n^{-2}I)^{-1}\tau_n^{-2} \\ &= X(\tau_n^2X^TX+I)^{-1}.\end{aligned}$$

Now write $x_j^T(I-\tilde{H})X = x_j^TX(\tau_n^2X^TX+I)^{-1} = \omega$, then $x_jX = \omega(\tau_n^2X^TX+I)$ and we want to show that for each element $\omega_k/\sqrt{n} < \infty$ for all large n. Now $x_j^Tx_k = n\rho_{jk}$ for any $j \in S^{*c}$ and $k \in S^*$, hence

$$\begin{aligned}x_jX &= n[\rho_{j1},\ldots,\rho_{jp_n}] = \omega(\tau_n^2X^TX+I) \\ &= [\omega_1,\ldots,\omega_{p_n}]\begin{bmatrix} n\tau_n^2+1, n\tau_n^2\rho_{12},\cdots,n\tau_n^2\rho_{1p_n} \\ n\tau_n^2\rho_{21}, n\tau_n^2+1,\cdots,n\tau_n^2\rho_{2p_n} \\ \cdots \\ n\tau_n^2\rho_{p_n1}, n\tau_n^2\rho_{p_n2},\cdots,n\tau_n^2+1\end{bmatrix} \\ &= n\tau_n^2[\omega_1,\cdots,\omega_{p_n}]\begin{bmatrix} 1+\frac{1}{n\tau_n^2}, \rho_{12},\cdots,\rho_{1p_n} \\ \rho_{21}, 1+\frac{1}{n\tau_n^2},\cdots,\rho_{2p_n} \\ \cdots \\ \rho_{p_n1},\rho_{p_n2},\cdots,1+\frac{1}{n\tau_n^2}\end{bmatrix} \\ &= n\left[\tau_n^2\omega_1\sum_{l=1,l\neq j}^{p_n}\rho_{1l}+\frac{\omega_1}{n},\cdots,\tau_n^2\omega_{p_n}\sum_{l=1,l\neq j}^{p_n}\rho_{p_nl}+\frac{\omega_{p_n}}{n}\right].\end{aligned}$$

Hence,

$$\begin{aligned}\max_{k\in S^*}|\omega_k| &= \max_{k\in S^*}\frac{|\rho_{jk}|}{\tau_n^2\left|\sum_{l=1,l\neq j}^{p_n}\rho_{kl}\right|+\frac{1}{n}} \\ &\leq \frac{1}{\tau_n^2}\max_{k\in S^*}\frac{|\rho_{jk}|}{\left|\sum_{l=1,l\neq j}^{p_n}\rho_{kl}\right|} \\ &\leq \frac{1}{\tau_n^2}\sup_{j\in S^{*c}}\max_{k\in S^*}\frac{|\rho_{jk}|}{\tau_n^2\left|\sum_{l=1,l\neq j}^{p_n}\rho_{kl}\right|} \\ &\leq \sqrt{n}.\end{aligned}$$

Proof of Lemma 4. This is an immediate consequence of (17) and Lemma 1; as we have $\mathbb{E}(\mu_{1j}) \to \beta_j^*$, $\mathrm{Var}(\mu_{1j}) \to 0$, $\mathbb{E}(\mu_{0j}) \to 0$ and $\mathrm{Var}(\mu_{0j}) \to 0$.

Proof of Lemma 5. Given

$$\phi_{0j} \sim \mathbf{N}\left(\frac{x_j^T(I-\tilde{H}_j)X\beta^*}{\sqrt{x_j^T(I-\tilde{H}_j)x_j+\tau_{0n}^{-2}}},\quad \sigma^2\,\frac{x_j^T(I-\tilde{H}_j)^2x_j}{x_j^T(I-\tilde{H}_j)x_j+\tau_{0n}^{-2}}\right)$$

we have

$$\begin{aligned}
\mathbb{E}\,\phi_{0j} &= \frac{x_j^T(I-\tilde{H}_j)x_j\beta_j^*}{\sqrt{x_j^T(I-\tilde{H}_j)x_j+\tau_{0n}^{-2}}}+\frac{x_j^T(I-\tilde{H}_j)X_{[-j]}\beta_{[-j]}^*}{\sqrt{x_j^T(I-\tilde{H}_j)x_j+\tau_{0n}^{-2}}}\\
&\sim \frac{n\beta_j^*}{\sqrt{n+\tau_{0n}^{-2}}}+\frac{x_j^T(I-\tilde{H}_j)X_{[-j]}\beta_{[-j]}^*}{\sqrt{n+\tau_{0n}^{-2}}}\\
&\le n\tau_{0n}\beta_j^*+\tau_{0n}x_j^T(I-\tilde{H}_j)X_{[-j]}\beta_{[-j]}^*\to 0.\\
\operatorname{Var}\phi_{0j} &\sim \frac{n}{n+\tau_{0n}^{-2}}=\frac{n\tau_{0n}^2}{n\tau_{0n}^2+1}\to 0.
\end{aligned}$$

Since $n\tau_{0n}^2\to 0$ and $n^{-1/2}x_j^T(I-\tilde{H}_j)X_{[-j]}\beta_{[-j]}^*=O(1)$, using Lemma 1, it is that $\phi_{0j}\xrightarrow{\mathbf{qm}}0$.

Proof of Theorem 1. From the model given in (2),

$$p(\beta_j|Y) \propto (1-q_{jn})F_{0j}\mathbf{t}(\beta_j|\nu,\mu_{0j},\psi_{0j})+q_{jn}F_{1j}\mathbf{t}(\beta_j|\nu,\mu_{1j},\psi_{1j})$$

where $\nu=n+2a$, and for $k\in\{0,1\}$,

$$\begin{aligned}
\mu_{kj} &= \frac{x_j^T(I-\tilde{H}_j)Y}{x_j^T(I-\tilde{H}_j)x_j+\tau_{kn}^{-2}}\\
\xi_{kj}^2 &= \frac{\sigma^2}{x_j^T(I-\tilde{H}_j)x_j+\tau_{kn}^{-2}}\\
\psi_{kj} &= \frac{b+\frac{1}{2}Y^T(I-\tilde{H}_j)Y-\frac{1}{2}\frac{(x_j^T(I-\tilde{H}_j)Y)^2}{x_j^T(I-\tilde{H}_j)x_j+\tau_{kn}^{-2}}}{(n+2a)(x_j^T(I-\tilde{H}_j)x_j+\tau_{kn}^{-2})}\\
F_{kj} &= \sqrt{\frac{\tau_{kn}^{-2}}{x_j^T(I-\tilde{H}_j)x_j+\tau_{kn}^{-2}}}\left(b+\frac{1}{2}Y^T(I-\tilde{H}_j)Y-\frac{1}{2}\phi_{kj}^2\right)^{-(\frac{n}{2})}\\
\mu_{kj} &\sim \mathbf{N}\left(\frac{x_j^T(I-\tilde{H}_j)X\beta^*}{x_j^T(I-\tilde{H}_j)x_j+\tau_{kn}^{-2}},\quad \sigma^2\,\frac{x_j^T(I-\tilde{H}_j)^2x_j}{(x_j^T(I-\tilde{H}_j)x_j+\tau_{kn}^{-2})^2}\right)\\
\phi_{kj} &= \sqrt{x_j^T(I-\tilde{H}_j)x_j+\tau_{kn}^{-2}}\cdot\mu_{kj}\\
&\sim \mathbf{N}\left(\frac{x_j^T(I-\tilde{H}_j)X\beta^*}{\sqrt{x_j^T(I-\tilde{H}_j)x_j+\tau_{kn}^{-2}}},\quad \sigma^2\,\frac{x_j^T(I-\tilde{H}_j)^2x_j}{x_j^T(I-\tilde{H}_j)x_j+\tau_{kn}^{-2}}\right).
\end{aligned}$$

Now assume $a = b = 0$ and $q_{jn} = q_n$, we want to look at the key terms

$$\frac{F_{1j}}{F_{0j}} = \sqrt{\frac{x_j^T(I-\tilde{H}_j)x_j\tau_{0n}^2+1}{x_j^T(I-\tilde{H}_j)x_j\tau_{1n}^2+1}}\left(\frac{Y^T(I-\tilde{H}_j)Y-\phi_{0j}^2}{Y^T(I-\tilde{H}_j)Y-\phi_{1j}^2}\right)^{\frac{n}{2}}.$$

First we consider $\beta_j^* = 0$. Now

$$\mathbb{P}(Z_j = 1|Y) = \frac{q_n\frac{F_{1j}}{F_{0j}}}{q_n\frac{F_{1j}}{F_{0j}}+(1-q_n)} \sim q_n\frac{F_{1j}}{F_{0j}} = O_P\left(q_n\frac{\tau_{0n}}{\tau_{1n}}\right)$$

since

$$\begin{aligned}\frac{F_{1j}}{F_{0j}} &= \sqrt{\frac{x_j^T(I-\tilde{H}_j)x_j\tau_{0n}^2+1}{x_j^T(I-\tilde{H}_j)x_j\tau_{1n}^2+1}}\left(1+\frac{\phi_{1j}^2\phi_{0j}^2}{Y^T(I-\tilde{H}_j)Y-\phi_{1j}^2}\right)^{\frac{n}{2}}\\ &\to O_P\left(\frac{\tau_{0n}}{\tau_{1n}}\right)\end{aligned}$$

provided

$$\frac{\phi_{1j}^2-\phi_{0j}^2}{Y^T(I-\tilde{H}_j)Y-\phi_{1j}^2} \le O_P\left(\frac{1}{n}\right).$$

To see this, we have

$$\begin{aligned}&\frac{\phi_{1j}^2-\phi_{0j}^2}{Y^T(I-\tilde{H}_j)Y-\phi_{1j}^2}\\ =\ &\frac{\frac{(x_j^T(I-\tilde{H}_j)Y)^2}{x_j^T(I-\tilde{H}_j)x_j+\tau_{1n}^{-2}}-\frac{(x_j^T(I-\tilde{H}_j)Y)^2}{x_j^T(I-\tilde{H}_j)x_j+\tau_{0n}^{-2}}}{Y^T(I-\tilde{H}_j)Y-\frac{(x_j^T(I-\tilde{H}_j)Y)^2}{x_j^T(I-\tilde{H}_j)x_j+\tau_{1n}^{-2}}}\\ =\ &\frac{(x_j^T(I-\tilde{H}_j)Y)^2\left(\frac{\tau_{0n}^{-2}-\tau_{1n}^{-2}}{x_j^T(I-\tilde{H}_j)x_j+\tau_{0n}^{-2}}\right)}{(Y^T(I-\tilde{H}_j)Y)(x_j^T(I-\tilde{H}_j)x_j+\tau_{1n}^{-2})-(x_j^T(I-\tilde{H}_j)Y)^2}\\ \le\ &\frac{(x_j^T(I-\tilde{H}_j)Y)^2}{(Y^T(I-\tilde{H}_j)Y)(x_j^T(I-\tilde{H}_j)x_j)-(x_j^T(I-\tilde{H}_j)Y)^2}\\ \sim\ &\frac{(x_j^T(I-\tilde{H}_j)Y)^2}{(Y^T(I-\tilde{H}_j)Y)(x_j^T(I-\tilde{H}_j)x_j)}\\ \sim\ &\left(\frac{1}{n}x_j^T(I-\tilde{H}_j)Y)\right)^2\\ =\ &\left(\frac{1}{n}x_j^T(I-\tilde{H}_j)X_{[-j]}\beta_{[-j]}^*+\frac{1}{n}x_j^T(I-\tilde{H}_j)\epsilon\right)^2\\ =\ &O_P\left(\frac{1}{n}\right).\end{aligned}$$

Here the last equation is due to Lemma 3 and the following equations:

$$\mathbb{E}\left[\frac{1}{n}x_j^T(I-\tilde{H}_j)\epsilon\right] = \frac{1}{n}x_j^T(I-\tilde{H}_j)\mathbb{E}[\epsilon] = 0$$
$$\text{Var}\left(\frac{1}{n}x_j^T(I-\tilde{H}_j)\epsilon\right) = \frac{1}{n^2}x_j^T(I-\tilde{H}_j)^2x_j\text{Var}\left(\epsilon\right) = O\left(\frac{1}{n}\right).$$

On the other hand, if $\beta_j^* \neq 0$,

$$\mathbb{P}(Z_j = 0|Y) = \frac{(1-q_n)}{q_n\frac{F_{1j}}{F_{0j}} + (1-q_n)} \sim \frac{1}{q_n}\frac{F_{0j}}{F_{1j}} = O_P\left(\frac{1}{q_n}\frac{\tau_{1n}}{\tau_{0n}}e^{-nd_n}\right)$$

since

$$\begin{aligned}\frac{F_{0j}}{F_{1j}} &= \sqrt{\frac{x_j^T(I-\tilde{H}_j)x_j\tau_{1n}^2+1}{x_j^T(I-\tilde{H}_j)x_j\tau_{0n}^2+1}}\left(1-\frac{\phi_{1j}^2-\phi_{0j}^2}{Y^T(I-\tilde{H}_j)Y-\phi_{0j}^2}\right)^{\frac{n}{2}} \\ &\to O_P\left(\frac{\tau_{1n}}{\tau_{0n}}e^{-nd_n}\right).\end{aligned}$$

To show the last argument, since we have $\phi_{0j} \xrightarrow{\mathbf{qm}} 0$, thus $\phi_{0j}^2 \xrightarrow{\mathbf{P}} 0$, it suffices to show the following:

$$\begin{aligned}&\frac{\phi_{1j}^2}{Y^T(I-\tilde{H}_j)Y} \\ =\ &\frac{(x_j^T(I-\tilde{H}_j)Y)^2}{(x_j^T(I-\tilde{H}_j)x_j+\tau_{1n}^{-2})(Y^T(I-\tilde{H}_j)Y)} \\ =\ &\frac{x_j^T(I-\tilde{H}_j)x_j}{x_j^T(I-\tilde{H}_j)x_j+\tau_{1n}^{-2}}\frac{(x_j^T(I-\tilde{H}_j)Y)^2}{(x_j^T(I-\tilde{H}_j)x_j)(Y^T(I-\tilde{H}_j)Y)} \\ \sim\ &\frac{n}{n+\tau_{1n}^{-2}}\left(\beta_j^* + \frac{1}{n}x_j^T(I-\tilde{H}_j)X_{[-j]}\beta_{[-j]}^* + \frac{1}{n}x_j^T(I-\tilde{H}_j)\epsilon\right)^2 \\ \sim\ &(\beta_j^*)^2 > 0.\end{aligned}$$

Now the above quantity will be 0 only if $x_j^T(I-\tilde{H}_j)Y = 0$, and we argue this is not the case. Given x_j being a true active covariate, x_j cannot be completely uncorrelated with Y which prevents $x_j^TY = 0$; x_j cannot be a linear combination of any of other x_k in the design matrix which prevents $x_j^T(I-\tilde{H}_j) = 0$; and Y should not be in column space of design matrix which prevents $(I-\tilde{H}_j)Y = 0$.

Proof of Theorem 2. We have

$$P(Z \neq T|Y) \leq P\left(\cup_{j=1}^{p_n}(Z_j \neq T_j)|Y\right) \leq \sum_{j=1}^{p_n}P(Z_j \neq T_j|Y)$$

$$\begin{aligned} &= \sum_{j\in S^*} P(Z_j = 0|Y) + \sum_{j\in S^{*C}} P(Z_j = 1|Y) \\ &= s^* O_P\left(\frac{1}{q_n}\frac{\tau_{1n}}{\tau_{0n}}e^{-nd_n}\right) + (p_n - s^*)\, O_P\left(q_n\frac{\tau_{0n}}{\tau_{1n}}\right) \\ &= O_P\left(\frac{1}{q_n}\frac{\tau_{1n}}{\tau_{0n}}e^{-nd_n}\right) + O_P\left(\frac{\tau_{0n}}{\tau_{1n}}\right). \end{aligned}$$

Let E_j be the event that the marginal posterior probability of jth covariate $P(Z_j = T_j|Y) > \epsilon$ for any $0 < \epsilon < 1$. We show that $P[\cup_{j=1}^{p_n} E_j^c] \to 0$, then $P[P(Z_j = T_j|Y) > \epsilon \text{ for all } j = 1, \ldots, p_n] = P[\cap_{j=1}^{p_n} E_j] = 1 - P[\cup_{j=1}^{p_n} E_j^c] \to 1$. Hence,

$$\begin{aligned} P[\cup_{j=1}^{p_n} E_j^c] &= P[P(Z_j = T_j|Y) \le \epsilon \text{ for some } j = 1, \ldots, p_n] \\ &\le \sum_{j=1}^{p_n} P[P(Z_j = T_j|Y) \le \epsilon] \\ &= \sum_{j=1}^{p_n} P[P(Z_j \ne T_j|Y) > \epsilon] \\ &= O\left(\frac{1}{q_n}\frac{\tau_{1n}}{\tau_{0n}}e^{-nd_n}\right) + O\left(\frac{\tau_{0n}}{\tau_{1n}}\right) \to 0. \end{aligned}$$

Acknowledgements

The authors are grateful for the comments of an Associate Editor and referee made on an earlier version of the paper.

References

[1] Barbieri, M. M., Berger, J. O. et al. (2004). Optimal predictive model selection. *Annals of Statistics* **32** 870–897. MR2065192

[2] 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. MR3036420

[3] Brown, P. J., Vannucci, M. and Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60** 627–641. MR1626005

[4] Candes, E., Tao, T. et al. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics* **35** 2313–2351. MR2382644

[5] Casella, G. and Moreno, E. (2006). Objective Bayesian variable selection. *Journal of the American Statistical Association* **101** 157–167. MR2268035

[6] Castillo, I., Schmidt-Hieber, J., van der Vaart, A. et al. (2015). Bayesian linear regression with sparse priors. *Annals of Statistics* **43** 1986–2018. MR3375874

[7] Clyde, M. and George, E. I. (2004). Model uncertainty. *Statistical Science* 81–94. MR2082148
[8] 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. MR2847969
[9] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360. MR1946581
[10] Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* **20** 101. MR2640659
[11] George, E. and Foster, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87** 731–747. MR1813972
[12] George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88** 881–889.
[13] George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica sinica* 339–373.
[14] Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *Annals of Statistics* 730–773. MR2163158
[15] Johnson, V. E. and Rossell, D. (2012). Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association* **107** 649–660. MR2980074
[16] 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. MR3174644
[17] Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association* **83** 1023–1032. MR0997578
[18] Narisetty, N. N., He, X. et al. (2014). Bayesian variable selection with shrinking and diffusing priors. *Annals of Statistics* **42** 789–817. MR3210987
[19] O'Hara, R. B., Sillanpää, M. J. et al. (2009). A review of Bayesian variable selection methods: what, how and which. *Bayesian Analysis* **4** 85–117. MR2486240
[20] Ročková, V. and George, E. I. (2018). The spike-and-slab lasso. *Journal of the American Statistical Association* **113** 431–444. MR3803476
[21] Song, R., Yi, F. and Zou, H. (2014). On varying-coefficient independence screening for high-dimensional varying-coefficient models. *Statistica Sinica* **24** 1735. MR3308660
[22] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288. MR1379242
[23] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101** 1418–1429. MR2279469