Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ORIGINAL ARTICLE

Bayesian selection of best subsets via hybrid search

Shiqiang Jin | Gyuhyeong Goh*

[1]Department of Statistics, Kansas State University, Manhattan, KS 66506-0802, U.S.A.

Correspondence
*Gyuhyeong Goh, Department of Statistics, 101 Dickens Hall, Kansas State University, 1116 Mid-Campus Drive N., Manhattan, KS 66506-0802, U.S.A. Email: ggoh@ksu.edu

Summary

Over the decades, variable selection with high-dimensional data has drawn increasing attention. With a large number of predictors, there rises a big challenge for model fitting and prediction. In this paper, we develop a new Bayesian method of best subset selection that quickly finds the best subset via a hybrid search algorithm that combines deterministic local search and stochastic global search. To avoid the computational burden in evaluating multiple candidate subsets for each update, we propose a novel strategy that enables us to calculate exact posterior probabilities of all neighbor models simultaneously in a single computation. In addition, we establish model selection consistency for the proposed method in the high-dimensional setting in which the number of possible predictors can increase faster than the sample size. Simulation study and real data example are conducted to investigate the performance of the newly-developed method.

KEYWORDS:
Bayesian model selection, best subset selection, extended Bayesian information criterion, hybrid search algorithm

1 | INTRODUCTION

Variable selection plays a key role in recent regression analysis. In many statistical applications, especially in genetics studies, researchers often face situations in which the number of candidate predictors is extremely large but the sample size is relatively small, often referred to as high-dimensional regression problems. The most pressing challenge in high-dimensional regression is to identify relevant predictor variables from the huge pool of candidates. In an attempt to perform high-dimensional variable selection, a lot of effort has been put into the development of penalized likelihood methods (e.g., Fan and Li 2001; Tibshirani 1996; Zhang 2010; Zou and Hastie 2005). By adding a penalty function to the likelihood criterion, the penalized likelihood method produces sparse solutions that eliminate the irrelevant predictors from the regression model using the zero-estimates.

In this paper, we are interested in selecting the k most important predictors out of p candidates, called best subset selection (Hocking & Leslie 1967). It is well known that best subset selection involves non-convex optimization problems that are computationally intractable in high-dimensional settings. Although some penalized likelihood approaches such as Lasso, elastic net, and MCP provide a convex surrogate for non-convex optimization, their applicability to best subset selection is still limited (Bertsimas, King, & Mazumder 2016). Meanwhile, a Bayesian approach to best subset selection, called Bayesian subset regression (BSR), has been proposed by Liang, Song, and Yu (2013). Using an adaptive Markov chain Monte Carlo (MCMC) algorithm, called the stochastic approximation Monte Carlo (Liang, Liu, & Carroll 2007), BSR finds the best subset by performing a global search over the entire model space. However, the global stochastic search with a large number of candidate predictors often raises computational challenges including heavy computation and slow convergence. To overcome this limitation, we introduce new Bayesian subset selection algorithms that quickly identify the best subset via hybrid algorithms that combine deterministic local search and stochastic global search in a Bayesian framework. The main attractive feature of our proposed method is that evaluating all possible candidate models for the next update, which is the most expensive part of Bayesian computation, is simultaneously accomplished in a single computation.

2 | BASIC SETUP

Consider a multiple linear regression model,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^\top$ is the n-dimensional response vector, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ is the $n \times p$ design matrix with $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^\top$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is a p-dimensional coefficient vector, and $\boldsymbol{\epsilon} \sim \text{Normal}(\mathbf{0}_n, \sigma^2\mathbf{I}_n)$. As we are interested in high-dimensional regression settings, we assume that $p > n$ and $\boldsymbol{\beta}$ contains many zero elements, i.e., $\boldsymbol{\beta}$ is a high-dimensional sparse vector. We further assume that the response and predictors are standardized so that the intercept is always excluded from our regression analysis. In this paper, our goal is to identify the k most important predictors in (1), where the best subset size, k, can be considered as being either fixed or varying. To formulate a Bayesian framework for best subset selection, let $\boldsymbol{\gamma} = \{j : \beta_j \neq 0\}$ be an index set of the active predictors. Given $\boldsymbol{\gamma}$, the full model (1) reduces to

$$\mathbf{y} = \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}} + \boldsymbol{\epsilon},$$

where $\mathbf{X}_{\boldsymbol{\gamma}}$ and $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ are a sub-matrix of $\mathbf{X}$ and a sub-vector of $\boldsymbol{\beta}$ that are determined by $\boldsymbol{\gamma}$, respectively. For algebraic and computational convenience, given $\boldsymbol{\gamma}$, we consider conjugate priors for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and σ^2 as follows:

$$\begin{aligned} \boldsymbol{\beta}_{\boldsymbol{\gamma}} | \sigma^2, \boldsymbol{\gamma} &\sim \text{Normal}(0, \tau\sigma^2\mathbf{I}_{|\boldsymbol{\gamma}|}), \\ \sigma^2 &\sim \text{Inverse-Gamma}(a_\sigma/2, b_\sigma/2), \end{aligned}$$

where τ, a_σ, and b_σ are hyperparameters and $|\boldsymbol{\gamma}|$ denotes the number of elements in the set $\boldsymbol{\gamma}$. To impose the constraint $|\boldsymbol{\gamma}| = k$, we define the prior distribution of $\boldsymbol{\gamma}$ by $\pi(\boldsymbol{\gamma}) \propto \mathbb{I}(|\boldsymbol{\gamma}| = k)$, where $\mathbb{I}(\cdot)$ is an indicator function. Let $m(\mathbf{y}|\boldsymbol{\gamma})$ be the marginal likelihood given $\boldsymbol{\gamma}$. Using the kernels of normal density and inverse gamma density, the marginal likelihood can be easily calculated as

$$m(\mathbf{y}|\boldsymbol{\gamma}) = \int f(\mathbf{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2)\pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\sigma^2, \boldsymbol{\gamma})\pi(\sigma^2) d\boldsymbol{\beta}_{\boldsymbol{\gamma}} d\sigma^2 \propto \frac{(\tau^{-1})^{\frac{|\boldsymbol{\gamma}|}{2}}}{|\mathbf{X}_{\boldsymbol{\gamma}}^\top\mathbf{X}_{\boldsymbol{\gamma}} + \tau^{-1}\mathbf{I}_{|\boldsymbol{\gamma}|}|^{\frac{1}{2}} \left(\mathbf{y}^\top\mathbf{H}_{\boldsymbol{\gamma}}\mathbf{y} + b_\sigma\right)^{\frac{a_\sigma+n}{2}}}, \tag{2}$$

where $f(\mathbf{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2)$ indicates the reduced model likelihood given $\boldsymbol{\gamma}$ and $\mathbf{H}_{\boldsymbol{\gamma}} = \mathbf{I}_n - \mathbf{X}_{\boldsymbol{\gamma}}(\mathbf{X}_{\boldsymbol{\gamma}}^\top\mathbf{X}_{\boldsymbol{\gamma}} + \tau^{-1}\mathbf{I}_{|\boldsymbol{\gamma}|})^{-1}\mathbf{X}_{\boldsymbol{\gamma}}^\top$. By Bayes' theorem, the posterior model probability of $\boldsymbol{\gamma}$ is proportional to $\pi(\boldsymbol{\gamma}|\mathbf{y}) \propto m(\mathbf{y}|\boldsymbol{\gamma})\pi(\boldsymbol{\gamma})$. Therefore, our Bayesian best subset selection can be performed by maximizing $m(\mathbf{y}|\boldsymbol{\gamma})$ over $\boldsymbol{\gamma}$ subject to the constraint $|\boldsymbol{\gamma}| = k$. Keep in mind that high-dimensional and non-convex optimization problems arise in our framework.

3 | BEST SUBSET SELECTION WITH A FIXED-SIZE

In this section, we develop a new Bayesian approach to best subset selection with a fixed subset size, k. Let $\hat{\boldsymbol{\gamma}}$ be the current estimate of the best model of size k. Our strategy is to update $\hat{\boldsymbol{\gamma}}$ by searching over neighbor models iteratively. To this end, for a given $\boldsymbol{\gamma}$, define $\mathcal{N}_+(\boldsymbol{\gamma}) = \{\boldsymbol{\gamma} \cup \{j\} : j \notin \boldsymbol{\gamma}\}$, which represents the set of larger neighbors of $\boldsymbol{\gamma}$ obtained by adding a new index to $\boldsymbol{\gamma}$. Similarly, define $\mathcal{N}_-(\boldsymbol{\gamma}) = \{\boldsymbol{\gamma} \setminus \{j\} : j \in \boldsymbol{\gamma}\}$ to be the set of smaller neighbors of $\boldsymbol{\gamma}$ obtained by deleting an index from $\boldsymbol{\gamma}$.

We introduce a deterministic search algorithm in Algorithm 1 that converges to a local maximum of $m(\mathbf{y}|\boldsymbol{\gamma})$ subject to $|\boldsymbol{\gamma}| = k$.

Algorithm 1 Deterministic best subset search with a fixed k

1. Define the initial model of size k, $\hat{\boldsymbol{\gamma}}$, and set $\hat{\boldsymbol{\gamma}}^{(0)} = \hat{\boldsymbol{\gamma}}$.
2. **Repeat** for $t = 1, 2, \ldots,$
   a) Compute $\tilde{\boldsymbol{\gamma}}^{(t)} = \arg\max_{\boldsymbol{\gamma} \in \mathcal{N}_+(\hat{\boldsymbol{\gamma}}^{(t-1)})} m(\mathbf{y}|\boldsymbol{\gamma})$;
   b) Compute $\hat{\boldsymbol{\gamma}}^{(t)} = \arg\max_{\boldsymbol{\gamma} \in \mathcal{N}_-(\tilde{\boldsymbol{\gamma}}^{(t)})} m(\mathbf{y}|\boldsymbol{\gamma})$;
   c) Update $\hat{\boldsymbol{\gamma}} \leftarrow \hat{\boldsymbol{\gamma}}^{(t)}$;

   until $\hat{\boldsymbol{\gamma}}^{(t-1)} = \hat{\boldsymbol{\gamma}}^{(t)}$.
3. Return $\hat{\boldsymbol{\gamma}}$.

The following theorem proves the convergence of the proposed deterministic search algorithm.

Theorem 1. The deterministic search in Algorithm 1 monotonically increases the objective function, $\mathsf{m}(\mathbf{y}|\boldsymbol{\gamma})$, subject to $|\boldsymbol{\gamma}| = \mathsf{k}$. In addition, the algorithm terminates in a finite number of iterations.

Proof of Theorem 1. Let $\hat{\boldsymbol{\gamma}}^{(\mathsf{t})}$ be the best subset of size k updated by the t^{th} iteration. Then, $\hat{\boldsymbol{\gamma}}^{(\mathsf{t}+1)} = \arg\max_{\boldsymbol{\gamma}\in\mathcal{N}_-(\tilde{\boldsymbol{\gamma}}^{(\mathsf{t})})} \mathsf{m}(\mathbf{y}|\boldsymbol{\gamma})$, where $\tilde{\boldsymbol{\gamma}}^{(\mathsf{t})} = \arg\max_{\boldsymbol{\gamma}\in\mathcal{N}_+(\hat{\boldsymbol{\gamma}}^{(\mathsf{t})})} \mathsf{m}(\mathbf{y}|\boldsymbol{\gamma})$. Since $\hat{\boldsymbol{\gamma}}^{(\mathsf{t})}$ also belongs to $\mathcal{N}_-(\tilde{\boldsymbol{\gamma}}^{(\mathsf{t})})$, we thus have $\mathsf{m}(\mathbf{y}|\hat{\boldsymbol{\gamma}}^{(\mathsf{t})}) \leq \mathsf{m}(\mathbf{y}|\hat{\boldsymbol{\gamma}}^{(\mathsf{t}+1)})$, which proves our first statement. Since the number of all possible state of $\boldsymbol{\gamma}$ satisfying $|\boldsymbol{\gamma}| = \mathsf{k}$ is finite, the algorithm terminates in a finite number of iterations. This completes our proof. □

Although the deterministic search algorithm converges quickly, a possible drawback is that the algorithm can get trapped in a local optimum. As an alternative, we introduce a stochastic search algorithm in Algorithm 2 that converges to a global maximum of $\mathsf{m}(\mathbf{y}|\boldsymbol{\gamma})$ with the constraint $|\boldsymbol{\gamma}| = \mathsf{k}$.

Algorithm 2 Stochastic best subset search with a fixed k

1. Define the initial model of size k, $\hat{\boldsymbol{\gamma}}$, and set $\hat{\boldsymbol{\gamma}}^{(0)} = \hat{\boldsymbol{\gamma}}$.
2. **Repeat** for $\mathsf{t} = 1, \ldots, \mathsf{T}$:
   a) Sample $\tilde{\boldsymbol{\gamma}}^{(\mathsf{t})}$ from $\mathcal{N}_+(\hat{\boldsymbol{\gamma}}^{(\mathsf{t}-1)})$ with probabilities proportional to $\mathsf{m}(\mathbf{y}|\boldsymbol{\gamma})\mathbb{I}\{\boldsymbol{\gamma} \in \mathcal{N}_+(\hat{\boldsymbol{\gamma}}^{(\mathsf{t}-1)})\}$;
   b) Sample $\hat{\boldsymbol{\gamma}}^{(\mathsf{t})}$ from $\mathcal{N}_-(\tilde{\boldsymbol{\gamma}}^{(\mathsf{t})})$ with probabilities proportional to $\mathsf{m}(\mathbf{y}|\boldsymbol{\gamma})\mathbb{I}\{\boldsymbol{\gamma} \in \mathcal{N}_-(\tilde{\boldsymbol{\gamma}}^{(\mathsf{t})})\}$;
   c) **If** $\mathsf{m}(\mathbf{y}|\hat{\boldsymbol{\gamma}}) < \mathsf{m}(\mathbf{y}|\hat{\boldsymbol{\gamma}}^{(\mathsf{t})})$, **then** update $\hat{\boldsymbol{\gamma}} \leftarrow \hat{\boldsymbol{\gamma}}^{(\mathsf{t})}$.
3. Return $\hat{\boldsymbol{\gamma}}$.

Note that the proposed stochastic search algorithm generates a Markov chain, $\{\hat{\boldsymbol{\gamma}}^{(\mathsf{t})}, \mathsf{t} = 1, \ldots, \mathsf{T}\}$, with the state space $\{\boldsymbol{\gamma} : |\boldsymbol{\gamma}| = \mathsf{k}\}$. Hence, if the current estimate $\hat{\boldsymbol{\gamma}}$ is not the global maximum, then we must observe that $\mathsf{m}(\mathbf{y}|\hat{\boldsymbol{\gamma}}) < \mathsf{m}(\mathbf{y}|\hat{\boldsymbol{\gamma}}^{(\mathsf{t})})$ with probability one as $\mathsf{T} \to \infty$. Therefore, as we iterate the algorithm, $\hat{\boldsymbol{\gamma}}$ converges to the global maximum.

However, while the stochastic search algorithm eventually reaches the global optimum, it requires the large number of iterations, which is computationally inefficient. To develop a computationally efficient global optimization algorithm, we propose a hybrid best subset selection algorithm that combines the stochastic global search algorithm and the deterministic local search algorithm into one. The proposed hybrid search algorithm is shown in Algorithm 3. In the proposed hybrid search algorithm, the deterministic search is first used to find a local optimum in an efficient manner. Then, the stochastic search is employed to check whether or not the deterministic search algorithm has reached the global optimum. To improve the performance of stochastic search, we introduce the tuning parameter $\alpha \in (0, 1]$ which acts as a precision parameter. As $\alpha \to 0$, the distribution will be more spread out so that the chance of getting stuck in the local maximum will be reduced in the stochastic search step. In our simulation study and real data analysis, we set $\alpha = \min\{1, \log(2)/\log(\mathsf{m}_{(1)}/\mathsf{m}_{(2)})\}$, where $\mathsf{m}_{(1)}$ and $\mathsf{m}_{(2)}$ denote the first and second largest values of $\{\mathsf{m}(\mathbf{y}|\boldsymbol{\gamma}) : \boldsymbol{\gamma} \in \mathcal{N}_+(\hat{\boldsymbol{\gamma}})\}$. In general, calculating marginal likelihoods for many candidate models, which is a necessary step in Bayesian model selection, is computationally expensive, even if an explicit form is available as in (2). The great merit of the proposed method is that evaluating the marginal likelihoods of all the candidates for the next update can be done *within a single computation*. To this end, let $\hat{\boldsymbol{\gamma}}$ be an index set of a model of size k and $\tilde{\boldsymbol{\gamma}}$ be an index set of a model of size $\mathsf{k}+1$. For any $\mathsf{i} \notin \hat{\boldsymbol{\gamma}}$, it can be shown that

$$\mathsf{m}(\mathbf{y}|\hat{\boldsymbol{\gamma}} \cup \{\mathsf{i}\}) \propto \left\{ \mathbf{y}^\top \mathbf{H}_{\hat{\boldsymbol{\gamma}}} \mathbf{y} - \frac{(\mathbf{x}_\mathsf{i}^\top \mathbf{H}_{\hat{\boldsymbol{\gamma}}} \mathbf{y})^2}{\tau^{-1} + \mathbf{x}_\mathsf{i}^\top \mathbf{H}_{\hat{\boldsymbol{\gamma}}} \mathbf{x}_\mathsf{i}} + \mathsf{b}_\sigma \right\}^{-\frac{\mathsf{a}_\sigma + \mathsf{n}}{2}} (\tau^{-1} + \mathbf{x}_\mathsf{i}^\top \mathbf{H}_{\hat{\boldsymbol{\gamma}}} \mathbf{x}_\mathsf{i})^{-1/2}, \tag{3}$$

where $\mathbf{x}_\mathsf{i}$ is the i^{th} column of $\mathbf{X}$. The proof of (3) is given in Appendix A. Note that $\mathbf{x}_\mathsf{i}^\top \mathbf{H}_{\hat{\boldsymbol{\gamma}}} \mathbf{x}_\mathsf{i}$ is the i^{th} diagonal element of $\mathbf{X}^\top \mathbf{H}_{\hat{\boldsymbol{\gamma}}} \mathbf{X}$ and that $\mathbf{x}_\mathsf{i}^\top \mathbf{H}_{\hat{\boldsymbol{\gamma}}} \mathbf{y}$ is the i^{th} element of $\mathbf{X}^\top \mathbf{H}_{\hat{\boldsymbol{\gamma}}} \mathbf{y}$. This implies that Eq. (3) is the i^{th} element of the following p-dimensional vector:

$$\mathbf{m}_+(\hat{\boldsymbol{\gamma}}) = \left\{ (\mathbf{y}^\top \mathbf{H}_{\hat{\boldsymbol{\gamma}}} \mathbf{y} + \mathsf{b}_\sigma)\mathbf{1}_\mathsf{p} - \frac{(\mathbf{X}^\top \mathbf{H}_{\hat{\boldsymbol{\gamma}}} \mathbf{y})^2}{\tau^{-1}\mathbf{1}_\mathsf{p} + \mathrm{diag}(\mathbf{X}^\top \mathbf{H}_{\hat{\boldsymbol{\gamma}}} \mathbf{X})} \right\}^{-\frac{\mathsf{a}_\sigma + \mathsf{n}}{2}} \left\{ \tau^{-1}\mathbf{1}_\mathsf{p} + \mathrm{diag}(\mathbf{X}^\top \mathbf{H}_{\hat{\boldsymbol{\gamma}}} \mathbf{X}) \right\}^{-1/2},$$

where $\mathbf{a}^\mathsf{x}$ and $\mathbf{a}/\mathbf{b}$ denote entrywise operations for generic vectors $\mathbf{a}$ and $\mathbf{b}$, accordingly, i.e., $\mathbf{a}^\mathsf{x} = (\mathsf{a}_1^\mathsf{x}, \ldots, \mathsf{a}_\mathsf{p}^\mathsf{x})$ and $\mathbf{a}/\mathbf{b} = (\mathsf{a}_1/\mathsf{b}_1, \ldots, \mathsf{a}_\mathsf{p}/\mathsf{b}_\mathsf{p})$. Since $\mathcal{N}_+(\hat{\boldsymbol{\gamma}}) = \{\hat{\boldsymbol{\gamma}} \cup \{\mathsf{i}\} : \mathsf{i} \notin \hat{\boldsymbol{\gamma}}\}$, evaluating $\mathsf{m}(\mathbf{y}|\boldsymbol{\gamma})$ for all $\boldsymbol{\gamma} \in \mathcal{N}_+(\hat{\boldsymbol{\gamma}})$ can be done simultaneously in a single computation by obtaining the

Algorithm 3 Hybrid best subset search with a fixed k

1. Define the initial model of size k, $\hat{\gamma}$
2. Set $\hat{\gamma}^{(0)} = \hat{\gamma}$.
3. **Repeat** for $t = 1, 2, \ldots$: `#deterministic search step`
   a) Compute $\tilde{\gamma}^{(t)} = \arg\max_{\gamma \in \mathcal{N}_+(\hat{\gamma}^{(t-1)})} m(\mathbf{y}|\gamma)$;
   b) Compute $\hat{\gamma}^{(t)} = \arg\max_{\gamma \in \mathcal{N}_-(\tilde{\gamma}^{(t)})} m(\mathbf{y}|\gamma)$;
   c) Update $\hat{\gamma} \leftarrow \hat{\gamma}^{(t)}$;

   until $\hat{\gamma}^{(t-1)} = \hat{\gamma}^{(t)}$.
4. Set $\hat{\gamma}^{(0)} = \hat{\gamma}$ and choose $\alpha \in (0, 1]$.
5. **Repeat** for $t = 1, \ldots, T$: `#stochastic search step`
   a) Sample $\tilde{\gamma}^{(t)}$ from $\mathcal{N}_+(\hat{\gamma}^{(t-1)})$ with probabilities proportional to $m(\mathbf{y}|\gamma)^\alpha \mathbb{I}\{\gamma \in \mathcal{N}_+(\hat{\gamma}^{(t-1)})\}$;
   b) Sample $\hat{\gamma}^{(t)}$ from $\mathcal{N}_-(\tilde{\gamma}^{(t)})$ with probabilities proportional to $m(\mathbf{y}|\gamma)^\alpha \mathbb{I}\{\gamma \in \mathcal{N}_-(\tilde{\gamma}^{(t)})\}$;
   c) **If** $m(\mathbf{y}|\hat{\gamma}) < m(\mathbf{y}|\hat{\gamma}^{(t)})$, **then** update $\hat{\gamma} \leftarrow \hat{\gamma}^{(t)}$, break the loop, and go to Step 2;
6. Return $\hat{\gamma}$.

sub-vector of $\mathbf{m}_+(\tilde{\gamma})$ for $\{i : i \notin \tilde{\gamma}\}$. Similarly, for any $j \in \tilde{\gamma}$, we can show that

$$m(\mathbf{y}|\tilde{\gamma} \setminus \{j\}) \propto \left\{ \mathbf{y}^\top \mathbf{H}_{\tilde{\gamma}} \mathbf{y} + \frac{(\mathbf{x}_j^\top \mathbf{H}_{\tilde{\gamma}} \mathbf{y})^2}{\tau^{-1} - \mathbf{x}_j^\top \mathbf{H}_{\tilde{\gamma}} \mathbf{x}_j} + b_\sigma \right\}^{-\frac{a_\sigma + n}{2}} (\tau^{-1} - \mathbf{x}_j^\top \mathbf{H}_{\tilde{\gamma}} \mathbf{x}_j)^{-1/2}. \tag{4}$$

The proof of (4) is given in Appendix B. Define

$$\mathbf{m}_-(\tilde{\gamma}) = \left\{ (\mathbf{y}^\top \mathbf{H}_{\tilde{\gamma}} \mathbf{y} + b_\sigma)\mathbf{1}_p + \frac{(\mathbf{X}^\top \mathbf{H}_{\tilde{\gamma}} \mathbf{y})^2}{\tau^{-1}\mathbf{1}_p - \mathrm{diag}(\mathbf{X}^\top \mathbf{H}_{\tilde{\gamma}} \mathbf{X})} \right\}^{-\frac{a_\sigma + n}{2}} \left\{ \tau^{-1}\mathbf{1}_p - \mathrm{diag}(\mathbf{X}^\top \mathbf{H}_{\tilde{\gamma}} \mathbf{X}) \right\}^{-1/2}.$$

It is easy to check that Eq. (4) is the j^{th} element of $\mathbf{m}_-(\tilde{\gamma})$. Since $\mathcal{N}_-(\tilde{\gamma}) = \{\tilde{\gamma} \setminus \{j\} : j \in \tilde{\gamma}\}$, evaluating $m(\mathbf{y}|\gamma)$ for all $\gamma \in \mathcal{N}_-(\tilde{\gamma})$ can be done simultaneously in a single computation by obtaining the sub-vector of $\mathbf{m}_-(\tilde{\gamma})$ for $\{j : j \in \tilde{\gamma}\}$. Using the aforementioned strategy, we can easily and quickly implement steps 3a), 3b), 5a), and 5b) of Algorithm 3. It is also worth noting that the proposed approach enables us to avoid multiple computations of inverse and determinant of matrices in Eq. (2).

4 | BEST SUBSET SELECTION WITH VARYING SIZES

In this section, we extend the proposed method to best subset selection for varying k with a prespecified upper bound, say K, which is a common setting for high-dimensional best subset selection (e.g., Bertsimas et al. 2016; Liang et al. 2013). In a Bayesian framework, this extension can be easily done by assigning an appropriate prior for unknown k. As a non-informative prior, one may consider a discrete uniform prior for k, that is, $k \sim \text{Uniform}\{1, \ldots, K\}$. However, the uniform prior tends to assign larger probability to a larger subset due to the fact that the total number of subsets of size k is $\binom{p}{k}$, which tends to increase exponentially as k increase. To resolve this issue, using a similar idea of Chen and Chen (2008), we consider

$$\pi(k) \propto \mathbb{I}(k \leq K) / \binom{p}{k}.$$

From Bayes' theorem, Bayesian best subset selection is then performed by maximizing

$$\pi(\gamma|\mathbf{y}) \propto m(\mathbf{y}|\gamma)\pi(k) \propto m(\mathbf{y}|\gamma)\mathbb{I}(|\gamma| \leq K) / \binom{p}{|\gamma|}. \tag{5}$$

Note that (5) is equivalent to maximizing

$$m(\mathbf{y}|\boldsymbol{\gamma})\mathbb{I}(|\boldsymbol{\gamma}| = k)/\binom{p}{k} \quad \text{subject to } k \leq K.$$

Hence, the optimization problem in Eq. (5) can be solved by proceeding the following steps:

1. For $k = 1, \ldots, K$, compute $\hat{\boldsymbol{\gamma}}_k = \arg\max_{\boldsymbol{\gamma}} \{m(\mathbf{y}|\boldsymbol{\gamma})\mathbb{I}(|\boldsymbol{\gamma}| = k)\}$ using Algorithm 3.
2. Compute $\hat{k} = \arg\max_{1 \leq k \leq K} \left\{\log m(\mathbf{y}|\hat{\boldsymbol{\gamma}}_k) - \log \binom{p}{k}\right\}$.
3. Return $\hat{\boldsymbol{\gamma}} = \hat{\boldsymbol{\gamma}}_{\hat{k}}$.

Note that if parallel computation is available, it can be applied to the first step of the above procedure. The following theorem shows that the proposed model selection approach achieves the model selection consistency in the high-dimensional setting that $p > n$.

Theorem 2. Let $\boldsymbol{\gamma}_*$ indicate the true model. Define $\Gamma = \{\boldsymbol{\gamma} : |\boldsymbol{\gamma}| \leq K, \boldsymbol{\gamma} \neq \boldsymbol{\gamma}_*\}$. Assume that $p = O(n^\xi)$ for $\xi \geq 1$. Under the asymptotic identifiability condition of Chen and Chen (2008), if $\tau \to \infty$ but $\tau = o(n)$, then the proposed Bayesian subset selection possesses the Bayesian model selection consistency, that is,

$$\pi(\boldsymbol{\gamma}_*|\mathbf{y}) > \max_{\boldsymbol{\gamma} \in \Gamma} \pi(\boldsymbol{\gamma}|\mathbf{y}) \tag{6}$$

in probability as $n \to \infty$.

Proof of Theorem 2. From the Laplace approximation of Kass and Raftery (1995), we have

$$\log m(\mathbf{y}|\boldsymbol{\gamma}) = \log f(\mathbf{y}|\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}, \hat{\sigma}^2) - \frac{|\boldsymbol{\gamma}|}{2}\log(n) + o_p(\log n),$$

where $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} = (\mathbf{X}_{\boldsymbol{\gamma}}^\top \mathbf{X}_{\boldsymbol{\gamma}})^{-1}\mathbf{X}_{\boldsymbol{\gamma}}^\top \mathbf{y}$ and $\hat{\sigma}^2 = \|\mathbf{H}_{\boldsymbol{\gamma}}\mathbf{y}\|^2/n$. Ignoring a smaller order term than $\log n$, our posterior criterion (5) can be approximated as

$$-2\log \pi(\boldsymbol{\gamma}|\mathbf{y}) \approx -2\log f(\mathbf{y}|\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}, \hat{\sigma}^2) + |\boldsymbol{\gamma}|\log(n) + 2\log\binom{p}{k}, \tag{7}$$

which is equivalent to the extended Bayesian information criterion (Chen & Chen 2008). Therefore, it follows from Theorem 1 of Chen and Chen (2008) that Eq. (6) holds in probability. □

5 | SIMULATION STUDY

In this section, we investigate the variable selection performance of our best subset selection algorithm on simulated high-dimensional data. For $n = 100$, we generate the data $\{(y_i, x_{i1}, \ldots, x_{ip}) : i = 1, \ldots, n\}$ from the following linear regression model:

$$y_i \overset{\text{ind}}{\sim} \text{Normal}\left(\sum_{j=1}^{p} \beta_j x_{ij}, 1\right),$$

where $(x_{i1}, \ldots, x_{ip})^\top \overset{\text{iid}}{\sim} \text{Normal}(\mathbf{0}_p, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = (\Sigma_{ij})_{p\times p}$ and $\Sigma_{ij} = \rho^{|i-j|}$, four β_j's are randomly selected and then generated from Uniform$\{-1, -2, 1, 2\}$ independently, and the remaining β-coefficients are set equal to 0.

We consider four different scenarios: (i) $p = 200$, $\rho = 0.1$, (ii) $p = 200$, $\rho = 0.9$, (iii) $p = 1000$, $\rho = 0.1$, and (iv) $p = 1000$, $\rho = 0.9$. We assume that there is no prior information. Hence, to make our prior distributions non-informative, we set $a_\sigma = b_\sigma = 1$ and $\tau = (\log p)^2$, which satisfies the sufficient condition for model selection consistency discussed in Theorem 2. We further assume that the true number of active predictors is unknown. Hence, we employ the proposed method with unknown k in Section 4 with $K = \lceil n^{2/3} \rceil = 22$. For the hybrid search algorithm, we set $T = 100$, which is much smaller compared to existing stochastic search algorithms (e.g., George and McCulloch 1993; Hans, Dobra, and West 2007; Kirkpatrick, Gelatt, and Vecchi 1983), and marginal correlations between the response and each predictor are used to define the initial value of $\hat{\boldsymbol{\gamma}}$. For comparison purposes, we also employ the most popular penalized likelihood methods, LASSO (Tibshirani 1996), the elastic net (Zou & Hastie 2005), SCAD (Fan & Li 2001), and MCP (Zhang 2010), where the regularization parameters or tuning parameters are determined by the extended BIC in (7) to achieve a fair comparison with the proposed method. To evaluate the variable selection performance, we calculate false discovery rate, percentage of selecting the exact true model, average size of the selected model, and mean of Hamming distance based on 2,000 Monte Carlo replications. The results are shown in Table 1. For every scenario, the proposed method outperforms all the penalized likelihood methods. The proposed method always achieves the smallest false discovery rate and this implies that the proposed method provides the minimum proportion of incorrectly identifying the true active predictors as inactive. According to the selected model size and the Hamming distance, we argue that the proposed method tends to select the closest model to the true model. In addition, the proposed method selects the exact true model with high probability, and this finite sample performance supports our theoretical result in Theorem 2.

6 | REAL DATA APPLICATION

In this section, we apply our proposed methods to Breast invasive carcinoma (BRCA) data generated by The Cancer Genome Atlas (TCGA) Research Network: `http://cancergenome.nih.gov`. We download BRCA data using `R` package "`curatedTCGAData`". The data set contains $17,814$ gene expression measurements (recorded on the log scale) of 526 patients with primary solid tumor in TCGA project. BRCA1 is a well-known tumor suppressor gene and its mutations predispose women to breast cancer (Findlay et al. 2018). Our goal here is to identify the best fitting model for estimating an association between BRCA1 (response variable) and the other genes (independent variables). After removing missing values, the data set reduces to $n = 526$ samples with $17,323$ genes.

As a pre-screening procedure, we first select the top $p = 5,000$ genes that are marginally correlated with BRCA1. Then, the proposed method and the penalized likelihood methods as in Section 5 are applied to the reduced data. To assess model fitting, for each method, we compute BIC (Schwarz 1978), extended BIC, and mean squared prediction error (MSPE) using the ordinary least square (OLS) estimate with the selected predictors, where MSPE is estimated by Monte Carlo cross-validation over 500 replications using 70% of training set and 30% of testing set.

Table 2 summarizes model comparison results. As both BIC and extended BIC are minimized at the resulting model from the proposed method, this implies that our Bayesian method is most strongly supported by the data. In addition, the proposed method has the smallest MSPE. Hence, the results demonstrate that the proposed method provides the best fit to the data. The heatmap in Figure 1 displays the OLS coefficient estimates and the p-values of the selected predictors for each method. As a result, the proposed method identifies 8 genes that are statistically related to the human tumor suppressor gene, BRCA1. According to the human gene database, called GeneCards, except for C10orf76, 7 among the 8 genes are associated with diseases including Myasthenic Syndrome, Pancreatic Cancer, Kenny-Caffey Syndrome, and Mental Retardation. The GeneCards database is publicly available at `https://www.genecards.org`.

7 | DISCUSSION

The proposed hybrid search approach is computationally attractive compared with traditional stochastic search algorithms that are commonly used in Bayesian variable selection. As mentioned in Section 4, parallel computing, which executes many calculations simultaneously using multiple nodes or multiple processors, can be employed in the proposed framework. Our hybrid search algorithm can be also extended to multivariate regression and binary regression in a Bayesian framework. Such works are in progress by the authors.

References

Bertsimas, D., King, A., & Mazumder, R. (2016). Best subset selection via a modern optimization lens. *The Annals of Statistics*, *44*(2), 813–852.

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

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

Findlay, G. M., Daza, R. M., Martin, B., Zhang, M. D., Leith, A. P., Gasperini, M., … Shendure, J. (2018). Accurate classification of brca1 variants with saturation genome editing. *Nature*, *562*(7726), 217.

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

Hans, C., Dobra, A., & West, M. (2007). Shotgun stochastic search for "large p" regression. *Journal of the American Statistical Association*, *102*(478), 507–516.

Hocking, R. R., & Leslie, R. N. (1967). Selection of the best subset in regression analysis. *Technometrics*, *9*(4), 531–540.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *science*, *220*(4598), 671–680.

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

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

Schwarz, G. (1978, 03). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. doi: 10.1214/aos/1176344136

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

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

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

How to cite this article: Shiqiang J. and G. Goh (20xx), Bayesian selection of best subsets in high-dimensional regression, *Stat*, *20xx;00:1–6.*

APPENDIX

A CALCULATION OF EQUATION (3)

For any $i \notin \hat{\gamma}$, from Eq. (2), we have

$$m(\mathbf{y}|\hat{\gamma} \cup \{i\}) \propto |\mathbf{X}_{\hat{\gamma}\cup\{i\}}^\top \mathbf{X}_{\hat{\gamma}\cup\{i\}} + \tau^{-1}\mathbf{I}_{k+1}|^{-1/2} \left(\mathbf{y}^\top \mathbf{H}_{\hat{\gamma}\cup\{i\}}\mathbf{y} + b_\sigma\right)^{-\frac{a_\sigma+n}{2}}. \tag{A1}$$

It follows from the Sherman-Morrison formula that

$$\mathbf{H}_{\hat{\gamma}\cup\{i\}} = \mathbf{H}_{\hat{\gamma}} - \frac{\mathbf{H}_{\hat{\gamma}}\mathbf{x}_i\mathbf{x}_i^\top\mathbf{H}_{\hat{\gamma}}}{\tau^{-1} + \mathbf{x}_i^\top\mathbf{H}_{\hat{\gamma}}\mathbf{x}_i}. \tag{A2}$$

Using the Sylvester's determinant identity and the Sherman-Morrison formula, we obtain

$$\begin{aligned}
|\mathbf{X}_{\hat{\gamma}\cup\{i\}}^\top \mathbf{X}_{\hat{\gamma}\cup\{i\}} + \tau^{-1}\mathbf{I}_{k+1}| &= \tau^{-(k+1)}|\tau\mathbf{X}_{\hat{\gamma}\cup\{i\}}\mathbf{X}_{\hat{\gamma}\cup\{i\}}^\top + \mathbf{I}_n| \\
&= \tau^{-(k+1)}|\tau\mathbf{X}_{\hat{\gamma}}\mathbf{X}_{\hat{\gamma}}^\top + \mathbf{I}_n + \tau\mathbf{x}_i\mathbf{x}_i^\top| \\
&= \tau^{-(k+1)}|\tau\mathbf{X}_{\hat{\gamma}}\mathbf{X}_{\hat{\gamma}}^\top + \mathbf{I}_n|\{1 + \tau\mathbf{x}_i^\top(\tau\mathbf{X}_{\hat{\gamma}}\mathbf{X}_{\hat{\gamma}}^\top + \mathbf{I}_n)^{-1}\mathbf{x}_i\} \\
&= |\mathbf{X}_{\hat{\gamma}}^\top\mathbf{X}_{\hat{\gamma}} + \tau^{-1}\mathbf{I}_k|\{\tau^{-1} + \mathbf{x}_i^\top\mathbf{H}_{\hat{\gamma}}\mathbf{x}_i\}.
\end{aligned} \tag{A3}$$

Applying (A2) and (A3) to (A1), we thus have

$$m(\mathbf{y}|\hat{\gamma}\cup\{i\}) \propto \left\{\mathbf{y}^\top\mathbf{H}_{\hat{\gamma}}\mathbf{y} - \frac{(\mathbf{x}_i^\top\mathbf{H}_{\hat{\gamma}}\mathbf{y})^2}{\tau^{-1} + \mathbf{x}_i^\top\mathbf{H}_{\hat{\gamma}}\mathbf{x}_i} + b_\sigma\right\}^{-\frac{a_\sigma+n}{2}} (\tau^{-1} + \mathbf{x}_i^\top\mathbf{H}_{\hat{\gamma}}\mathbf{x}_i)^{-1/2}$$

for any $i \notin \hat{\gamma}$.

B CALCULATION OF EQUATION (4)

For any $j \in \tilde{\gamma}$, Eq. (2) leads to

$$m(\mathbf{y}|\tilde{\gamma}\setminus\{j\}) \propto |\mathbf{X}_{\tilde{\gamma}\setminus\{j\}}^\top \mathbf{X}_{\tilde{\gamma}\setminus\{j\}} + \tau^{-1}\mathbf{I}_k|^{-1/2} \left(\mathbf{y}^\top \mathbf{H}_{\tilde{\gamma}\setminus\{j\}}\mathbf{y} + b_\sigma\right)^{-\frac{a_\sigma+n}{2}}. \tag{B4}$$

From the Sherman-Morrison formula, we have

$$\mathbf{H}_{\tilde{\gamma}\setminus\{j\}} = \mathbf{H}_{\tilde{\gamma}} + \frac{\mathbf{H}_{\tilde{\gamma}}\mathbf{x}_j\mathbf{x}_j^\top\mathbf{H}_{\tilde{\gamma}}}{\tau^{-1} - \mathbf{x}_j^\top\mathbf{H}_{\tilde{\gamma}}\mathbf{x}_j}. \tag{B5}$$

From the Sylvester's determinant identity and the Sherman-Morrison formula, we obtain

$$\begin{aligned}
|\mathbf{X}_{\tilde{\gamma}\setminus\{j\}}^\top \mathbf{X}_{\tilde{\gamma}\setminus\{j\}} + \tau^{-1}\mathbf{I}_k| &= \tau^{-k}|\tau\mathbf{X}_{\tilde{\gamma}\setminus\{j\}}\mathbf{X}_{\tilde{\gamma}\setminus\{j\}}^\top + \mathbf{I}_n| \\
&= \tau^{-k}|\tau\mathbf{X}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}}^\top + \mathbf{I}_n - \tau\mathbf{x}_j\mathbf{x}_j^\top| \\
&= \tau^{-k}|\tau\mathbf{X}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}}^\top + \mathbf{I}_n|\{1 - \tau\mathbf{x}_j^\top(\tau\mathbf{X}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}}^\top + \mathbf{I}_n)^{-1}\mathbf{x}_j\} \\
&= \tau^2|\mathbf{X}_{\tilde{\gamma}}^\top\mathbf{X}_{\tilde{\gamma}} + \tau^{-1}\mathbf{I}_{k+1}|\{\tau^{-1} - \mathbf{x}_j^\top\mathbf{H}_{\tilde{\gamma}}\mathbf{x}_j\}.
\end{aligned} \tag{B6}$$

Hence, applying (B5) and (B6) to (B4), we have

$$m(\mathbf{y}|\tilde{\gamma}\setminus\{j\}) \propto \left\{\mathbf{y}^\top\mathbf{H}_{\tilde{\gamma}}\mathbf{y} + \frac{(\mathbf{x}_j^\top\mathbf{H}_{\tilde{\gamma}}\mathbf{y})^2}{\tau^{-1} - \mathbf{x}_j^\top\mathbf{H}_{\tilde{\gamma}}\mathbf{x}_j} + b_\sigma\right\}^{-\frac{a_\sigma+n}{2}} (\tau^{-1} - \mathbf{x}_j^\top\mathbf{H}_{\tilde{\gamma}}\mathbf{x}_j)^{-1/2}$$

for any $j \in \tilde{\gamma}$.

TABLE 1 Simulation study results based on $2,000$ Monte Carlo replications for scenarios i–iv.

Scenario	Method	FDR (s.e.)	TRUE% (s.e.)	SIZE (s.e.)	HAM (s.e.)
i	Proposed	0.006 (0.001)	96.900 (0.388)	4.032 (0.004)	0.032 (0.004)
	SCAD	0.034 (0.002)	85.200 (0.794)	4.188 (0.011)	0.188 (0.011)
	MCP	0.035 (0.002)	84.750 (0.804)	4.191 (0.011)	0.191 (0.011)
	ENET	0.016 (0.001)	92.700 (0.582)	4.087 (0.007)	0.087 (0.007)
	LASSO	0.020 (0.002)	91.350 (0.629)	4.109 (0.009)	0.109 (0.009)
ii	Proposed	0.023 (0.002)	88.750 (0.707)	3.985 (0.006)	0.203 (0.014)
	SCAD	0.059 (0.003)	74.150 (0.979)	4.107 (0.015)	0.480 (0.022)
	MCP	0.137 (0.004)	55.400 (1.112)	4.264 (0.020)	1.098 (0.034)
	ENET	0.501 (0.004)	0.300 (0.122)	7.716 (0.072)	5.018 (0.052)
	LASSO	0.276 (0.004)	15.550 (0.811)	5.308 (0.033)	2.038 (0.034)
iii	Proposed	0.004 (0.001)	98.100 (0.305)	4.020 (0.003)	0.020 (0.003)
	SCAD	0.027 (0.002)	87.900 (0.729)	4.145 (0.010)	0.145 (0.010)
	MCP	0.031 (0.002)	86.550 (0.763)	4.172 (0.013)	0.172 (0.013)
	ENET	0.035 (0.002)	84.850 (0.802)	4.181 (0.013)	0.206 (0.012)
	LASSO	0.014 (0.001)	93.850 (0.537)	4.073 (0.007)	0.073 (0.007)
iv	Proposed	0.023 (0.002)	89.850 (0.675)	4.005 (0.005)	0.190 (0.013)
	SCAD	0.068 (0.003)	74.250 (0.978)	4.196 (0.014)	0.493 (0.023)
	MCP	0.152 (0.004)	53.750 (1.115)	4.226 (0.017)	1.202 (0.035)
	ENET	0.417 (0.005)	0.150 (0.087)	6.228 (0.068)	4.089 (0.043)
	LASSO	0.265 (0.004)	19.500 (0.886)	5.139 (0.029)	1.909 (0.035)

FDR: false discovery rate; TRUE%: percentage of the true model detected; SIZE : selected model size; HAM: Hamming distance; s.e.: standard error.

TABLE 2 Model comparison results for BRCA data

	BIC	extended BIC	MSPE
Proposed	984.45	1099.50	0.60
SCAD	1104.69	1166.47	0.68
MCP	1104.69	1166.47	0.68
ENET	1110.65	1186.25	0.68
LASSO	1104.69	1166.47	0.68

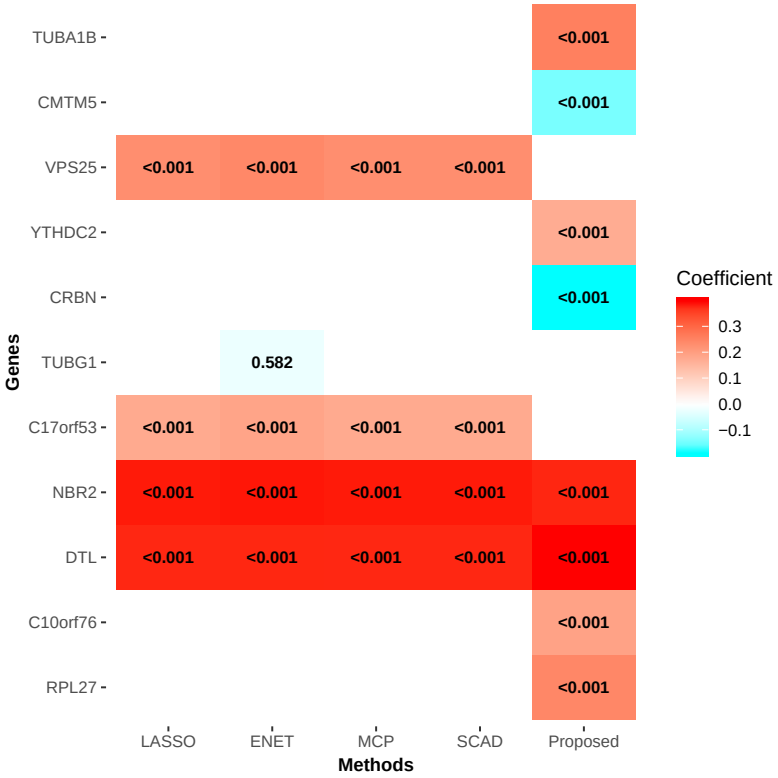

FIGURE 1 Heatmap of OLS coefficient estimates with p-values for the selected predictors.