**Computational Statistics manuscript No.**
(will be inserted by the editor)

# Bayesian selection of best subsets via hybrid search

**Shiqiang Jin · Gyuhyeong Goh**

Received: date / Accepted: date

**Abstract** Over the past decades, variable selection for 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 using a hybrid search algorithm that combines a deterministic local search and a stochastic global search. To reduce the computational cost of evaluating multiple candidate subsets for each update, we propose a novel strategy that enables us to calculate exact marginal likelihoods 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 analysis are conducted to investigate the performance of the proposed method.

**Keywords** Bayesian variable selection · Best subset selection · High-dimensional regression analysis · 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 where 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 a large 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., Tibshirani 1996; Fan and Li 2001; Zou and Hastie 2005; Zhang 2010). By adding a penalty function to the likelihood criterion, the penalized likelihood method produces sparse solutions that automatically 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 and Leslie 1967). It is well known that best subset selection involves nonconvex 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 nonconvex optimization, their applicability to best subset selection is still limited (Bertsimas et al. 2016). Meanwhile, a Bayesian approach to best subset selection, called Bayesian subset regression (BSR), has been proposed by Liang et al. (2013). Using an adaptive Markov chain Monte Carlo (MCMC) algorithm, called the stochastic approximation Monte Carlo (Liang et al. 2007), BSR finds the best subset by performing a stochastic search over the entire model space. However, the stochastic search with a large number of candidate predictors often raises computational challenges including extremely heavy computation and slow convergence. To overcome this limitation, we introduce a hybrid best subset search algorithm that combines a deterministic local search and a stochastic global search in a Bayesian framework.

Shiqiang Jin
Department of Statistics, Kansas State University, Manhattan, KS 66506, U.S.A.
E-mail: jinsq@ksu.edu

Gyuhyeong Goh
Department of Statistics, Kansas State University, Manhattan, KS 66506, U.S.A.
E-mail: ggoh@ksu.edu

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 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 $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top \sim \text{Normal}(\mathbf{0}_n, \sigma^2\mathbf{I}_n)$. Throughout this paper, we restrict our attention to high-dimensional regression problems and thus it is always assumed that $p > n$ and $\boldsymbol{\beta}$ contains many zero elements. We further assume that $\mathbf{y}$ and columns of $\mathbf{X}$ are standardized so that the intercept is 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 subvector 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 $\sigma^2$ 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 $\tau$, $a_\sigma$, and $b_\sigma$ 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 calculated as

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

where $f(\mathbf{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2)$ denotes the reduced model likelihood function given $\boldsymbol{\gamma}$ and $\mathbf{H}_{\boldsymbol{\gamma}} = \mathbf{I}_n - \mathbf{X}_{\boldsymbol{\gamma}}(\mathbf{X}_{\boldsymbol{\gamma}}^\top\mathbf{X}_{\boldsymbol{\gamma}} + \tau^{-1}\mathbf{I}_{|\boldsymbol{\gamma}|})^{-1}\mathbf{X}_{\boldsymbol{\gamma}}^\top$.

From Bayes' theorem, the posterior model probability of $\boldsymbol{\gamma}$ is proportional to $\pi(\boldsymbol{\gamma}|\mathbf{y}) \propto m(\mathbf{y}|\boldsymbol{\gamma})\pi(\boldsymbol{\gamma}) \propto m(\mathbf{y}|\boldsymbol{\gamma})\mathbb{I}(|\boldsymbol{\gamma}| = k)$. Therefore, 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 subset of size $k$. Our strategy is to update $\hat{\boldsymbol{\gamma}}$ by searching over neighbor models iteratively. To this end, for an index set $\boldsymbol{\gamma}$, define $\mathcal{N}_+(\boldsymbol{\gamma}) = \{\boldsymbol{\gamma} \cup \{i\} : i \notin \boldsymbol{\gamma}\}$, 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 $\hat{\gamma}$ to be an initial subset of size $k$; and set $\hat{\gamma}^{(0)} = \hat{\gamma}$.
2. Repeat for $t = 1, 2, \ldots,$
     a) Compute $\tilde{\gamma}^{(t)} = \arg\max_{\gamma \in \mathcal{N}_+(\hat{\gamma}^{(t-1)})} m(\mathbf{y}|\gamma)$;
        # i.e., select the locally best subset of size k+1
     b) Compute $\hat{\gamma}^{(t)} = \arg\max_{\gamma \in \mathcal{N}_-(\tilde{\gamma}^{(t)})} m(\mathbf{y}|\gamma)$;
        # i.e., select the locally best subset of size k
     c) Update $\hat{\gamma} \leftarrow \hat{\gamma}^{(t)}$;
   until $\hat{\gamma}^{(t-1)} = \hat{\gamma}^{(t)}$. # i.e., terminate immediately if no update is made
3. Return $\hat{\gamma}$.
```

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, $m(\mathbf{y}|\boldsymbol{\gamma})$, subject to $|\boldsymbol{\gamma}| = k$. In addition, the algorithm terminates in a finite number of iterations.*

*Proof (of Theorem 1)* Let $\hat{\boldsymbol{\gamma}}^{(t-1)}$ be the best subset of size $k$ updated by the $t-1$-th iteration. Then, $\hat{\boldsymbol{\gamma}}^{(t)} = \arg\max_{\boldsymbol{\gamma} \in \mathcal{N}_-(\tilde{\boldsymbol{\gamma}}^{(t)})} m(\mathbf{y}|\boldsymbol{\gamma})$, where

$$\tilde{\boldsymbol{\gamma}}^{(t)} = \arg\max_{\boldsymbol{\gamma} \in \mathcal{N}_+(\hat{\boldsymbol{\gamma}}^{(t-1)})} m(\mathbf{y}|\boldsymbol{\gamma}).$$

Since $\hat{\boldsymbol{\gamma}}^{(t-1)}$ also belongs to $\mathcal{N}_-(\tilde{\boldsymbol{\gamma}}^{(t)})$, we thus have $m(\mathbf{y}|\hat{\boldsymbol{\gamma}}^{(t-1)}) \leq m(\mathbf{y}|\hat{\boldsymbol{\gamma}}^{(t)})$, which proves our first statement. Since the number of all possible states of $\boldsymbol{\gamma}$ satisfying $|\boldsymbol{\gamma}| = k$ is finite, the algorithm terminates in a finite number of iterations. This completes our proof.

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

**Algorithm 2** Stochastic best subset search with a fixed $k$

```
1. Define $\hat{\boldsymbol{\gamma}}$ to be an initial subset of size $k$; and set $\hat{\boldsymbol{\gamma}}^{(0)} = \hat{\boldsymbol{\gamma}}$.
2. Repeat for $t = 1, \ldots, T$:
   a) Update $\tilde{\boldsymbol{\gamma}}^{(t)}$ by selecting one from $\mathcal{N}_+(\hat{\boldsymbol{\gamma}}^{(t-1)})$ with
      probability
         $\frac{m(\mathbf{y}|\boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma}' \in \mathcal{N}_+(\hat{\boldsymbol{\gamma}}^{(t-1)})} m(\mathbf{y}|\boldsymbol{\gamma}')}, \ \boldsymbol{\gamma} \in \mathcal{N}_+(\hat{\boldsymbol{\gamma}}^{(t-1)});$
   b) Update $\hat{\boldsymbol{\gamma}}^{(t)}$ by selecting one from $\mathcal{N}_-(\tilde{\boldsymbol{\gamma}}^{(t)})$ with
      probability
         $\frac{m(\mathbf{y}|\boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma}' \in \mathcal{N}_-(\tilde{\boldsymbol{\gamma}}^{(t)})} m(\mathbf{y}|\boldsymbol{\gamma}')}, \ \boldsymbol{\gamma} \in \mathcal{N}_-(\tilde{\boldsymbol{\gamma}}^{(t)});$
   c) If $m(\mathbf{y}|\hat{\boldsymbol{\gamma}}) < m(\mathbf{y}|\hat{\boldsymbol{\gamma}}^{(t)})$, then update $\hat{\boldsymbol{\gamma}} \leftarrow \hat{\boldsymbol{\gamma}}^{(t)}$.
3. Return $\hat{\boldsymbol{\gamma}}$.
```

Although 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 given 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. Note that if the stochastic search finds a better subset $\hat{\boldsymbol{\gamma}}^{(t)}$ than the current best subset $\hat{\boldsymbol{\gamma}}$ (i.e., when $m(\mathbf{y}|\hat{\boldsymbol{\gamma}}) < m(\mathbf{y}|\hat{\boldsymbol{\gamma}}^{(t)})$ occurs), then we immediately stop the stochastic search procedure and go back to the deterministic search step with the updated state of $\hat{\boldsymbol{\gamma}}$. To improve the performance of stochastic search, we introduce a 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. If we set $\alpha = 1$, the stochastic search step is analogous to Algorithm 2. In our simulation study and real data analysis, we set $\alpha = \min\{1, \log(2)/\log(m_{(1)}/m_{(2)})\}$, where $m_{(1)}$ and $m_{(2)}$ are, respectively, the first and second largest values of $\{m(\mathbf{y}|\boldsymbol{\gamma}) : \boldsymbol{\gamma} \in \mathcal{N}_+(\hat{\boldsymbol{\gamma}})\}$.

**Algorithm 3** Hybrid best subset search with a fixed $k$

```
1. Define an initial model of size $k$ as $\hat{\boldsymbol{\gamma}}$.
2. Set $\hat{\boldsymbol{\gamma}}^{(0)} = \hat{\boldsymbol{\gamma}}$.
# deterministic search step
3. 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 $\hat{\boldsymbol{\gamma}}^{(t-1)} = \hat{\boldsymbol{\gamma}}^{(t)}$.
4. Set $\hat{\boldsymbol{\gamma}}^{(0)} = \hat{\boldsymbol{\gamma}}$ and choose a tuning parameter $\alpha \in (0, 1]$.
# stochastic search step
5. Repeat for $t = 1, \ldots, T$:
   a) Update $\tilde{\boldsymbol{\gamma}}^{(t)}$ by drawing one from $\mathcal{N}_+(\hat{\boldsymbol{\gamma}}^{(t-1)})$ with
      probability
         $\frac{m(\mathbf{y}|\boldsymbol{\gamma})^\alpha}{\sum_{\boldsymbol{\gamma} \in \mathcal{N}_+(\hat{\boldsymbol{\gamma}}^{(t-1)})} m(\mathbf{y}|\boldsymbol{\gamma})^\alpha}, \ \boldsymbol{\gamma} \in \mathcal{N}_+(\hat{\boldsymbol{\gamma}}^{(t-1)});$
   b) Update $\hat{\boldsymbol{\gamma}}^{(t)}$ by drawing one from $\mathcal{N}_-(\tilde{\boldsymbol{\gamma}}^{(t)})$ with
      probability
         $\frac{m(\mathbf{y}|\boldsymbol{\gamma})^\alpha}{\sum_{\boldsymbol{\gamma}' \in \mathcal{N}_-(\tilde{\boldsymbol{\gamma}}^{(t)})} m(\mathbf{y}|\boldsymbol{\gamma}')^\alpha}, \ \boldsymbol{\gamma} \in \mathcal{N}_-(\tilde{\boldsymbol{\gamma}}^{(t)});$
   c) If $m(\mathbf{y}|\hat{\boldsymbol{\gamma}}) < m(\mathbf{y}|\hat{\boldsymbol{\gamma}}^{(t)})$, then update $\hat{\boldsymbol{\gamma}} \leftarrow \hat{\boldsymbol{\gamma}}^{(t)}$ and
      immediately go to Step 2.
6. Return $\hat{\boldsymbol{\gamma}}$.
```

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 simultaneously *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 $k+1$. For any $i \notin \hat{\boldsymbol{\gamma}}$, it can be shown that

$$
\begin{aligned}
& m(\mathbf{y}|\hat{\boldsymbol{\gamma}} \cup \{i\}) \\
& \propto \left\{ \mathbf{y}^\top \mathbf{H}_{\hat{\boldsymbol{\gamma}}} \mathbf{y} - \frac{(\mathbf{x}_i^\top \mathbf{H}_{\hat{\boldsymbol{\gamma}}} \mathbf{y})^2}{\tau^{-1} + \mathbf{x}_i^\top \mathbf{H}_{\hat{\boldsymbol{\gamma}}} \mathbf{x}_i} + b_\sigma \right\}^{-\frac{a_\sigma + n}{2}} \\
& \quad \times (\tau^{-1} + \mathbf{x}_i^\top \mathbf{H}_{\hat{\boldsymbol{\gamma}}} \mathbf{x}_i)^{-1/2}, \qquad (3)
\end{aligned}
$$

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

$$\begin{aligned}&\mathbf{m}_+(\hat{\boldsymbol\gamma}) = \\ &\left\{(\mathbf{y}^\top \mathbf{H}_{\hat{\gamma}} \mathbf{y} + b_\sigma)\mathbf{1}_p - \frac{(\mathbf{X}^\top \mathbf{H}_{\hat{\gamma}} \mathbf{y})^2}{\frac{1}{\tau}\mathbf{1}_p + \mathrm{diag}(\mathbf{X}^\top \mathbf{H}_{\hat{\gamma}} \mathbf{X})}\right\}^{-\frac{a_\sigma+n}{2}} \\ &\quad\times \left\{\frac{1}{\tau}\mathbf{1}_p + \mathrm{diag}(\mathbf{X}^\top \mathbf{H}_{\hat{\gamma}} \mathbf{X})\right\}^{-1/2},\end{aligned}$$

where $\mathbf{a}^x$ and $\mathbf{a}/\mathbf{b}$ indicate entrywise operations for generic vectors $\mathbf{a}$ and $\mathbf{b}$, accordingly. For example, $\mathbf{a}^x = (a_1^x, \ldots, a_p^x)$ and $\mathbf{a}/\mathbf{b} = (a_1/b_1, \ldots, a_p/b_p)$. Note that $\mathcal{N}_+(\hat{\boldsymbol\gamma}) = \{\hat{\boldsymbol\gamma} \cup \{i\} : i \notin \hat{\boldsymbol\gamma}\}$. Therefore, evaluating $m(\mathbf{y}|\boldsymbol\gamma)$ for all $\boldsymbol\gamma \in \mathcal{N}_+(\hat{\boldsymbol\gamma})$ can be done simultaneously in a single computation by obtaining the sub-vector of $\mathbf{m}_+(\hat{\boldsymbol\gamma})$ for $\{i : i \notin \hat{\boldsymbol\gamma}\}$. Similarly, for any $j \in \tilde{\boldsymbol\gamma}$, we can show that

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

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

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

It is easy to check that Eq. (4) is the $j$-th element of $\mathbf{m}_-(\tilde{\boldsymbol\gamma})$. Since $\mathcal{N}_-(\tilde{\boldsymbol\gamma}) = \{\tilde{\boldsymbol\gamma} \setminus \{j\} : j \in \tilde{\boldsymbol\gamma}\}$, evaluating $m(\mathbf{y}|\boldsymbol\gamma)$ for all $\boldsymbol\gamma \in \mathcal{N}_-(\tilde{\boldsymbol\gamma})$ can be done simultaneously in a single computation by obtaining the sub-vector of $\mathbf{m}_-(\tilde{\boldsymbol\gamma})$ for $\{j : j \in \tilde{\boldsymbol\gamma}\}$. Using the aforementioned calculation strategy, we can easily and quickly implement steps 3a), 3b), 5a), and 5b) of Algorithm 3. It is also worth noting that the proposed calculation method 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 handle the case of varying $k \leq K$, where $K$ is a prespecified upper bound. This is a common setting for high-dimensional best subset selection problems (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, $k \sim \mathrm{Uniform}\{1, \ldots, K\}$. 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 $k$ increases. 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

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

Note that (5) is equivalent to the following optimization problem:

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

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

1. For $k = 1, \ldots, K$, compute
$$\hat{\boldsymbol\gamma}_k = \arg\max_{\boldsymbol\gamma : |\boldsymbol\gamma| = k} m(\mathbf{y}|\boldsymbol\gamma)$$
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 a parallel or cluster computing environment is available, then it can be applied to the first step of the above procedure to run the for-loop over $k$ in parallel. The following theorem shows that the proposed Bayesian approach achieves the model selection consistency in the high-dimensional setting that $p > n$.

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

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

*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 and 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. We consider two cases. In case I, for $n = 100$, we generate the data $\{(y_i, x_{i1}, \ldots, x_{ip}) : i = 1, \ldots, n\}$ from the following linear regression model:

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

where $(x_{i1}, \ldots, x_{ip})^{\top} \overset{iid}{\sim} \text{Normal}(\mathbf{0}_p, \boldsymbol{\varSigma})$ with $\boldsymbol{\varSigma} = (\varSigma_{ij})_{p\times p}$ and $\varSigma_{ij} = \rho^{|i-j|}$, four $\beta_j$'s are randomly selected and then generated from Uniform$\{-1, -2, 1, 2\}$ independently, and the remaining $\beta$-coefficients are set equal to 0. In case II, for $n = 100$, we generate the data $\{(y_i, x_{i1}, \ldots, x_{ip}) : i = 1, \ldots, n\}$ from the following linear regression model:

$$y_i \overset{ind}{\sim} \text{Normal}\left(\sum_{j=1}^{p}\beta_j x_{ij}, \sigma_y^2\right),$$

where $\sigma_{y_i}^2 = 1{+}0.01\sqrt{x_{i1}}$, $(x_{i1}, \ldots, x_{ip})^{\top} \overset{iid}{\sim} \text{Normal}(\mathbf{0}_p, \boldsymbol{\varSigma})$ with $\boldsymbol{\varSigma} \sim \mathcal{IW}(\boldsymbol{\varPsi}, p{+}2)$, $\boldsymbol{\varPsi}, = (\psi_{ij})_{p\times p}$ and $\psi_{ij} = \rho^{|i-j|}$, four $\beta_j$'s are randomly selected and then generated from Uniform$\{-1, -2, 1, 2\}$ independently, and the remaining $\beta$-coefficients are set equal to 0.

For both cases, 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 in Section 4 for varying $k$ with $K = \lceil n^{2/3}\rceil = 22$. For the hybrid search algorithm given in Algorithm 3, we set $T = 100$, which is a remarkably small number of iterations compared to existing stochastic search algorithms (e.g., Kirkpatrick et al. 1983; George and McCulloch 1993; Hans et al. 2007), 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 four popular penalized likelihood methods: LASSO (Tibshirani 1996), the elastic net (Zou and Hastie 2005), SCAD (Fan and 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. In this simulation study, LASSO and ENET are implemented by R package `glmnet` and MCP and SCAD are performed by R package `ncvreg`. 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. False discovery rate is defined as $|\hat{\boldsymbol{\gamma}} \setminus \boldsymbol{\gamma}_*|/|\hat{\boldsymbol{\gamma}}|$ and Hamming distance is defined as $|\hat{\boldsymbol{\gamma}} \setminus \boldsymbol{\gamma}_*| + |\boldsymbol{\gamma}_* \setminus \hat{\boldsymbol{\gamma}}|$, where $\boldsymbol{\gamma}_*$ denotes the index set of the true model.

The results are shown in Table 1. For every case, 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. Hence, this finite sample performance supports our theoretical result in Theorem 2.

## 6 Real data application

In this section, we apply the proposed method to Breast invasive carcinoma (BRCA) data, which are 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 $5,000$ genes that are marginally correlated with

**Table 1** Simulation study results based on $2,000$ Monte Carlo replications for cases i–iv. Notation: FDR—false discovery rate; TRUE—percentage of the true model detected; SIZE—selected model size; HAM—Hamming distance; s.e.—standard error.

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

BRCA1. Then, the proposed method and the penalized likelihood methods as in Section 5 are applied to the reduced data with $n = 526$ and $p = 5,000$. To assess model fitting, for each method, we compute BIC (Schwarz 1978), extended BIC, and mean squared prediction error (MSPE) for the ordinary least squares (OLS) estimate that are obtained by using the selected predictors, where MSPE is estimated by Monte Carlo cross-validation over 500 replications based on 70% of training set and 30% of testing set.

**Table 2** Model comparison results for BRCA data

| | BIC | extended BIC | MSPE |
|---|---|---|---|
| Proposed | 984.449 | 1099.504 | 0.599 |
| SCAD | 1104.693 | 1166.472 | 0.681 |
| MCP | 1104.693 | 1166.472 | 0.681 |
| ENET | 1110.653 | 1186.245 | 0.683 |
| LASSO | 1104.693 | 1166.472 | 0.681 |

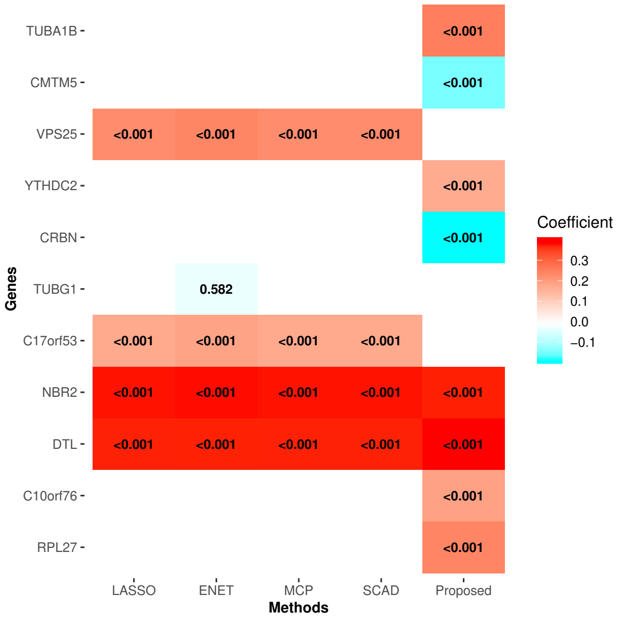

**Fig. 1** A heatmap of OLS coefficient estimates for the selected predictors with p-values.

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`.

| | NSP | MSE_Pred | BIC | EBIC |
|---|---|---|---|---|
| Proposed | 8.00 | 0.60 | 985.93 | 1120.87 |
| SCAD | 4.00 | 0.68 | 1104.69 | 1176.42 |
| MCP | 4.00 | 0.68 | 1104.69 | 1176.42 |
| ENET | 5.00 | 0.68 | 1110.65 | 1198.68 |
| LASSO | 4.00 | 0.68 | 1104.69 | 1176.42 |

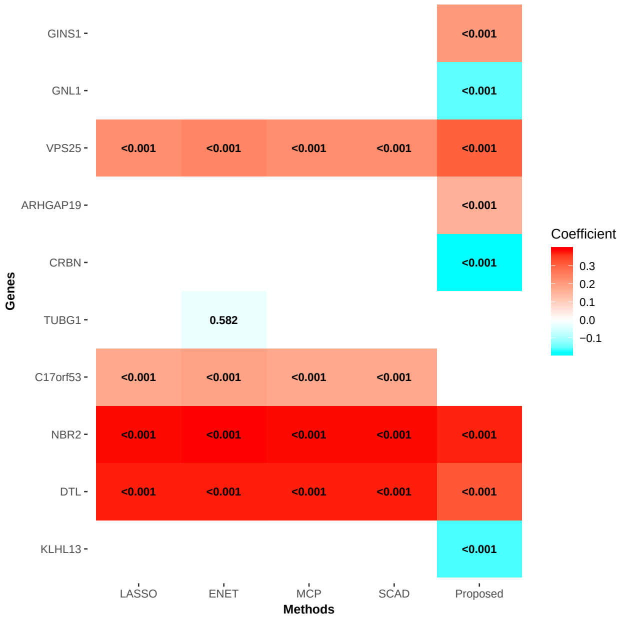

**Fig. 2** A heatmap of OLS coefficient estimates for the selected predictors with p-values without screening.

## 7 Discussion

The proposed hybrid search approach is computationally beneficial and practically effective compared with traditional stochastic search algorithms that are commonly used in Bayesian variable selection. Although the idea of searching over the neighborhood is somewhat similar to that of the MCMC model composition (MC$^3$) algorithm (Madigan and York 1995) and the shotgun stochastic search (SSS) algorithm (Hans et al. 2007), our search strategy is completely different from the existing methods. While MC$^3$ and SSS perform a stochastic search over the model space of all possible candidates, we implement a hybrid search over the sub-model space of size $k$. In addition, evaluating multiple candidates for the next move can be done simultaneously in a single computation in the proposed method.

In addition, 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.

## A Calculation of Equation (3)

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

$$m(\mathbf{y}|\hat{\gamma}\cup\{i\}) \propto |\mathbf{X}^{\top}_{\hat{\gamma}\cup\{i\}}\mathbf{X}_{\hat{\gamma}\cup\{i\}} + \tau^{-1}\mathbf{I}_{k+1}|^{-1/2} \times \left(\mathbf{y}^{\top}\mathbf{H}_{\hat{\gamma}\cup\{i\}}\mathbf{y} + b_{\sigma}\right)^{-\frac{a_\sigma+n}{2}}. \tag{8}$$

It follows from the Sherman-Morrison formula that

$$\mathbf{H}_{\hat{\gamma}\cup\{i\}} = \mathbf{H}_{\hat{\gamma}} - \frac{\mathbf{H}_{\hat{\gamma}}\mathbf{x}_i\mathbf{x}_i^{\top}\mathbf{H}_{\hat{\gamma}}}{\tau^{-1} + \mathbf{x}_i^{\top}\mathbf{H}_{\hat{\gamma}}\mathbf{x}_i}. \tag{9}$$

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

$$\begin{aligned}
&|\mathbf{X}^{\top}_{\hat{\gamma}\cup\{i\}}\mathbf{X}_{\hat{\gamma}\cup\{i\}} + \tau^{-1}\mathbf{I}_{k+1}| \\
&= \tau^{-(k+1)}|\tau\mathbf{X}_{\hat{\gamma}\cup\{i\}}\mathbf{X}^{\top}_{\hat{\gamma}\cup\{i\}} + \mathbf{I}_n| \\
&= \tau^{-(k+1)}|\tau\mathbf{X}_{\hat{\gamma}}\mathbf{X}^{\top}_{\hat{\gamma}} + \mathbf{I}_n + \tau\mathbf{x}_i\mathbf{x}_i^{\top}| \\
&= \tau^{-(k+1)}|\tau\mathbf{X}_{\hat{\gamma}}\mathbf{X}^{\top}_{\hat{\gamma}} + \mathbf{I}_n|\{1 + \tau\mathbf{x}_i^{\top}(\tau\mathbf{X}_{\hat{\gamma}}\mathbf{X}^{\top}_{\hat{\gamma}} + \mathbf{I}_n)^{-1}\mathbf{x}_i\} \\
&= |\mathbf{X}^{\top}_{\hat{\gamma}}\mathbf{X}_{\hat{\gamma}} + \tau^{-1}\mathbf{I}_k|(\tau^{-1} + \mathbf{x}_i^{\top}\mathbf{H}_{\hat{\gamma}}\mathbf{x}_i).
\end{aligned} \tag{10}$$

Applying (9) and (10) to (8), we thus have

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

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

## B Calculation of Equation (4)

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

$$m(\mathbf{y}|\tilde{\gamma}\setminus\{j\}) \propto |\mathbf{X}^{\top}_{\tilde{\gamma}\setminus\{j\}}\mathbf{X}_{\tilde{\gamma}\setminus\{j\}} + \tau^{-1}\mathbf{I}_{k}|^{-1/2} \times \left(\mathbf{y}^{\top}\mathbf{H}_{\tilde{\gamma}\setminus\{j\}}\mathbf{y} + b_{\sigma}\right)^{-\frac{a_\sigma+n}{2}}. \tag{11}$$

From the Sherman-Morrison formula, we have

$$\mathbf{H}_{\tilde{\gamma}\setminus\{j\}} = \mathbf{H}_{\tilde{\gamma}} + \frac{\mathbf{H}_{\tilde{\gamma}}\mathbf{x}_j\mathbf{x}_j^{\top}\mathbf{H}_{\tilde{\gamma}}}{\tau^{-1} - \mathbf{x}_j^{\top}\mathbf{H}_{\tilde{\gamma}}\mathbf{x}_j}. \tag{12}$$

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

$$\begin{aligned}
&|\mathbf{X}^{\top}_{\tilde{\gamma}\setminus\{j\}}\mathbf{X}_{\tilde{\gamma}\setminus\{j\}} + \tau^{-1}\mathbf{I}_{k}| \\
&= \tau^{-k}|\tau\mathbf{X}_{\tilde{\gamma}\setminus\{j\}}\mathbf{X}^{\top}_{\tilde{\gamma}\setminus\{j\}} + \mathbf{I}_n| \\
&= \tau^{-k}|\tau\mathbf{X}_{\tilde{\gamma}}\mathbf{X}^{\top}_{\tilde{\gamma}} + \mathbf{I}_n - \tau\mathbf{x}_j\mathbf{x}_j^{\top}| \\
&= \tau^{-k}|\tau\mathbf{X}_{\tilde{\gamma}}\mathbf{X}^{\top}_{\tilde{\gamma}} + \mathbf{I}_n|\{1 - \tau\mathbf{x}_j^{\top}(\tau\mathbf{X}_{\tilde{\gamma}}\mathbf{X}^{\top}_{\tilde{\gamma}} + \mathbf{I}_n)^{-1}\mathbf{x}_j\} \\
&= \tau^{2}|\mathbf{X}^{\top}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}} + \tau^{-1}\mathbf{I}_{k+1}|(\tau^{-1} - \mathbf{x}_j^{\top}\mathbf{H}_{\tilde{\gamma}}\mathbf{x}_j).
\end{aligned} \tag{13}$$

Hence, applying (12) and (13) to (11), we have

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

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

## 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 GM, Daza RM, Martin B, Zhang MD, Leith AP, Gasperini M, Janizek JD, Huang X, Starita LM, Shendure J (2018) Accurate classification of BRCA1 variants with saturation genome editing. Nature 562(7726):217–222

George EI, McCulloch RE (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 RR, Leslie RN (1967) Selection of the best subset in regression analysis. Technometrics 9(4):531–540

Kass RE, Raftery AE (1995) Bayes factors. Journal of the American Statistical Association 90(430):773–795

Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. Science 220(4598):671–680

Liang F, Liu C, Carroll RJ (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

Madigan D, York J (1995) Bayesian graphical models for discrete data. International Statistical Review 63(2):215–232

Schwarz G (1978) Estimating the dimension of a model. The Annals of Statistics 6(2):461–464

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

Zhang CH (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