Bayesian selection of best subsets in high-dimensional regression

Shiqiang Jin* Gyuhyeong Goh†

April 20, 2019

Abstract

With large numbers of predictors, regression analysis has faced many challenges including model fitting and prediction. As a power tool for dimension reduction, high-dimensional variable selection has drawn increasing attention in recent years. In this paper, we develop a new Bayesian approach to best subset selection that quickly identifies the best subset from a high-dimensional set of candidates via a hybrid search algorithm of deterministic local search and stochastic global search. To reduce the computational cost of evaluating many candidate subsets needed for each iteration, we propose an efficient and fast computation strategy that enables us to calculate exact posterior probabilities of all neighbor models simultaneously. The proposed Bayesian method possesses model selection consistency in the high-dimensional setting in which the number of candidate predictors is allowed to grow faster than the sample size. Simulation study and real data analysis are performed to assess and validate the proposed method.

Key words: Bayesian subset selection

1 Introduction

Variable selection plays a key role in recent regression analysis. In many statistical applications, especially in genetics studies, researchers have faced the situation in which the number of candidate predictors is extremely large but the sample size is relatively small, often referred to as a high-dimensional regression problem. The most pressing challenge in high-dimensional regression is to identify relevant predictor variables

*Department of Statistics, Kansas State University, Manhattan, KS 66506, U.S.A.
†Department of Statistics, Kansas State University, Manhattan, KS 66506, U.S.A.

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., Tibshirani 1996; Fan and Li 2001; Zou and Hastie 2005; Zhang et al. 2010). By adding a penalty function to the likelihood criterion, a penalized likelihood method produces sparse solutions that eliminate irrelevant predictors corresponding to the zero-estimated coefficients in the regression model.

In this paper, we are interested in selecting k important predictors out of p candidates, called the best subset selection problem (Hocking and Leslie, 1967). It is well known that the best subset selection involves non-convex optimization, which is computationally intractable in high-dimensional settings (i.e., when p is large). Although some penalized likelihood approaches such as Lasso, elastic net, and MCP provide a convex surrogate for the non-convex optimization problem, their applicability to the best subset selection is still limited (Bertsimas et al., 2016). Recently, 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 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 of deterministic local search and stochastic global search. 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 MCMC computation, is simultaneously accomplished.

2 Basic setup

Consider a multiple linear regression model,

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

where $\mathbf{y}$ is the n-dimensional response vector, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ is the $n \times p$ design matrix, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ 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 problems, 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. Here, our goal is to identify the most important k predictors in (1). To this end, let $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$ be a vector of binary variables such that $\gamma_j = 1$ if the j^{th} predictor, $\mathbf{x}_j$, is active and $\gamma_j = 0$ otherwise for $j = 1, \ldots, p$. For a given $\boldsymbol{\gamma}$, we can reduce the model (1) 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 $\boldsymbol{\beta}$ corresponding to the non-zero elements of $\boldsymbol{\gamma}$, respectively. Given $\boldsymbol{\gamma}$, we assign conjugate priors for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and σ^2 that are most commonly used in Bayesian linear regression,

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

where τ, a_σ, and b_σ denote hyperparameters and $|\boldsymbol{\gamma}| = \sum_{j=1}^{p} \gamma_j$ indicates the number of active predictors in the reduced model. 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

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

where $f(\mathbf{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2)$ denotes the likelihood given $\boldsymbol{\gamma}$ and $\mathbf{H}_{\boldsymbol{\gamma}} = \mathbf{I}_n - \mathbf{X}_{\boldsymbol{\gamma}} (\mathbf{X}_{\boldsymbol{\gamma}}^{\mathrm{T}} \mathbf{X}_{\boldsymbol{\gamma}} + \tau^{-1} \mathbf{I}_{|\boldsymbol{\gamma}|})^{-1} \mathbf{X}_{\boldsymbol{\gamma}}^{\mathrm{T}}$. By Bayes' theorem, the posterior model probability of $\boldsymbol{\gamma}$ is proportional to

$$\pi(\boldsymbol{\gamma}|\mathbf{y}) \propto m(\mathbf{y}|\boldsymbol{\gamma}) \pi(\boldsymbol{\gamma}) \propto \frac{(\tau^{-1})^{\frac{|\boldsymbol{\gamma}|}{2}}}{|\mathbf{X}_{\boldsymbol{\gamma}}^{\mathrm{T}} \mathbf{X}_{\boldsymbol{\gamma}} + \tau^{-1} \mathbf{I}_{|\boldsymbol{\gamma}|}|^{\frac{1}{2}} \left(\mathbf{y}^{\mathrm{T}} \mathbf{H}_{\boldsymbol{\gamma}} \mathbf{y} + b_\sigma\right)^{\frac{a_\sigma + n}{2}}} \mathbb{I}(|\boldsymbol{\gamma}| = k) := Q(\boldsymbol{\gamma}).$$

Therefore, our Bayesian best subset selection can be achieved by maximizing $Q(\boldsymbol{\gamma})$.

3 Best subset selection with fixed-size

In this section, we introduce new algorithms to reach at $\hat{\boldsymbol{\gamma}} = \arg\max_{\boldsymbol{\gamma}} \pi(\boldsymbol{\gamma}|\mathbf{y})$ for the fixed subset size, k. Let $\mathcal{N}_+(\boldsymbol{\gamma})$ be the set of all larger neighbors of $\boldsymbol{\gamma}$ obtained by adding one new predictor variable to $\boldsymbol{\gamma}$. Similarly, let $\mathcal{N}_-(\boldsymbol{\gamma})$ be the set of all smaller neighbors of $\boldsymbol{\gamma}$ obtained by deleting one variable from $\boldsymbol{\gamma}$. Suppose that $\tilde{\boldsymbol{\gamma}}$ represents the current best subset of size k. Our main idea is to update $\tilde{\boldsymbol{\gamma}}$ via the following two-step procedure until convergence:

> **Addition-step:** Using the marginal likelihood, *evaluate all models* in $\mathcal{N}_+(\tilde{\boldsymbol{\gamma}})$ *simultaneously* and then find the best subset of size $k+1$,
>
> $$\tilde{\boldsymbol{\eta}} = \arg \max_{\boldsymbol{\eta} \in \mathcal{N}_+(\tilde{\boldsymbol{\gamma}})} m(\mathbf{y}|\boldsymbol{\eta}).$$
>
> **Deletion-step:** Using the marginal likelihood, *evaluate all models* in $\mathcal{N}_-(\tilde{\boldsymbol{\eta}})$ *simultaneously* and then update the best subset of size k as
>
> $$\tilde{\boldsymbol{\gamma}} = \arg \max_{\boldsymbol{\gamma} \in \mathcal{N}_-(\tilde{\boldsymbol{\eta}})} m(\mathbf{y}|\boldsymbol{\gamma}).$$

The following theorem proves the convergence of the proposed algorithm.

Theorem 1. *The proposed algorithm monotonically increases the posterior probability, $\pi(\tilde{\boldsymbol{\gamma}}|\mathbf{y})$. In addition, the algorithm terminates in a finite number of iterations.*

The proof is given in Appendix A. The great merit of our approach is that evaluating all the candidates for each step can be done simultaneously. To explain this in more details, let $\mathcal{G}_0 = \{i : \tilde{\gamma}_i = 0\}$ be an index set of inactive predictors associated with $\tilde{\boldsymbol{\gamma}}$, where $\tilde{\gamma}_i$ is the i^{th} element of $\tilde{\boldsymbol{\gamma}}$. For each $i \in \mathcal{G}_0$, define $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{ip})$ by replacing the i^{th} element of $\tilde{\boldsymbol{\gamma}}$ with 1, that is, $\eta_{ij} = 1$ if $j = i$ and $\eta_{ij} = \tilde{\gamma}_j$ if $j \neq i$. For every $i \in \mathcal{G}_0$, using (2), it can be shown that

$$m(\mathbf{y}|\boldsymbol{\eta}_i) \propto \left[\mathbf{y}^{\mathrm{T}}\mathbf{H}_{\tilde{\boldsymbol{\gamma}}}\mathbf{y} - \frac{(\mathbf{x}_i^{\mathrm{T}}\mathbf{H}_{\tilde{\boldsymbol{\gamma}}}\mathbf{y})^2}{\tau^{-1} + \mathbf{x}_i^{\mathrm{T}}\mathbf{H}_{\tilde{\boldsymbol{\gamma}}}\mathbf{x}_i} + b_\sigma\right]^{-\frac{a_\sigma + n}{2}} (\tau^{-1} + \mathbf{x}_i^{\mathrm{T}}\mathbf{H}_{\tilde{\boldsymbol{\gamma}}}\mathbf{x}_i)^{-1/2}, \tag{3}$$

where $\mathbf{x}_i$ is the i^{th} column of $\mathbf{X}$. The details of our calculation in (3) are shown by Appendix B. It is important to note that $\mathbf{x}_i^{\mathrm{T}}\mathbf{H}_{\tilde{\boldsymbol{\gamma}}}\mathbf{x}_i$ is the i^{th} diagonal element of

$\mathbf{X}^{\mathrm{T}}\mathbf{H}_{\tilde{\boldsymbol{\gamma}}}\mathbf{X}$ and that $\mathbf{x}_i^{\mathrm{T}}\mathbf{H}_{\tilde{\boldsymbol{\gamma}}}\mathbf{y}$ is the i^{th} element of $\mathbf{X}^{\mathrm{T}}\mathbf{H}_{\tilde{\boldsymbol{\gamma}}}\mathbf{y}$. Let $s_{\tilde{\boldsymbol{\gamma}}} = \mathbf{y}^{\mathrm{T}}\mathbf{H}_{\tilde{\boldsymbol{\gamma}}}\mathbf{y}$, $\mathbf{d}_{\tilde{\boldsymbol{\gamma}}} = \mathrm{diag}(\mathbf{X}^{\mathrm{T}}\mathbf{H}_{\tilde{\boldsymbol{\gamma}}}\mathbf{X})$, and $\mathbf{v}_{\tilde{\boldsymbol{\gamma}}} = \mathbf{X}^{\mathrm{T}}\mathbf{H}_{\tilde{\boldsymbol{\gamma}}}\mathbf{y}$. Define

$$\mathbf{m}_{\tilde{\boldsymbol{\gamma}}}^{+} = \left[(s_{\tilde{\boldsymbol{\gamma}}} + b_\sigma)\mathbf{1}_p - \frac{\mathbf{v}_{\tilde{\boldsymbol{\gamma}}}^2}{\tau^{-1}\mathbf{1}_p + \mathbf{d}_{\tilde{\boldsymbol{\gamma}}}}\right]^{-\frac{a_\sigma+n}{2}} (\tau^{-1}\mathbf{1}_p + \mathbf{d}_{\tilde{\boldsymbol{\gamma}}})^{-1/2},$$

where $\mathbf{a}^x = (a_1^x, \ldots, a_p^x)$ and $\mathbf{a}/\mathbf{b} = (a_1/b_1, \ldots, a_p/b_p)$ for generic vectors $\mathbf{a}$ and $\mathbf{b}$. Note that $\mathcal{N}_+(\tilde{\boldsymbol{\gamma}}) = \{\boldsymbol{\eta}_i : i \in \mathcal{G}_0\}$. Hence, in the addition-step, evaluating all the models in $\mathcal{N}_+(\tilde{\boldsymbol{\gamma}})$ (i.e., calculating all the marginal likelihoods) can be done at once by extracting the sub-vector corresponding to $\mathcal{G}_0$ from $\mathbf{m}_{\tilde{\boldsymbol{\gamma}}}^{+}$. Similarly, let $\mathcal{E}_1 = \{\ell : \tilde{\eta}_\ell = 1\}$ be an index set of active predictors associated with $\tilde{\boldsymbol{\eta}}$, where $\tilde{\eta}_\ell$ denotes the ℓ^{th} element of $\tilde{\boldsymbol{\eta}}$. For each $\ell \in \mathcal{E}_1$, we define $\boldsymbol{\gamma}_\ell$ by replacing the ℓ^{th} element of $\tilde{\boldsymbol{\eta}}$ with 0. We can show that

$$m(\mathbf{y}|\boldsymbol{\gamma}_\ell) \propto \left[\mathbf{y}^{\mathrm{T}}\mathbf{H}_{\tilde{\boldsymbol{\eta}}}\mathbf{y} + \frac{(\mathbf{x}_\ell^{\mathrm{T}}\mathbf{H}_{\tilde{\boldsymbol{\eta}}}\mathbf{y})^2}{\tau^{-1} - \mathbf{x}_\ell^{\mathrm{T}}\mathbf{H}_{\tilde{\boldsymbol{\eta}}}\mathbf{x}_\ell} + b_\sigma\right]^{-\frac{a_\sigma+n}{2}} (\tau^{-1} - \mathbf{x}_\ell^{\mathrm{T}}\mathbf{H}_{\tilde{\boldsymbol{\eta}}}\mathbf{x}_\ell)^{-1/2}. \tag{4}$$

Some details of our calculation for (4) are shown in Appendix C. Define

$$\mathbf{m}_{\tilde{\boldsymbol{\eta}}}^{-} = \left[(s_{\tilde{\boldsymbol{\eta}}} + b_\sigma)\mathbf{1}_p + \frac{\mathbf{v}_{\tilde{\boldsymbol{\eta}}}^2}{\tau^{-1}\mathbf{1}_p - \mathbf{d}_{\tilde{\boldsymbol{\eta}}}}\right]^{-\frac{a_\sigma+n}{2}} (\tau^{-1}\mathbf{1}_p - \mathbf{d}_{\tilde{\boldsymbol{\eta}}})^{-1/2}.$$

Note that $\mathcal{N}_-(\tilde{\boldsymbol{\eta}}) = \{\boldsymbol{\gamma}_\ell : \ell \in \mathcal{E}_1\}$. Hence, as in the addition-step, evaluating all the candidates in the deletion-step can be done at once by extracting the sub-vector corresponding to $\mathcal{E}_1$ from $\mathbf{m}_{\tilde{\boldsymbol{\eta}}}^{-}$.

In general, hyperparameter τ can be chosen to be a large value so that the Gaussian prior for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ becomes approximately non-informative (or flat). However, when predictors are strongly correlated (i.e., when multicollinearity occurs), hyperparameter optimization for τ plays a crucial role to improve the prediction accuracy (Hoerl and Kennard, 1970). Since multicollinearly is common in high-dimensional settings, we propose to estimate the optimal value of τ using an empirical Bayesian approach. Note that, given $\boldsymbol{\gamma}$, the marginal likelihood in (2) can be considered as a function of τ. Taking the negative logarithm of the right hand side of (2) and applying Lemma

1 in Appendix D, we define

$$Q(\tau|\boldsymbol{\gamma}) = \frac{|\boldsymbol{\gamma}|}{2}\log\tau + \frac{1}{2}\sum_{j=1}^{p}\log(\lambda_{\boldsymbol{\gamma}j}+\tau^{-1}) + \frac{a_\sigma+n}{2}\log\left\{\mathbf{y}^{\mathrm{T}}\mathbf{y} - \sum_{j=1}^{p}\frac{\lambda_{\boldsymbol{\gamma}j}(\mathbf{u}_{\boldsymbol{\gamma}j}^{\mathrm{T}}\mathbf{y})^2}{\lambda_{\boldsymbol{\gamma}j}+\tau^{-1}} + b_\sigma\right\}, \quad (5)$$

where $\lambda_{\boldsymbol{\gamma}j}$ is the j^{th} largest eigenvalue of $\mathbf{X}_{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{X}_{\boldsymbol{\gamma}}$ and $\mathbf{u}_{\boldsymbol{\gamma}j}$ is the eigenvector of $\mathbf{X}_{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{X}_{\boldsymbol{\gamma}}$ corresponding to $\lambda_{\boldsymbol{\gamma}j}$. Then, for given $\tilde{\boldsymbol{\gamma}}$, we estimate the optimal value of τ, say $\tilde{\tau}$, by minimizing $Q(\tau|\tilde{\boldsymbol{\gamma}})$. Note that $\tilde{\tau}$ can be easily and quickly obtained by using a simple numerical optimization method such as the Newton-Raphson method. We now introduce our Bayesian best subset algorithm as follows:

Algorithm 1 Deterministic best subset search with a fixed k

1. Initialize $\tilde{\boldsymbol{\gamma}}$ and $\tilde{\tau}$.
2. **Repeat**

   Update $\tilde{\boldsymbol{\eta}} \leftarrow \arg\max_{\boldsymbol{\eta}\in\mathcal{N}_+(\tilde{\boldsymbol{\gamma}})} m(\mathbf{y}|\boldsymbol{\eta},\tilde{\tau})$;
   Update $\tilde{\boldsymbol{\gamma}} \leftarrow \arg\max_{\boldsymbol{\gamma}\in\mathcal{N}_-(\tilde{\boldsymbol{\eta}})} m(\mathbf{y}|\boldsymbol{\gamma},\tilde{\tau})$;
   Update $\tilde{\tau} \leftarrow \arg\min_{\tau\in(0,\infty)} Q(\tau|\tilde{\boldsymbol{\gamma}})$;

   until convergence.
3. Return $\hat{\boldsymbol{\gamma}} = \tilde{\boldsymbol{\gamma}}$.

A possible drawback of our deterministic search is that the proposed algorithm can get trapped in a local optimum. To resolve this issue, we propose to add a stochastic global search algorithm to the proposed deterministic local search algorithm. This additional step will be used only to check whether or not the local search algorithm reaches the global optimum. To develop a fast global search algorithm, we employ a notion of the shotgun stochastic search (Hans et al., 2007). However, while the shotgun stochastic search evaluates the candidate models using parallel computing, our simultaneous evaluation is based on analytical calculations. Our stochastic search algorithm can be implemented by generating a Markov chain via the following iterative sampling scheme:

1. Given $\boldsymbol{\gamma}^*$, generate $\boldsymbol{\eta}^*$ with probability,

$$\pi(\boldsymbol{\eta}|\mathbf{y}, \boldsymbol{\gamma}^*) = \frac{m(\mathbf{y}|\boldsymbol{\eta})}{\sum_{\boldsymbol{\eta}\in\mathcal{N}_+(\boldsymbol{\gamma}^*)} m(\mathbf{y}|\boldsymbol{\eta})}\mathbb{I}\{\boldsymbol{\eta} \in \mathcal{N}_+(\boldsymbol{\gamma}^*)\}.$$

2. Given $\boldsymbol{\eta}^*$, generate $\boldsymbol{\gamma}^*$ with probability

$$\pi(\boldsymbol{\gamma}|\mathbf{y}, \boldsymbol{\eta}^*) = \frac{m(\mathbf{y}|\boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma}\in\mathcal{N}_-(\boldsymbol{\eta}^*)} m(\mathbf{y}|\boldsymbol{\gamma})}\mathbb{I}\{\boldsymbol{\gamma} \in \mathcal{N}_-(\boldsymbol{\eta}^*)\}.$$

Recall that using (3) and (4), calculating $\pi(\boldsymbol{\eta}|\mathbf{y}, \boldsymbol{\gamma}^*)$ for all $\boldsymbol{\eta} \in \mathcal{N}_+(\boldsymbol{\gamma}^*)$ can be done simultaneously. Similarly, we can calculate $\pi(\boldsymbol{\gamma}|\mathbf{y}, \boldsymbol{\eta}^*)$ for all $\boldsymbol{\gamma} \in \mathcal{N}_-(\boldsymbol{\eta}^*)$ at once. Using the notion of the Metropolis-Hasting algorithm, it is easy to check that the states of generated Markov chain by our sampling method is equivalent to the support of $\pi(\boldsymbol{\gamma}|\mathbf{y})$(Revise it later). Let $\boldsymbol{\gamma}^{(t)}$ be the sample generated at iteration t. If the outcome from Algorithm 1, $\hat{\boldsymbol{\gamma}}$, is not the global maximum, then we must observe at least one case that $\pi(\hat{\boldsymbol{\gamma}}|\mathbf{y}) < \pi(\boldsymbol{\gamma}^{(t)}|\mathbf{y})$ (or equivalently $m(\mathbf{y}|\hat{\boldsymbol{\gamma}}) < m(\mathbf{y}|\boldsymbol{\gamma}^{(t)})$) with probability one as $t \to \infty$. Therefore, if this case does not occur for sufficiently large t, we can treat $\hat{\boldsymbol{\gamma}}$ as the global maximum. However, this global maximum is unefficient because it require large enough t. To improve the efficiency of global search, We applied the idea from Simulated Annealing to Algorithm 2:

1. Given $\boldsymbol{\gamma}^*$, generate $\boldsymbol{\eta}^*$ with probability,

$$\pi(\boldsymbol{\eta}|\mathbf{y}, \boldsymbol{\gamma}^*) = \frac{\tilde{m}(\mathbf{y}|\boldsymbol{\eta})}{\sum_{\boldsymbol{\eta}\in\mathcal{N}_+(\boldsymbol{\gamma}^*)} \tilde{m}(\mathbf{y}|\boldsymbol{\eta})}\mathbb{I}\{\boldsymbol{\eta} \in \mathcal{N}_+(\boldsymbol{\gamma}^*)\},$$

   where $\tilde{m}(\mathbf{y}|\boldsymbol{\eta}) = \exp\{\frac{1}{\alpha}\log(m(\mathbf{y}|\boldsymbol{\eta}))\}$.

2. Given $\boldsymbol{\eta}^*$, generate $\boldsymbol{\gamma}^*$ with probability

$$\pi(\boldsymbol{\gamma}|\mathbf{y}, \boldsymbol{\eta}^*) = \frac{\tilde{m}(\mathbf{y}|\boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma}\in\mathcal{N}_-(\boldsymbol{\eta}^*)} \tilde{m}(\mathbf{y}|\boldsymbol{\gamma})}\mathbb{I}\{\boldsymbol{\gamma} \in \mathcal{N}_-(\boldsymbol{\eta}^*)\},$$

   where $\tilde{m}(\mathbf{y}|\boldsymbol{\gamma}) = \exp\{\frac{1}{\alpha}\log(m(\mathbf{y}|\boldsymbol{\gamma}))\}$.

Algorithm 2 Hybrid best subset search with a fixed k

1. Initialize $\tilde{\boldsymbol{\gamma}}$ and $\tilde{\tau}$.
2. **Repeat**

   Update $\tilde{\boldsymbol{\eta}} \leftarrow \arg\max_{\boldsymbol{\eta} \in \mathcal{N}_+(\tilde{\boldsymbol{\gamma}})} m(\mathbf{y}|\boldsymbol{\eta}, \tilde{\tau})$;
   Update $\tilde{\boldsymbol{\gamma}} \leftarrow \arg\max_{\boldsymbol{\gamma} \in \mathcal{N}_-(\tilde{\boldsymbol{\eta}})} m(\mathbf{y}|\boldsymbol{\gamma}, \tilde{\tau})$;

   until convergence.
3. Set $\boldsymbol{\gamma}^{(0)} = \tilde{\boldsymbol{\gamma}}$.
4. **Repeat** for $t = 1, \ldots, T$:

   Generate $\boldsymbol{\eta}^{(t)} \sim \pi(\boldsymbol{\eta}|\mathbf{y}, \boldsymbol{\gamma}^{(t-1)})$;
   Generate $\boldsymbol{\gamma}^{(t)} \sim \pi(\boldsymbol{\gamma}|\mathbf{y}, \boldsymbol{\eta}^{(t)})$;
   If $m(\mathbf{y}|\tilde{\boldsymbol{\gamma}}) < m(\mathbf{y}|\boldsymbol{\gamma}^{(t)})$, **then** set $\tilde{\boldsymbol{\gamma}} = \boldsymbol{\gamma}^{(t)}$, break the loop, and go to 2.
5. Return $\hat{\boldsymbol{\gamma}} = \tilde{\boldsymbol{\gamma}}$.

4 Best subset selection with bounded size

In this section, we extend the proposed method to best subset selection with varying $k(\leq K)$ for a prespecified upper bound K, which is a common setting for high-dimensional best subset selection (e.g., Bertsimas et al. 2016; Liang et al. 2013). In our 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 is $\binom{p}{k}$ for k. To remove this issue, using a similar idea of Chen and Chen (2008), we define

$$\pi(k) \propto 1/\binom{p}{k}\mathbb{I}(k \leq K).$$

Let $\boldsymbol{\gamma}_k = (\gamma_{k1}, \ldots, \gamma_{kp})$ indicate a subset of size k. Then, our best subset selection can be performed by maximizing

$$\pi(\boldsymbol{\gamma}_k|\mathbf{y}) \propto m(\mathbf{y}|\boldsymbol{\gamma}_k)/\binom{p}{k},$$

subject to $k \leq K$. Hence, our best subset search algorithms with varying k can be summarized as follows:

Algorithm 3 Best subset search with varying k

1. **Repeat** for $k = 1, \ldots, K$:
   a. Initialize $\tilde{\boldsymbol{\gamma}}$ and $\tilde{\tau}$.
   b. Implement step 2 of Algorithm 1 for deterministic search.
   (or Implement steps 2 – 4 of Algorithm 2 for hybrid search.)
   c. Set $\hat{\boldsymbol{\gamma}}_k = \tilde{\boldsymbol{\gamma}}$.
2. Return $\hat{\boldsymbol{\gamma}} = \hat{\boldsymbol{\gamma}}_{\hat{k}}$, where

$$\hat{k} = \arg \max_{1 \leq k \leq K} \left\{ \log m(\mathbf{y}|\hat{\boldsymbol{\gamma}}_k) - \log \binom{p}{k} \right\}. \tag{6}$$

From Kass and Raftery (1995), for sufficiently large n, we have

$$\log m(\mathbf{y}|\boldsymbol{\gamma}_{k_1}) - \log m(\mathbf{y}|\boldsymbol{\gamma}_{k_2}) \approx \hat{l}(\mathbf{y}|\boldsymbol{\gamma}_{k_1}) - \hat{l}(\mathbf{y}|\boldsymbol{\gamma}_{k_2}) - \frac{1}{2}(k_1 - k_2)\log(n),$$

where $\hat{l}(\mathbf{y}|\boldsymbol{\gamma})$ denotes the log-likelihood evaluated at the maximum likelihood estimates given $\boldsymbol{\gamma}$. This implies that our posterior criterion in (6) is asymptotically equivalent to the following extended BIC,

$$\text{EBIC}(\boldsymbol{\gamma}_k) = -2\hat{l}(\mathbf{y}|\boldsymbol{\gamma}_k) + k\log(n) + 2\log \binom{p}{k}. \tag{7}$$

Consequently, by Theorem 1 of Chen and Chen (2008), our method possesses the model selection consistency in the high-dimensional setting with $p = O(n^{\xi})$ for $\xi \geq 1$.

5 Simulation study

In this section, we investigate the variable selection performance of our best subset selection algorithm on simulated high-dimensional data. For given $n = 100$, 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 $(x_{i1}, \ldots, x_{ip})^{\mathrm{T}} \overset{iid}{\sim} \text{Normal}(\mathbf{0}_p, \mathbf{\Sigma})$ with $\mathbf{\Sigma} = (\Sigma_{ij})_{p\times p}$ 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 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 distribution non-informative, we define the hyperparameters of inverse gamma prior as $a_\sigma = b_\sigma = 1$. Recall that we have proposed four algorithms: 1) deterministic search with fixed subset size, called DS_k, 2) deterministic search with bounded subset size, called $\text{DS}_{\leq K}$, 3) hybrid search with fixed subset size, called HS_k, 4) hybrid search with bounded subset size, called $\text{HS}_{\leq K}$. In $\text{DS}_{\leq K}$ and $\text{HS}_{\leq K}$, we define the upper bound of subset size by $K = \lceil n^{2/3} \rceil = 22$. In DS_k and HS_k, we define $k = 4$, which is the true number of active predictors in the data generating model. For the hybrid search algorithms (HS_k and $\text{HS}_{\leq K}$), the stochastic search iteration size, T, is set to $T = 100$. We use the marginal correlation between the response and each predictor to define the initial values. All simulation studies are implemented in the R statistical software with computer(Intel Core i7-7700, 32GB, Window 10 Enterprise)

For comparison purposes, we also apply the most popular penalized likelihood methods, LASSO (Tibshirani, 1996), the elastic net (Zou and Hastie, 2005), SCAD (Fan and Li, 2001), and MCP (Zhang et al., 2010), to the simulated data, where the regularization parameters (i.e., tuning parameters) are determined by the extended BIC in (7). To evaluate the variable selection performance, we calculate false discovery rate (FDR), percentage of selecting the exact true model (TRUE%), size of the selected model (SIZE), and Hamming distance (HAM) based on 2,000 replications.

Table 1: Simulation study results based on 2,000 replications; FDR (false discovery rate), TRUE% (percentage of the true model detected), SIZE (selected model size), HAM (Hamming distance), and s.e. (standard error).

Scenario	Method	FDR (s.e.)	TRUE% (s.e.)	SIZE (s.e.)	HAM (s.e.)
i	Proposed	0.016 (0.001)	92.600 (0.585)	4.087 (0.007)	0.087 (0.007)
	local(new)	0.01(0.001)	96.90 (0.388)	4.03 (0.004)	0.03 (0.004)
	Proposed(new)	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	0.030 (0.002)	86.250 (0.770)	4.034 (0.007)	0.236 (0.014)
	local(new)	0.03 (0.002)	86.15 (0.773)	3.99 (0.007)	0.25 (0.015)
	Proposed(new)	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	0.014 (0.001)	93.450 (0.553)	4.073 (0.007)	0.073 (0.007)
	local(new)	0.00 (0.001)	98.10 (0.305)	4.02 (0.003)	0.02(0.003)
	Proposed(new)	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	0.032 (0.002)	86.700 (0.759)	4.051 (0.007)	0.236 (0.015)
	local(new)	0.04 (0.002)	85.30 (0.792)	4.02 (0.006)	0.28 (0.016)
	Proposed(new)	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)

6 Real data example using mRNA gene expression data

6.1 Data and data precessing

"Breast Invasive Carcinoma" (BRCA) is mRNA gene expression dataset from the R package "curatedTCGAData" in Ramos (2019). This package provides us with publicly available data from The Cancer Genome Atlas (TCGA) Bioconductor Mul-

tiAssayExperiment class objects. BRCA dataset comprises 526 patients with primary solid tumor and 17814 gene expression measurements, including a gene called BRCA1. 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 some missing data, we have $n = 526$ samples with $17,323$ genes. Then it is further screened and normalized and we selected $p = 5000$ genes that are most marginally correlated with the response BRCA1 gene expression measurement.

6.2 Results

In our proposed method, the upper bound of subset size $K = \lceil n^{2/3} \rceil = 66$, $\tau = (\log p)^2 = 72.54$. To evaluate the fit of model, we computed the average of prediction MSE, BIC and EBIC under the our proposed method compared with SCAD, MCP, elastic net and LASSO. We used cross validation method to randomly split data into 70% training set and remaining part as testing set. For selected genes by each method, we fitted the linear model and compute the MSE, BIC and EBIC. Repeated this procedure 500 times and compute the average.

Table 2 shows our proposed method selects 8 genes (NSP: number of selected predictors) while others 4 or 5. In addition, it indicates that our proposed method outperforms other models since it has smallest values of prediction MSE, BIC and EBIC. Fig 1 gives us a heat map of the specific gene's names selected and the magnitude of coefficients estimated by each methods under the linear regression model with the whole data. "X" represents non-significance of coefficient.

Fig **??** is $\log m(\boldsymbol{\gamma}|\mathbf{y})$ with k predictors selected by proposed method. It increases to be maximized at when $k = 8$ and then decreases all the way down.

7 Comments

We have presented a novel approach for best subsets selection, a hybrid algorithm for exploring model spaces via local search and stochastic global search to consider

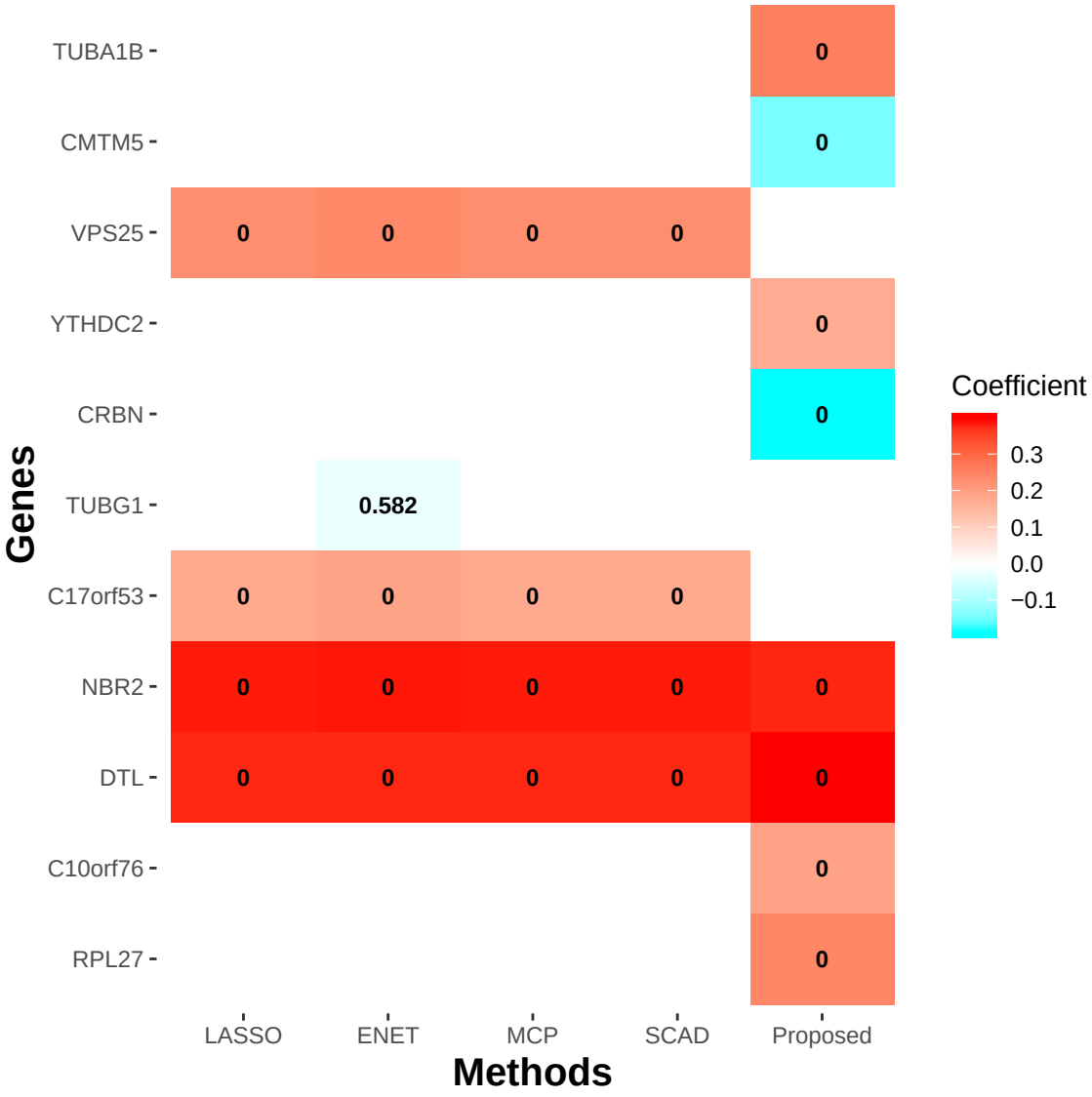

Figure 1: Heap map

Table 2: Predicion MSE, BIC and EBIC of Proposed method and other methods

	NSP	MSE_Pred	BIC	EBIC
Proposed	8.00	0.60	984.45	1099.50
SCAD	4.00	0.68	1104.69	1166.47
MCP	4.00	0.68	1104.69	1166.47
ENET	5.00	0.68	1110.65	1186.25
LASSO	4.00	0.68	1104.69	1166.47

and evaluate thousands of possible predictors simultaneously. The ability of identification of model spaces rapidly and accurately will be more practical than some traditional methods, such as MCMC stochastic search, which are inefficient due to high dimensions and high noise data. Furthermore, this approach can be applied into multivariate response regression model and binary regression model with logit link and probit link since marginal likelihood of them can be assessed or approximated quickly.

In future, we plan to introduce distributed computing (or parallel computing) to further improve the speed, e.g. evaluating best subset for each k by parallel computing. Besides, we plan to apply this algorithm into multivariate linear regression model, probit model and then multivariate mixed response regression model so that this model can be used in clinical data with both continuous and binary responses data.

A Proof

Proof of Theorem 1. Let $\tilde{\boldsymbol{\gamma}}^{(i)}$ be the best subset of size k updated by the i^{th} iteration. Then,

$$\tilde{\boldsymbol{\eta}}^{(i)} = \arg \max_{\boldsymbol{\eta} \in \mathcal{N}_+(\tilde{\boldsymbol{\gamma}}^{(i)})} m(\mathbf{y}|\boldsymbol{\gamma}) \quad \text{and} \quad \tilde{\boldsymbol{\gamma}}^{(i+1)} = \arg \max_{\boldsymbol{\gamma} \in \mathcal{N}_-(\tilde{\boldsymbol{\eta}}^{(i)})} m(\mathbf{y}|\boldsymbol{\gamma}).$$

Due to the fact that $\tilde{\boldsymbol{\gamma}}^{(i)} \in \mathcal{N}_-(\tilde{\boldsymbol{\eta}}^{(i)})$, we have

$$m(\mathbf{y}|\tilde{\boldsymbol{\gamma}}^{(i)}) \leq m(\mathbf{y}|\tilde{\boldsymbol{\gamma}}^{(i+1)}).$$

From Bayes' theorem, this implies

$$\pi(\tilde{\boldsymbol{\gamma}}^{(i)}|\mathbf{y}) \leq \pi(\tilde{\boldsymbol{\gamma}}^{(i+1)}|\mathbf{y}),$$

which proves our first statement. Since the number of all possible $\boldsymbol{\gamma}$ satisfying $|\boldsymbol{\gamma}| = k$ is finite, the algorithm terminates in a finite number of iterations and this completes our proof. □

B Calculation of (3)

$$\begin{aligned}
\pi(\boldsymbol{\gamma}^{+i}|\mathbf{y}, \boldsymbol{\gamma}) &\propto \frac{(\tau^{-1})^{(k+1)/2}}{|\mathbf{X}_{\boldsymbol{\gamma}^{+i}}^{\mathrm{T}} \mathbf{X}_{\boldsymbol{\gamma}^{+i}} + \tau^{-1} I_{k+1}|^{1/2}} \left(\mathbf{y}^{\mathrm{T}} \mathbf{H}_{\boldsymbol{\gamma}^{+i}} \mathbf{y} + b_\sigma\right)^{-\frac{a_\sigma + n}{2}} \mathbb{I}\left\{|\boldsymbol{\gamma}^{+i}| = k+1\right\} \\
&\propto |\mathbf{X}_{\boldsymbol{\gamma}^{+i}}^{\mathrm{T}} \mathbf{X}_{\boldsymbol{\gamma}^{+i}} + \tau^{-1} I_{k+1}|^{-1/2} \left(\mathbf{y}^{\mathrm{T}} (\mathbf{H}_{\boldsymbol{\gamma}} - \frac{\mathbf{H}_{\boldsymbol{\gamma}} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} \mathbf{H}_{\boldsymbol{\gamma}}}{\tau^{-1} + \mathbf{x}_i^{\mathrm{T}} \mathbf{H}_{\boldsymbol{\gamma}} \mathbf{x}_i}) \mathbf{y} + b_\sigma\right)^{-\frac{a_\sigma + n}{2}} \\
&= |\mathbf{X}_{\boldsymbol{\gamma}}^{\mathrm{T}} \mathbf{X}_{\boldsymbol{\gamma}} + \tau^{-1} \mathbf{I}_k|^{-1/2} \left(\tau^{-1} + \mathbf{x}_i^{\mathrm{T}} \mathbf{H}_{\boldsymbol{\gamma}} \mathbf{x}_i\right)^{-1/2} \left(\mathbf{y}^{\mathrm{T}} (\mathbf{H}_{\boldsymbol{\gamma}} - \frac{\mathbf{H}_{\boldsymbol{\gamma}} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} \mathbf{H}_{\boldsymbol{\gamma}}}{\tau^{-1} + \mathbf{x}_i^{\mathrm{T}} \mathbf{H}_{\boldsymbol{\gamma}} \mathbf{x}_i}) \mathbf{y} + b_\sigma\right)^{-\frac{a_\sigma + n}{2}} \\
&\propto \left[\mathbf{y}^{\mathrm{T}} \mathbf{H}_{\boldsymbol{\gamma}} \mathbf{y} - \frac{(\mathbf{x}_i^{\mathrm{T}} \mathbf{H}_{\boldsymbol{\gamma}} \mathbf{y})^2}{\tau^{-1} + \mathbf{x}_i^{\mathrm{T}} \mathbf{H}_{\boldsymbol{\gamma}} \mathbf{x}_i} + b_\sigma\right]^{-\frac{a_\sigma + n}{2}} (\tau^{-1} + \mathbf{x}_i^{\mathrm{T}} \mathbf{H}_{\boldsymbol{\gamma}} \mathbf{x}_i)^{-1/2} \qquad (8)
\end{aligned}$$

The second proportion holds by Lemma 1 in **??** and the equation holds by Lemma 4.1 in **??**

C Calculation of (4)

The details are shown in Appendix C.

$$
\begin{aligned}
& \pi(\boldsymbol{\gamma}^{*-\ell}|\mathbf{y},\boldsymbol{\gamma}^{*}) \\
\propto\ & \frac{(\tau^{-1})^{k/2}}{|\mathbf{X}_{\boldsymbol{\gamma}^{*-\ell}}^{\mathrm{T}}\mathbf{X}_{\boldsymbol{\gamma}^{*-\ell}}+\tau^{-1}I_k|^{1/2}}\left(\mathbf{y}^{\mathrm{T}}\mathbf{H}_{\boldsymbol{\gamma}^{*-\ell}}\mathbf{y}+b_\sigma\right)^{-\frac{a_\sigma+n}{2}}\mathbb{I}\left\{|\boldsymbol{\gamma}^{*-\ell}|=k\right\} \\
\propto\ & |\mathbf{X}_{\boldsymbol{\gamma}^{*-\ell}}^{\mathrm{T}}\mathbf{X}_{\boldsymbol{\gamma}^{*-\ell}}+\tau^{-1}I_k|^{-1/2}\left(\mathbf{y}^{\mathrm{T}}(\mathbf{H}_{\boldsymbol{\gamma}^{*}}+\frac{\mathbf{H}_{\boldsymbol{\gamma}^{*}}\mathbf{x}_\ell\mathbf{x}_\ell^{\mathrm{T}}\mathbf{H}_{\boldsymbol{\gamma}^{*}}}{\tau^{-1}-\mathbf{x}_\ell^{\mathrm{T}}\mathbf{H}_{\boldsymbol{\gamma}^{*}}\mathbf{x}_\ell})\mathbf{y}+b_\sigma\right)^{-\frac{a_\sigma+n}{2}} \\
=\ & |\mathbf{X}_{\boldsymbol{\gamma}^{*}}^{\mathrm{T}}\mathbf{X}_{\boldsymbol{\gamma}^{*}}+\tau^{-1}\mathbf{I}_{k+1}|^{-1/2}(\tau^{-1}-\mathbf{x}_\ell^{\mathrm{T}}\mathbf{H}_{\boldsymbol{\gamma}^{*}}\mathbf{x}_\ell)^{-1/2}\left[\mathbf{y}^{\mathrm{T}}\mathbf{H}_{\boldsymbol{\gamma}^{*}}\mathbf{y}+\frac{(\mathbf{x}_\ell^{\mathrm{T}}\mathbf{H}_{\boldsymbol{\gamma}^{*}}\mathbf{y})^2}{\tau^{-1}-\mathbf{x}_\ell^{\mathrm{T}}\mathbf{H}_{\boldsymbol{\gamma}^{*}}\mathbf{x}_\ell}+b_\sigma\right]^{-\frac{a_\sigma+n}{2}} \\
\propto\ & \left[\mathbf{y}^{\mathrm{T}}\mathbf{H}_{\boldsymbol{\gamma}^{*}}\mathbf{y}+\frac{(\mathbf{x}_\ell^{\mathrm{T}}\mathbf{H}_{\boldsymbol{\gamma}^{*}}\mathbf{y})^2}{\tau^{-1}-\mathbf{x}_\ell^{\mathrm{T}}\mathbf{H}_{\boldsymbol{\gamma}^{*}}\mathbf{x}_\ell}+b_\sigma\right]^{-\frac{a_\sigma+n}{2}}(\tau^{-1}-\mathbf{x}_\ell^{\mathrm{T}}\mathbf{H}_{\boldsymbol{\gamma}^{*}}\mathbf{x}_\ell)^{-1/2}, \qquad (9)
\end{aligned}
$$

D Lemmas

Lemma 1. *Let* $\mathbf{X}=UDV^{\mathrm{T}}$. *We have*

$$
\begin{aligned}
|\mathbf{X}^{\mathrm{T}}\mathbf{X}+\tau^{-1}\mathbf{I}_p| &= |VD^{\mathrm{T}}U^{\mathrm{T}}UDV^{\mathrm{T}}+\tau^{-1}VV^{\mathrm{T}}| \\
&= |VD^{\mathrm{T}}DV^{\mathrm{T}}+V\tau^{-1}\mathbf{I}_pV^{\mathrm{T}}| \\
&= |V(D^{\mathrm{T}}D+\tau^{-1}\mathbf{I}_p)V^{\mathrm{T}}| \\
&= |V||D^{\mathrm{T}}D+\tau^{-1}\mathbf{I}_p||V^{\mathrm{T}}| \\
&= |VV^{\mathrm{T}}||D^{\mathrm{T}}D+\tau^{-1}\mathbf{I}_p| \\
&= |D^{\mathrm{T}}D+\tau^{-1}\mathbf{I}_p| \\
&= \prod_{j=1}^{p}(d_j^2+\tau^{-1}) \qquad (10)
\end{aligned}
$$

where d_j is j^{th} diagonal element of D. Also, we have

$$\begin{aligned}
& \mathbf{y}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\tau^{-1}\mathbf{I}_p)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \\
= \; & \mathbf{y}^{\mathrm{T}}UDV^{\mathrm{T}}(VD^{\mathrm{T}}U^{\mathrm{T}}UDV^{\mathrm{T}}+V\tau^{-1}\mathbf{I}_pV^{\mathrm{T}})^{-1}VD^{\mathrm{T}}U^{\mathrm{T}}\mathbf{y} \\
= \; & \mathbf{y}^{\mathrm{T}}UDV^{\mathrm{T}}(V(D^{\mathrm{T}}D+\tau^{-1}\mathbf{I}_p)V^{\mathrm{T}})^{-1}VD^{\mathrm{T}}U^{\mathrm{T}}\mathbf{y} \\
= \; & \mathbf{y}^{\mathrm{T}}UD(D^{\mathrm{T}}D+\tau^{-1}\mathbf{I}_p)^{-1}D^{\mathrm{T}}U^{\mathrm{T}}\mathbf{y} \\
= \; & \mathbf{y}^{\mathrm{T}}UD(\mathrm{diag}(d_j^2+\tau^{-1}))^{-1}D^{\mathrm{T}}U^{\mathrm{T}}\mathbf{y} \\
= \; & \mathbf{y}^{\mathrm{T}}U\mathrm{diag}\left(\frac{d_j^2}{d_j^2+\tau^{-1}}\right)U^{\mathrm{T}}\mathbf{y} \\
= \; & \sum_j \frac{d_j^2 z_j^2}{d_j^2+\tau^{-1}} \qquad (11)
\end{aligned}$$

where z_j is the j^{th} component of $U^{\mathrm{T}}\mathbf{y}$.

References

Bertsimas, D., A. King, R. Mazumder, et al. (2016). Best subset selection via a modern optimization lens. *The annals of statistics 44*(2), 813–852.

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

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

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

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

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

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*(1), 55–67.

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

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

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

Ramos, M. (2019). *curatedTCGAData: Curated Data From The Cancer Genome Atlas (TCGA) as MultiAssayExperiment Objects*. R package version 1.4.3.

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

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

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