Computational Statistics
https://doi.org/10.1007/s00180-023-01368-y

ORIGINAL PAPER

A hybrid deterministic–deterministic approach for high-dimensional Bayesian variable selection with a default prior

Jieun Lee[1] · **Gyuhyeong Goh**[1]

Received: 16 October 2022 / Accepted: 15 May 2023

Abstract
Identifying relevant variables among numerous potential predictors has been of primary interest in modern regression analysis. While stochastic search algorithms have surged as a dominant tool for Bayesian variable selection, when the number of potential predictors is large, their practicality is constantly challenged due to high computational cost as well as slow convergence. In this paper, we propose a new Bayesian variable selection scheme by using hybrid deterministic–deterministic variable selection (HD-DVS) algorithm that asymptotically ensures a rapid convergence to the global mode of the posterior model distribution. A key feature of HD-DVS is that it allows us to circumvent the iterative computation of inverse matrices, which is a common computational bottleneck in Bayesian variable selection. A simulation study is conducted to demonstrate that our proposed method outperforms existing Bayesian and frequentist methods. An analysis of the Bardet–Biedl syndrome gene expression data is presented to illustrate the applicability of HD-DVS to real data.

Keywords Forward selection · Greedy algorithm · High-dimensional Bayesian linear regression · Highest probability model (HPM)

1 Introduction

With the increasing influence of high-dimensional data, variable selection has emerged as an essential task in regression analysis. In a Bayesian framework, variable selection is generally performed by using posterior model probabilities that quantify the plausibility of candidate models given the data. However, when the

✉ Gyuhyeong Goh
ggoh@ksu.edu

Jieun Lee
jieunl@ksu.edu

[1] Department of Statistics, Kansas State University, 1116 Mid-Campus Drive N., Manhattan, KS 66506-0802, USA

Published online: 27 May 2023

total number of potential predictors, p, is large, the implementation of Bayesian variable selection becomes extremely expensive and difficult due to the multimodality of posterior model distributions.

The use of Markov chain Monte Carlo (MCMC) sampling has long been the dominant way of Bayesian variable selection, often referred to as a stochastic search method. For example, George and McCulloch (1993) propose to use a Gibbs sampler with spike-and-slab Gaussian mixture priors. For the problem of large p regression, Hans et al. (2007) introduce the idea of stochastic shotgun search that exploits parallel computing to evaluate many candidate models simultaneously. A comprehensive review of MCMC methods for Bayesian variable selection can be found in Tadesse and Vannucci (2021).

Despite the substantial developments in stochastic search methods, the demand for new approaches is still strong due to the intrinsic limitations of MCMC techniques including convergence issues and massive computational costs for dealing with high-dimensional data. In response to this need, we propose a new deterministic search approach for high-dimensional Bayesian regression problems. To the best of our knowledge, the first attempt to exploit a deterministic approach for high-dimensional Bayesian variable selection was made by Ročková and George (2014) in which an expectation–maximization (EM) algorithm is used to quickly identify posterior modes under the spike-and-slab Gaussian mixture prior. While there are several extensions (e.g., Zhao and Lian 2016; Koslovsky et al. 2018; Ročková and George 2018) existing deterministic approaches for Bayesian variable selection are limited to finding the mode of the posterior parameter distribution rather than the mode of the posterior model distribution.

In this paper, we develop a novel Bayesian variable selection scheme called hybrid deterministic-deterministic variable selection (HD-DVS) that quickly estimates the global mode of the posterior model distribution by alternating between two deterministic search algorithms. For the prior of the regression coefficient vector, we consider Zellner's g-prior, the most common choice for Bayesian model selection. A distinctive feature of HD-DVS is that a fast convergence to the global optimum is guaranteed asymptotically. A long-standing problem in high-dimensional Bayesian variable selection is the computational cost associated with iterative calculations of inverse matrices (Bhattacharya et al. 2016). Inspired by Jin and Goh (2021), we develop a fast HD-DVS implementation strategy that bypasses the direct calculation of inverse matrices. However, our work is different from Jin and Goh (2021) especially in that the proposed method in this paper is fully deterministic while the aforementioned method relies on a stochastic search component.

The remainder of this paper is organized as follows. Section 2 states the basic model settings, notations, and assumptions for the problem of high-dimensional Bayesian linear regression. Section 3 describes the key idea of HD-DVS and its important properties. Section 4 introduces the fast implementation strategy for HD-DVS with technical details. In Sect. 5, the finite sample performance of our proposed method is investigated via a comparative simulation study. Section 6 illustrates the reliability and practicality of HD-DVS based on the analysis of Bardet–Biedl

syndrome gene expression data. Section 7 concludes with some remarks and directions for future research. The proofs of theorems are given in Appendices.

2 Bayesian formulations and challenges in high-dimensional regression

We consider the problem of variable selection in the context of linear regression modeling with a response variable, Y, and p potential predictor variables, $X_1, \ldots, X_p$. Suppose that $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ is an $n \times 1$ vector of independent realizations of Y and $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^{\mathrm{T}}$ is an $n \times 1$ vector of independent realizations of X_j for $j = 1, \ldots, p$. The full linear regression model is given as

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

where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_p]$ is the $n \times p$ design matrix, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ denotes the $p \times 1$ unknown regression coefficient vector, and $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$. Suppose that many potential predictors are available (i.e., p is large), but most of them are redundant or irrelevant in explaining the variation of the response variable. In such a case, it is essential to perform variable selection for accurate estimation and prediction as well as dimension reduction purpose. In our paper, the main objective is to develop a fast Bayesian variable selection method in high-dimensional regression settings. Throughout the paper, we assume that $\mathbf{y}$ is centered (i.e., $\mathbf{y}^{\mathrm{T}}\mathbf{1}_n = 0$) and $\mathbf{x}_1, \ldots, \mathbf{x}_p$ are standardized (i.e., $\mathbf{x}_j^{\mathrm{T}}\mathbf{1}_n = 0$ and $n^{-1}\|\mathbf{x}_j\|^2 = 1$, $j = 1, \ldots, p$) so that the intercept term can always be omitted in linear regression models.

To indicate important predictors, let $\gamma \subset \{1, \ldots, p\}$ be the set of their indices in model (1). Let p_γ denote the number of elements in γ. Given γ, we define the $n \times p_\gamma$ design matrix $\mathbf{X}_\gamma$ whose columns correspond to $\mathbf{x}_j$ for $j \in \gamma$. If γ is known, the problem of high-dimensional regression can be solved by reducing the full model (1) to

$$\mathbf{y} = \mathbf{X}_\gamma \boldsymbol{\beta}_\gamma + \boldsymbol{\epsilon}, \tag{2}$$

where $\boldsymbol{\beta}_\gamma$ denotes the $p_\gamma \times 1$ vector of regression coefficients. This implies that the variable selection problem is equivalent to the problem of estimating γ. From a Bayesian perspective, γ is treated as a random variable; therefore, variable selection can be performed by using the posterior model probabilities, $p(\gamma \mid \mathbf{y})$ for $\gamma \in \mathcal{G}$, where $\mathcal{G}$ represents the set of candidate models under consideration. Following Chen and Chen (2008) and Yang et al. (2016), we assume $\mathcal{G} = \{\gamma : p_\gamma \leq K_n\}$ for some $K_n = o(n)$.

Let $p(\boldsymbol{\beta}_\gamma, \sigma^2 \mid \gamma)$ be the joint prior density of $(\boldsymbol{\beta}_\gamma, \sigma^2)$ given γ. The marginal likelihood under the reduced model (2) is calculated as

$$m(\mathbf{y} \mid \gamma) = \int \int f(\mathbf{y} \mid \boldsymbol{\beta}_\gamma, \sigma^2, \gamma) p(\boldsymbol{\beta}_\gamma, \sigma^2 \mid \gamma) d\boldsymbol{\beta}_\gamma d\sigma^2, \tag{3}$$

where $f(\mathbf{y} \mid \boldsymbol{\beta}_\gamma, \sigma^2, \gamma)$ denotes the likelihood function under the reduced model (2). Let $p(\gamma)$ represent the prior model probabilities for $\gamma \in \mathcal{G}$. By Bayes' theorem, the posterior probability of γ can be expressed as

$$p(\gamma \mid \mathbf{y}) = \frac{m(\mathbf{y} \mid \gamma)p(\gamma)}{\sum_{\gamma \in \mathcal{G}} m(\mathbf{y} \mid \gamma)p(\gamma)} \propto m(\mathbf{y} \mid \gamma)p(\gamma),$$

where the symbol $\propto$ means "proportional to". It is well known that the optimal Bayesian decision under the 0–1 loss function is to estimate γ by the highest probability model (HPM), which is given by

$$\gamma_{\text{HPM}} = \arg\max_{\gamma \in \mathcal{G}} p(\gamma \mid \mathbf{y}) = \arg\max_{\gamma \in \mathcal{G}} m(\mathbf{y} \mid \gamma)p(\gamma). \tag{4}$$

On the other hand, when Bayesian variable selection is performed for prediction, the median probability model (MPM) is shown to be the optimal choice (Barbieri and Berger 2004). It is worth noting that when $p(\gamma_{\text{HPM}} \mid \mathbf{y}) \geq 0.5$, HPM is also MPM. Since the priors used in this paper will be specified to achieve the posterior model consistency in Moreno et al. (2015), we are asymptotically guaranteed to have $p(\gamma_{\text{HPM}} \mid \mathbf{y}) \geq 0.5$. Hence, we restrict our attention to finding γ_{HPM}.

The implementation of Bayesian variable selection requires us to specify the priors, $p(\boldsymbol{\beta}_\gamma, \sigma^2 \mid \gamma)$ and $p(\gamma)$. In the context of high-dimensional variable selection, the determination of $p(\boldsymbol{\beta}_\gamma, \sigma^2 \mid \gamma)$ should be performed with special attention paid to the computational complexity of the marginal likelihood in (3). For the convenience of marginal likelihood computation, we consider the g-prior (Zellner 1986):

$$\begin{aligned} \boldsymbol{\beta}_\gamma \mid \sigma^2, \gamma &\sim N_{p_\gamma}\left\{\mathbf{0}, \sigma^2 g(\mathbf{X}_\gamma^{\mathrm{T}}\mathbf{X}_\gamma)^{-1}\right\}, \\ p(\sigma^2 \mid \gamma) &\propto 1/\sigma^2, \end{aligned}$$

where $g > 0$ is a hyperparameter. Following Kass and Wasserman (1995), we set $g = n$ which induces the unit information prior that contains the amount of information of one data point. A key merit of the g-prior is that we can obtain a closed-form expression for the marginal likelihood as follows:

$$\begin{aligned} m(\mathbf{y} \mid \gamma) &= \frac{\Gamma(n/2)}{\pi^{n/2}(1+g)^{p_\gamma/2}\left\{\|\mathbf{y}\|^2 - \frac{g}{1+g}\mathbf{y}^{\mathrm{T}}\mathbf{X}_\gamma(\mathbf{X}_\gamma^{\mathrm{T}}\mathbf{X}_\gamma)^{-1}\mathbf{X}_\gamma^{\mathrm{T}}\mathbf{y}\right\}^{n/2}} \\ &\propto (1+g)^{-p_\gamma/2}\left\{\|\mathbf{y}\|^2 - \frac{g}{1+g}\mathbf{y}^{\mathrm{T}}\mathbf{X}_\gamma(\mathbf{X}_\gamma^{\mathrm{T}}\mathbf{X}_\gamma)^{-1}\mathbf{X}_\gamma^{\mathrm{T}}\mathbf{y}\right\}^{-n/2}, \end{aligned} \tag{5}$$

where $\|\mathbf{y}\|^2 = \mathbf{y}^{\mathrm{T}}\mathbf{y}$. Since variable selection is the problem of multiple testing, controlling for multiplicity is crucial, especially in high dimensional data settings. To induce a multiplicity-correction effect, we consider the hierarchical uniform model prior (Scott and Berger 2010; Moreno et al. 2015) for γ:

$$p(\gamma) \propto \binom{p}{p_\gamma}^{-1} \mathbb{I}(\gamma \in \mathcal{G}), \tag{6}$$

where $\mathbb{I}(\cdot)$ is the indicator function. Define

$$D(\gamma) = n \log\left\{ \|\mathbf{y}\|^2 - \frac{n}{1+n} \mathbf{y}^{\mathrm{T}} \mathbf{P}_\gamma \mathbf{y} \right\} + p_\gamma \log(n+1) + 2 \log \binom{p}{p_\gamma}, \tag{7}$$

where $\mathbf{P}_\gamma = \mathbf{X}_\gamma (\mathbf{X}_\gamma^{\mathrm{T}} \mathbf{X}_\gamma)^{-1} \mathbf{X}_\gamma^{\mathrm{T}}$. Applying (5) and (6) with $g = n$ to (4), it is easy to see that

$$\gamma_{\mathrm{HPM}} = \arg\max_{\gamma \in \mathcal{G}} p(\gamma \mid \mathbf{y}) = \arg\min_{\gamma \in \mathcal{G}} D(\gamma).$$

Hence, for computational and notational convenience, we will use $D(\gamma)$ as an equivalence of the posterior model probability, $p(\gamma \mid \mathbf{y})$. From the above aspect, $D(\cdot)$ can be interpreted as a Bayesian model selection measure that quantifies the level of surprise of choosing γ given the data. It is important to note that our Bayesian variable selection framework is equivalent to the following hierarchical model with the point-mass spike-and-slab prior:

$$\begin{aligned}
\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2 &\sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n), \\
\boldsymbol{\beta}_\gamma \mid \sigma^2, \gamma &\sim N_p\left\{ \mathbf{0}, \sigma^2 g (\mathbf{X}_\gamma^{\mathrm{T}} \mathbf{X}_\gamma)^{-1} \right\}, \\
\beta_j \mid \sigma^2, \gamma &\sim \delta_0, \quad j \notin \gamma, \\
p(\sigma^2 \mid \gamma) &\propto 1/\sigma^2, \\
p(\gamma) &\propto \binom{p}{p_\gamma}^{-1} \mathbb{I}(\gamma \in \mathcal{G}),
\end{aligned}$$

where δ_0 represents the degenerate distribution at 0.

The implementation of Bayesian model selection using $D(\cdot)$ seems to be straightforward. However, when p is large, finding the global minimum is computationally intractable due to the non-convex nature of $D(\cdot)$. Specifically, when the number of competing models is enormous, the exhaustive search strategy is computationally infeasible. The MCMC-based stochastic search approach has been the most popular solution to the problem of Bayesian variable selection (e.g., George and McCulloch 1993; Casella and Moreno 2006; Hans et al. 2007). The stochastic search method exploits the fact that HPM has the highest probability when we generate a random variate from $p(\gamma \mid \mathbf{y})$. In our framework, posterior sampling from $p(\gamma \mid \mathbf{y})$ is possible via the Gibbs sampler that cycles through the full conditionals, $p(z_j \mid \mathbf{z}_{-j}, \mathbf{y})\,(j = 1, \ldots, p)$, where $\mathbf{z} = (z_1, \ldots, z_p)^{\mathrm{T}}$ is the p-dimensional binary vector with $z_j = \mathbb{I}(j \in \gamma)$ and $\mathbf{z}_{-j}$ is the sub-vector of $\mathbf{z}$ obtained by deleting the j-th element of $\mathbf{z}$. Let $\{\mathbf{z}^{(t)} : t = 1, \ldots, T\}$ be a posterior sample generated by the Gibbs sampler. Then, the posterior sample from $p(\gamma \mid \mathbf{y})$ is obtained as $\{\gamma^{(t)} = \{j : z_j^{(t)} = 1\} : t = 1, \ldots, T\}$. As the number of the Gibbs itera-

tions, T, goes to infinity, $\hat{\gamma} = \arg\max_{\gamma^{(t)}:t=1,\ldots,T} p(\gamma^{(t)} \mid \mathbf{y})$ converges to the posterior mode, γ_{HPM}. When p is large, however, the element-wise Gibbs sampling scheme results in a slow mixing rate of the Markov chain. To improve the convergence rate of the Gibbs sampler, various continuous approximations to the point-mass spike-and-slab prior have been proposed. For example, the following normal spike-and-slab is most commonly used in high-dimensional Bayesian variable selection:

$$\beta_j \mid \sigma^2, \gamma \sim (1 - z_j)N(0, \sigma^2 v_0) + z_j N(0, \sigma^2 v_1),$$

where v_0 and v_1 are prespecified hyperparameters such that $0 < v_0 \leq v_1$. The major advantage of continuous spike-and-slab priors is that the sampling of the entire vector of $\mathbf{z}$ (or equivalently γ) is feasible via the Gibbs sampling scheme of George and McCulloch (1993). However, when p is large, the computational cost of the Gibbs sampler becomes prohibitively expensive due to the problem of the $p \times p$ inverse matrix computation (Bhattacharya et al. 2016). While there are considerable advances in the MCMC-based Bayesian variable selection (Narisetty et al. 2018; Johndrow et al. 2020), the existing MCMC-based algorithms still require a *sufficiently large number of iterations* to ensure the convergence to the stationary distribution. Recently, shrinkage priors such as the Bayesian lasso (Park and Casella 2008) and the horseshoe priors (Carvalho et al. 2009, 2010) have received considerable attention for Bayesian high-dimensional regression. However, they are not applicable to finding the mode of posterior model distribution, which is the main objective of our Bayesian model selection framework. To address the aforementioned limitations of existing Bayesian variable selection methods, in what follows, we propose a novel Bayesian variable selection scheme that allows fast *deterministic* searching for γ_{HPM}.

3 The HD-DVS approach

For high-dimensional Bayesian variable selection, a deterministic model search algorithm can be exploited to overcome the computational burden of stochastic search algorithms. However, it is well known that the deterministic search approach is limited to finding a locally optimal solution in the problem of nonconvex optimization. To address such fatal weakness of the deterministic search method, we propose hybrid deterministic–deterministic variable selection (HD-DVS) which identifies the *asymptotically* global optimum model. Our HD-DVS framework can be divided into two steps:

1. *Locally optimal model search* Find a local best model by minimizing $D(\cdot)$.
2. *Global optimality check* Investigate if there exists a better model than the locally optimal model found in Step 1.

For a fast and easy implementation of the first step of HD-DVS, we propose to use a greedy algorithm that finds a locally optimal solution by iteratively searching over the neighborhood of the current solution. In the context of variable selection, the neighborhood for the greedy search can be constructed by adding a new predictor to or deleting one from the current best model. Let $\hat{\gamma}$ be the current best model. Then, the neighborhood of $\hat{\gamma}$ is defined by $\mathcal{N}(\hat{\gamma}) = \mathcal{N}_+(\hat{\gamma}) \cup \mathcal{N}_-(\hat{\gamma})$ such that

$$\mathcal{N}_+(\hat{\gamma}) = \{\hat{\gamma} \cup \{j\} : j \notin \hat{\gamma}\} \quad \text{and} \quad \mathcal{N}_-(\hat{\gamma}) = \{\hat{\gamma} \backslash \{j\} : j \in \hat{\gamma}\}.$$

The greedy algorithm for searching the local best model proceeds by repeating the following two-step procedure:

(a) Compute $\tilde{\gamma} = \arg\min_{\gamma \in \mathcal{N}(\hat{\gamma})} D(\gamma)$ (the best choice at the moment).
(b) If $D(\tilde{\gamma}) < D(\hat{\gamma})$, then update $\hat{\gamma} \leftarrow \tilde{\gamma}$ and go to Step (a). Otherwise, return $\hat{\gamma}$.

After finding the local best model using the above greedy algorithm, the next step in HD-DVS is to test whether the result is globally optimal. It is worth noting that testing the global optimality of the local best model is computationally expensive or even infeasible due to the nonconvexity of our objective function $D(\cdot)$ as well as the high-dimensionality of the candidate model space. To address such issue, we introduce the following asymptotic property of our Bayesian model selection criterion $D(\cdot)$:

Theorem 1 *Consider the globally optimal model* $\gamma_{\text{HPM}} = \arg\min_{\gamma \in \mathcal{G}} D(\gamma)$. *Suppose that* $\gamma_1 \in \mathcal{G}$ *and* $\gamma_2 \in \mathcal{G}$ *are two arbitrary candidate models. Under the regularity conditions ofChen and Chen* (2008), *which include* $p = O(n^\kappa)$ *for a constant* $\kappa > 0$, *we have the following results*:

a. If $\gamma_{\text{HPM}} \subsetneq \gamma_1 \subsetneq \gamma_2$, *then*

$$D(\gamma_{\text{HPM}}) < D(\gamma_1) < D(\gamma_2)$$

in probability as $n \to \infty$.

b. If $\gamma_{\text{HPM}} \subset \gamma_1$ *and* $\gamma_{\text{HPM}} \not\subset \gamma_2$, *then*

$$D(\gamma_{\text{HPM}}) < D(\gamma_1) < D(\gamma_2)$$

in probability as $n \to \infty$.

In Theorem 1, the convergence in probability is with respect to the true data generating distribution, and this argument will be implicitly used throughout the paper. The proof of Theorem 1 is given in Appendix A. By Theorem 1a, the local best model, $\hat{\gamma}$, returned from the greedy search step of HD-DVS must satisfy either (i) $\hat{\gamma} = \gamma_{\text{HPM}}$ or (ii) $\hat{\gamma} \not\supset \gamma_{\text{HPM}}$ for sufficiently large n. Then, by Theorem 1b, when n is sufficiently large, $\hat{\gamma}$ is locally optimal but not globally optimal if and only if there

exists $\tilde{\gamma}$ ($\supset \hat{\gamma}$) such that $D(\tilde{\gamma}) < D(\hat{\gamma})$. This motivates us to perform the following forward search algorithm for testing the global optimality:

(a) Set $\tilde{\gamma} = \hat{\gamma}$, where $\hat{\gamma}$ is the local best model from the greedy algorithm.
(b) (Forward search) Compute $\tilde{\gamma}_+ = \arg\min_{\gamma \in \mathcal{N}_+(\tilde{\gamma})} D(\gamma)$.

(c) (i) If $D(\tilde{\gamma}_+) > D(\hat{\gamma})$ and $\tilde{\gamma}_+ \in \mathcal{G}$, then update $\tilde{\gamma} \leftarrow \tilde{\gamma}_+$ and go to Step (b).
(ii) If $D(\tilde{\gamma}_+) \leq D(\hat{\gamma})$ and $\tilde{\gamma}_+ \in \mathcal{G}$, then update $\hat{\gamma} \leftarrow \tilde{\gamma}_+$ and return $R = 1$.
(iii) Otherwise (i.e., $\tilde{\gamma}_+ \notin \mathcal{G}$), return $R = 0$.

If the forward search algorithm returns $R = 0$, then we can conclude that $\hat{\gamma}$ is globally optimal. Hence, we terminate our HD-DVS method. Note that $R = 1$ means that $\hat{\gamma}$ was updated by a better model found within the forward search algorithm. In this case, we move back to the greedy algorithm starting with the updated $\hat{\gamma}$. This prevents us from being stuck in the local trap. The full algorithm of HD-DVS is given in Algorithm 1.

Algorithm 1 Hybrid Deterministic-Deterministic Variable Selection (HD-DVS)

Input: $\hat{\gamma}$ (initial model).

1. **Local best model search:**
   (a) Compute $\tilde{\gamma} = \arg\min_{\gamma \in \mathcal{N}(\hat{\gamma})} D(\gamma)$.
   (b) If $D(\tilde{\gamma}) < D(\hat{\gamma})$, then update $\hat{\gamma} \leftarrow \tilde{\gamma}$ and go to Step 1(a).
2. **Global optimality check:**
   (a) Set $\tilde{\gamma} = \hat{\gamma}$.
   (b) Compute $\tilde{\gamma}_+ = \arg\min_{\gamma \in \mathcal{N}_+(\tilde{\gamma})} D(\gamma)$.
   (c) (i) If $D(\tilde{\gamma}_+) > D(\hat{\gamma})$ and $\tilde{\gamma}_+ \in \mathcal{G}$, then update $\tilde{\gamma} \leftarrow \tilde{\gamma}_+$ and go to Step 2(b).
   (ii) If $D(\tilde{\gamma}_+) \leq D(\hat{\gamma})$ and $\tilde{\gamma}_+ \in \mathcal{G}$, then update $\hat{\gamma} \leftarrow \tilde{\gamma}_+$ and set $R = 1$.
   (iii) Otherwise (i.e., $\tilde{\gamma}_+ \notin \mathcal{G}$), set $R = 0$.
3. **If** $R = 1$, then go to Step 1. **Otherwise** (i.e., $R = 0$), terminate.

Output: $\hat{\gamma}$.

In Algorithm 1, the initial model and the candidate model space are chosen beforehand by a user based on his/her prior information or expertise. For example, in our simulation study and real data analysis, they are defined as $\hat{\gamma} = \emptyset$ (null model) and $\mathcal{G} = \{\gamma : p_\gamma < n^{2/3}\}$ assuming that we have no prior knowledge about important predictors.

After we obtain the HPM estimate, $\hat{\gamma}$, using HD-DVS, Bayesian inference about $\boldsymbol{\beta}_{\hat{\gamma}}$ and σ^2 is made by using the posterior distribution under the g-prior. Under the g-prior, the joint posterior of $\boldsymbol{\beta}_{\hat{\gamma}}$ and σ^2 given $\mathbf{y}$ and $\hat{\gamma}$ can be factorized as

$$
\begin{aligned}
p(\boldsymbol{\beta}_{\hat{\gamma}}, \sigma^2 \mid \mathbf{y}, \hat{\gamma}) &= p(\boldsymbol{\beta}_{\hat{\gamma}} \mid \mathbf{y}, \sigma^2, \hat{\gamma}) p(\sigma^2 \mid \mathbf{y}, \hat{\gamma}) \\
&= N\left(\boldsymbol{\beta}_{\hat{\gamma}} \mid \frac{g}{1+g}\hat{\boldsymbol{\beta}}_{\hat{\gamma}}, \sigma^2 \frac{g}{1+g}(\mathbf{X}_{\hat{\gamma}}^{\mathrm{T}}\mathbf{X}_{\hat{\gamma}})^{-1}\right) \\
&\quad \times \text{Inv-Gamma}\left(\sigma^2 \mid \frac{n}{2}, \frac{1}{2}\left\{\|\mathbf{y}\|^2 - \frac{g}{1+g}\mathbf{y}^{\mathrm{T}}\mathbf{P}_{\hat{\gamma}}\mathbf{y}\right\}\right),
\end{aligned}
$$

where $\hat{\boldsymbol{\beta}}_{\hat{\gamma}} = (\mathbf{X}_{\hat{\gamma}}^{\mathrm{T}}\mathbf{X}_{\hat{\gamma}})^{-1}\mathbf{X}_{\hat{\gamma}}^{\mathrm{T}}\mathbf{y}$. Hence, Bayesian inference can be performed by generating samples from the inverse gamma posterior $p(\sigma^2 \mid \mathbf{y}, \hat{\gamma})$ and the normal posterior $p(\boldsymbol{\beta}_{\hat{\gamma}} \mid \mathbf{y}, \sigma^2, \hat{\gamma})$, sequentially.

One of the great merits of HD-DVS is the use of a *fully deterministic* search scheme, which does not suffer from the slow convergence issues that are inherent to existing stochastic search methods. Another important feature of HD-DVS is that the global best model (i.e., HPM) can be consistently found as follows:

Theorem 2 *Let $\hat{\gamma}$ be the best model returned from HD-DVS algorithm* (*Algorithm* 1). *Under the regularity conditions of Theorem* 1 *andWang* (2009), *we have*

$$D(\hat{\gamma}) = \min_{\gamma \in \mathcal{G}} D(\gamma)$$

in probability as $n \to \infty$.

The proof of Theorem 2 is shown in Appendix B. While the implementation of HD-DVS algorithm is straightforward, there is a computational bottleneck when p is very large. Specifically, in Step 1(a) of Algorithm 1, minimizing $D(\gamma)$ over $\gamma \in \mathcal{N}(\hat{\gamma})$ tends to be demanding as p increases. Similarly, the computational cost of Step 2(b) grows dramatically as p grows. To alleviate such scalability issues of HD-DVS, we propose a fast implementation strategy for HD-DVS in the following section.

4 Fast HD-DVS

4.1 Fast implementation of local best model search step

From (7), it is easy to see that when we evaluate $D(\gamma)$ for each $\gamma \in \mathcal{N}(\hat{\gamma})$, the largest burden occurs in the calculation of the inverse matrix $(\mathbf{X}_{\gamma}^{\mathrm{T}}\mathbf{X}_{\gamma})^{-1}$. Since the number of models in $\mathcal{N}(\hat{\gamma})$ is always p regardless of the size of $\hat{\gamma}$, the computational cost of Step 1(a) of Algorithm 1 could be prohibitively expensive when p is large. To reduce the computational complexity, we develop a fast implementation strategy for the local best model search in HD-DVS using the notion of the matrix ridge approximation (Zhang 2014). The key idea of our approach is to replace $(\mathbf{X}_{\gamma}^{\mathrm{T}}\mathbf{X}_{\gamma})^{-1}$ with $(\mathbf{X}_{\gamma}^{\mathrm{T}}\mathbf{X}_{\gamma} + \delta\mathbf{I}_{p_{\gamma}})^{-1}$ for a small positive value δ. We therefore define a substitution for $D(\cdot)$ as follows:

$$D_\delta(\gamma) = n \log\left\{ \|\mathbf{y}\|^2 - \frac{n}{1+n}\mathbf{y}^{\mathrm{T}}\mathbf{P}_{\gamma,\delta}\mathbf{y} \right\} + p_\gamma \log(n+1) + 2\log\begin{pmatrix} p \\ p_\gamma \end{pmatrix}, \tag{8}$$

where $\mathbf{P}_{\gamma,\delta} = \mathbf{X}_\gamma(\mathbf{X}_\gamma^{\mathrm{T}}\mathbf{X}_\gamma + \delta\mathbf{I}_{p_\gamma})^{-1}\mathbf{X}_\gamma^{\mathrm{T}}$. The following lemma shows an important relationship between $D_\delta(\cdot)$ and $D(\cdot)$.

Lemma 1 *Let γ_1 and γ_2 be two candidate models such that $D(\gamma_1) < D(\gamma_2)$. Then there always exists a $\delta_0 > 0$ such that*

$$D(\gamma_1) < D(\gamma_2) \quad \Leftrightarrow \quad D_\delta(\gamma_1) < D_\delta(\gamma_2)$$

for any $\delta \in (0, \delta_0]$.

Proof of Lemma 1 Since $\mathbf{y}^{\mathrm{T}}\mathbf{P}_{\gamma,\delta}\mathbf{y}$ is a continuous function of δ, the result is immediate from the fact that $\mathbf{y}^{\mathrm{T}}\mathbf{P}_{\gamma,\delta}\mathbf{y} \to \mathbf{y}^{\mathrm{T}}\mathbf{P}_\gamma\mathbf{y}$ as $\delta \to 0$, where $\mathbf{P}_\gamma = \mathbf{X}_\gamma(\mathbf{X}_\gamma^{\mathrm{T}}\mathbf{X}_\gamma)^{-1}\mathbf{X}_\gamma^{\mathrm{T}}$. □

Suppose that γ_1 and γ_2 are two models in $\mathcal{N}_+(\hat{\gamma}) = \{\hat{\gamma} \cup \{j\} : j \notin \hat{\gamma}\}$. From the fact that $p_{\gamma_1} = p_{\gamma_2}$, Lemma 1 leads to the following equivalence relations:

$$D(\gamma_1) < D(\gamma_2) \Leftrightarrow D_\delta(\gamma_1) < D_\delta(\gamma_2) \Leftrightarrow \mathbf{y}^{\mathrm{T}}\mathbf{P}_{\gamma_1,\delta}\mathbf{y} > \mathbf{y}^{\mathrm{T}}\mathbf{P}_{\gamma_2,\delta}\mathbf{y} \tag{9}$$

for some $\delta > 0$. Without loss of generality, let $\gamma_1 = \hat{\gamma} \cup \{j_1\}$ and $\gamma_2 = \hat{\gamma} \cup \{j_2\}$ for $j_1, j_2 \notin \hat{\gamma}$. Using the push-through identity and the Sherman–Morrison formula, it can be shown that

$$\begin{aligned}
\mathbf{I}_n - \mathbf{P}_{\gamma_k,\delta} &= \delta(\mathbf{X}_{\gamma_k}\mathbf{X}_{\gamma_k}^{\mathrm{T}} + \delta\mathbf{I}_n)^{-1} \\
&= \delta\left\{ (\mathbf{X}_{\hat{\gamma}}\mathbf{X}_{\hat{\gamma}}^{\mathrm{T}} + \delta\mathbf{I}_n)^{-1} - \frac{(\mathbf{X}_{\hat{\gamma}}\mathbf{X}_{\hat{\gamma}}^{\mathrm{T}} + \delta\mathbf{I}_n)^{-1}\mathbf{x}_{j_k}\mathbf{x}_{j_k}^{\mathrm{T}}(\mathbf{X}_{\hat{\gamma}}\mathbf{X}_{\hat{\gamma}}^{\mathrm{T}} + \delta\mathbf{I}_n)^{-1}}{1 + \mathbf{x}_{j_k}^{\mathrm{T}}(\mathbf{X}_{\hat{\gamma}}\mathbf{X}_{\hat{\gamma}}^{\mathrm{T}} + \delta\mathbf{I}_n)^{-1}\mathbf{x}_{j_k}} \right\} \\
&= (\mathbf{I}_n - \mathbf{P}_{\hat{\gamma},\delta}) - \frac{(\mathbf{I}_n - \mathbf{P}_{\hat{\gamma},\delta})\mathbf{x}_{j_k}\mathbf{x}_{j_k}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_{\hat{\gamma},\delta})}{\delta + \mathbf{x}_{j_k}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_{\hat{\gamma},\delta})\mathbf{x}_{j_k}}
\end{aligned}$$

for $k = 1, 2$. From (9), this implies that

$$D(\gamma_1) < D(\gamma_2) \Leftrightarrow \frac{\{\mathbf{y}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_{\hat{\gamma},\delta})\mathbf{x}_{j_1}\}^2}{\delta + \mathbf{x}_{j_1}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_{\hat{\gamma},\delta})\mathbf{x}_{j_1}} > \frac{\{\mathbf{y}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_{\hat{\gamma},\delta})\mathbf{x}_{j_2}\}^2}{\delta + \mathbf{x}_{j_2}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_{\hat{\gamma},\delta})\mathbf{x}_{j_2}}.$$

Since the above relation holds for any arbitrary γ_1 and γ_2 in $\mathcal{N}_+(\hat{\gamma})$, we can generalize it to find the solution to $\min_{\gamma \in \mathcal{N}_+(\hat{\gamma})} D(\gamma)$ as follows:

$$\min_{\gamma \in \mathcal{N}_+(\hat{\gamma})} D(\gamma) \Leftrightarrow \min_{\gamma \in \mathcal{N}_+(\hat{\gamma})} D_\delta(\gamma) \Leftrightarrow \max_{j : j \notin \hat{\gamma}} \frac{(\mathbf{y}^{\mathrm{T}}\mathbf{P}^{\perp}_{\hat{\gamma},\delta}\mathbf{x}_j)^2}{\mathbf{x}_j^{\mathrm{T}}\mathbf{P}^{\perp}_{\hat{\gamma},\delta}\mathbf{x}_j + \delta}, \tag{10}$$

where $\mathbf{P}^{\perp}_{\hat{\gamma},\delta} = \mathbf{I}_n - \mathbf{P}_{\hat{\gamma},\delta}$. Following a similar technique, it is straightforward to show that

$$\min_{\gamma \in \mathcal{N}_-(\hat{\gamma})} D(\gamma) \Leftrightarrow \min_{\gamma \in \mathcal{N}_-(\hat{\gamma})} D_\delta(\gamma) \Leftrightarrow \max_{j:j\in\hat{\gamma}} \frac{(\mathbf{y}^{\mathrm{T}}\mathbf{P}^{\perp}_{\hat{\gamma},\delta}\mathbf{x}_j)^2}{\mathbf{x}_j^{\mathrm{T}}\mathbf{P}^{\perp}_{\hat{\gamma},\delta}\mathbf{x}_j - \delta}. \tag{11}$$

If $\mathbf{A}_{\hat{\gamma},\delta} = (\mathbf{X}^{\mathrm{T}}_{\hat{\gamma}}\mathbf{X}_{\hat{\gamma}} + \delta\mathbf{I}_{p_{\hat{\gamma}}})^{-1}$ is readily available, then $\mathbf{y}^{\mathrm{T}}\mathbf{P}^{\perp}_{\hat{\gamma},\delta}\mathbf{x}_j$ and $\mathbf{x}_j^{\mathrm{T}}\mathbf{P}^{\perp}_{\hat{\gamma},\delta}\mathbf{x}_j$ can be quickly and efficiently computed in the following ways:

$$\mathbf{y}^{\mathrm{T}}\mathbf{P}^{\perp}_{\hat{\gamma},\delta}\mathbf{x}_j = \mathbf{y}^{\mathrm{T}}\mathbf{x}_j - \mathbf{y}^{\mathrm{T}}\mathbf{X}_{\hat{\gamma}}\mathbf{A}_{\hat{\gamma},\delta}\mathbf{X}^{\mathrm{T}}_{\hat{\gamma}}\mathbf{x}_j,$$
$$\mathbf{x}_j^{\mathrm{T}}\mathbf{P}^{\perp}_{\hat{\gamma},\delta}\mathbf{x}_j = \|\mathbf{x}_j\|^2 - \mathbf{x}_j^{\mathrm{T}}\mathbf{X}_{\hat{\gamma}}\mathbf{A}_{\hat{\gamma},\delta}\mathbf{X}^{\mathrm{T}}_{\hat{\gamma}}\mathbf{x}_j.$$

In the next lemma, we introduce an efficient way of precomputing $\mathbf{A}_{\hat{\gamma},\delta}$ in our local best model search framework.

Lemma 2 *Let*

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}^{*-1} & \mathbf{b} \\ \mathbf{b}^{\mathrm{T}} & c \end{bmatrix},$$

where $\mathbf{A}^*$ *is a* $d \times d$ *matrix*, $\mathbf{b}$ *is a* $d \times 1$ *vector*, *and* c *is a scalar.*

a. If $\mathbf{A}^*$, $\mathbf{b}$, *and* c *are given*, *then* $\mathbf{A}$ *can be computed by*

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}^* + c^*\mathbf{A}^*\mathbf{b}\mathbf{b}^{\mathrm{T}}\mathbf{A}^* & -c^*\mathbf{A}^*\mathbf{b} \\ -c^*\mathbf{b}^{\mathrm{T}}\mathbf{A}^* & c^* \end{bmatrix},$$

where $c^* = 1/(c - \mathbf{b}^{\mathrm{T}}\mathbf{A}^*\mathbf{b})$.

b. If $\mathbf{A}$, $\mathbf{b}$, *and* c *are given*, *then* $\mathbf{A}^*$ *can be computed by*

$$\mathbf{A}^* = \mathbf{A}_{11} - \mathbf{a}_{12}\mathbf{a}_{12}^{\mathrm{T}}/a_{22},$$

where $\mathbf{A}_{11}$, $\mathbf{a}_{12}$, *and* a_{22} *are the* $d \times d$ *matrix*, $d \times 1$ *vector, and scalar, respectively, such that*

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{a}_{12} \\ \mathbf{a}_{12}^{\mathrm{T}} & a_{22} \end{bmatrix}.$$

Proof of Lemma 2 The results are obtained by the standard 2×2 block matrix inverse formula (e.g., Lu and Shiou 2002). □

Using (10) and (11) with Lemma 2, the fast implementation of the local best model search step in HD-DVS can be achieved by modifying Step 1 of Algorithm 1 as follows:

1'. *Fast local best model search*:

(a) Define $\tilde{\gamma}_+ = \hat{\gamma} \cup \{j_+\}$ such that

$$j_+ = \arg\max_{j:j\notin\hat{\gamma}} \frac{(\mathbf{y}^{\mathrm{T}}\mathbf{x}_j - \mathbf{y}^{\mathrm{T}}\mathbf{X}_{\hat{\gamma}}\mathbf{A}_{\hat{\gamma},\delta}\mathbf{X}_{\hat{\gamma}}^{\mathrm{T}}\mathbf{x}_j)^2}{\|\mathbf{x}_j\|^2 - \mathbf{x}_j^{\mathrm{T}}\mathbf{X}_{\hat{\gamma}}\mathbf{A}_{\hat{\gamma},\delta}\mathbf{X}_{\hat{\gamma}}^{\mathrm{T}}\mathbf{x}_j + \delta}, \tag{12}$$

and compute

$$\mathbf{A}_{\tilde{\gamma}_+,\delta} = \begin{bmatrix} \mathbf{A}_{\hat{\gamma},\delta} + c^*\mathbf{A}_{\hat{\gamma},\delta}\mathbf{b}\mathbf{b}^{\mathrm{T}}\mathbf{A}_{\hat{\gamma},\delta} & -c^*\mathbf{A}_{\hat{\gamma},\delta}\mathbf{b} \\ -c^*\mathbf{b}^{\mathrm{T}}\mathbf{A}_{\hat{\gamma},\delta} & c^* \end{bmatrix},$$

where $\mathbf{b} = \mathbf{X}_{\hat{\gamma}}^{\mathrm{T}}\mathbf{x}_{j_+}$ and $c^* = 1/(\|\mathbf{x}_{j_+}\|^2 + \delta - \mathbf{b}^{\mathrm{T}}\mathbf{A}_{\hat{\gamma},\delta}\mathbf{b})$.

(b) Define $\tilde{\gamma}_- = \hat{\gamma}\backslash\{j_-\}$ such that

$$j_- = \arg\max_{j:j\in\hat{\gamma}} \frac{(\mathbf{y}^{\mathrm{T}}\mathbf{x}_j - \mathbf{y}^{\mathrm{T}}\mathbf{X}_{\hat{\gamma}}\mathbf{A}_{\hat{\gamma},\delta}\mathbf{X}_{\hat{\gamma}}^{\mathrm{T}}\mathbf{x}_j)^2}{\|\mathbf{x}_j\|^2 - \mathbf{x}_j^{\mathrm{T}}\mathbf{X}_{\hat{\gamma}}\mathbf{A}_{\hat{\gamma},\delta}\mathbf{X}_{\hat{\gamma}}^{\mathrm{T}}\mathbf{x}_j - \delta}, \tag{13}$$

and compute

$$\mathbf{A}_{\tilde{\gamma}_-,\delta} = \mathbf{A}_{11} - \mathbf{a}_{12}\mathbf{a}_{12}^{\mathrm{T}}/a_{22},$$

where $\mathbf{A}_{11}$, $\mathbf{a}_{12}$ and a_{22} are obtained from $\mathbf{A}_{\hat{\gamma},\delta}$ accordingly.

(c) If $D_\delta(\tilde{\gamma}_+) < D_\delta(\tilde{\gamma}_-)$, then set $\tilde{\gamma} = \tilde{\gamma}_+$ and $\mathbf{A}_{\tilde{\gamma},\delta} = \mathbf{A}_{\tilde{\gamma}_+,\delta}$.
Otherwise, set $\tilde{\gamma} = \tilde{\gamma}_-$ and $\mathbf{A}_{\tilde{\gamma},\delta} = \mathbf{A}_{\tilde{\gamma}_-,\delta}$.

(d) If $D_\delta(\tilde{\gamma}) < D_\delta(\hat{\gamma})$, then update $\hat{\gamma} \leftarrow \tilde{\gamma}$ and $\mathbf{A}_{\hat{\gamma},\delta} \leftarrow \mathbf{A}_{\tilde{\gamma},\delta}$, and go to Step 1'(a).

Note that $\mathbf{A}_{\hat{\gamma},\delta}$ is obtained from the previous iteration by Step 1'(d). Hence, the proposed fast search scheme can be implemented very quickly and efficiently even when p is large.

4.2 Fast implementation of global optimality check step

As mentioned earlier, the global optimality check step of HD-DVS also poses the heavy burden of inverse matrix computation due to the iterative implementation of Step 2(b) in Algorithm 1. Following the proposed idea in Sect. 4.1, we use $D_\delta(\cdot)$ in (8) as a substitute for $D(\cdot)$. Then, by (10), the fast implementation of the

global optimality check step can be achieved by replacing Step 2 of Algorithm 1 as follows:

2'. *Fast global optimality check*:

(a) Set $\tilde{\gamma} = \hat{\gamma}$ (and $\mathbf{A}_{\tilde{\gamma},\delta} = \mathbf{A}_{\hat{\gamma},\delta}$).
(b) Define $\tilde{\gamma}_+ = \hat{\gamma} \cup \{j_+\}$ such that

$$j_+ = \arg\max_{j:j\notin\tilde{\gamma}} \frac{(\mathbf{y}^{\mathrm{T}}\mathbf{x}_j - \mathbf{y}^{\mathrm{T}}\mathbf{X}_{\tilde{\gamma}}\mathbf{A}_{\tilde{\gamma},\delta}\mathbf{X}_{\tilde{\gamma}}^{\mathrm{T}}\mathbf{x}_j)^2}{\|\mathbf{x}_j\|^2 - \mathbf{x}_j^{\mathrm{T}}\mathbf{X}_{\tilde{\gamma}}\mathbf{A}_{\tilde{\gamma},\delta}\mathbf{X}_{\tilde{\gamma}}^{\mathrm{T}}\mathbf{x}_j + \delta},$$

and compute

$$\mathbf{A}_{\tilde{\gamma}_+,\delta} = \begin{bmatrix} \mathbf{A}_{\tilde{\gamma},\delta} + c^*\mathbf{A}_{\tilde{\gamma},\delta}\mathbf{b}\mathbf{b}^{\mathrm{T}}\mathbf{A}_{\tilde{\gamma},\delta} & -c^*\mathbf{A}_{\tilde{\gamma},\delta}\mathbf{b} \\ -c^*\mathbf{b}^{\mathrm{T}}\mathbf{A}_{\tilde{\gamma},\delta} & c^* \end{bmatrix},$$

where $\mathbf{b} = \mathbf{X}_{\tilde{\gamma}}^{\mathrm{T}}\mathbf{x}_{j_+}$ and $c^* = 1/(\|\mathbf{x}_{j_+}\|^2 + \delta - \mathbf{b}^{\mathrm{T}}\mathbf{A}_{\tilde{\gamma},\delta}\mathbf{b})$.

(c)
(i) If $D_\delta(\tilde{\gamma}_+) > D_\delta(\hat{\gamma})$ and $\tilde{\gamma}_+ \in \mathcal{G}$, then update $\tilde{\gamma} \leftarrow \tilde{\gamma}_+$ (and $\mathbf{A}_{\tilde{\gamma},\delta} \leftarrow \mathbf{A}_{\tilde{\gamma}_+,\delta}$) and go to Step 2'(b).
(ii) If $D_\delta(\hat{\gamma}_+) \le D_\delta(\hat{\gamma})$ and $\tilde{\gamma}_+ \in \mathcal{G}$, then update $\hat{\gamma} \leftarrow \tilde{\gamma}_+$ (and $\mathbf{A}_{\hat{\gamma},\delta} \leftarrow \mathbf{A}_{\tilde{\gamma}_+,\delta}$) and set $R = 1$.
(iii) Otherwise (i.e., $\tilde{\gamma}_+ \notin \mathcal{G}$), set $R = 0$.

By removing the inverse matrix computation throughout all the iterations, the proposed fast algorithm indeed leads to a dramatic reduction in execution time.

4.3 Choice of δ

The use of our fast HD-DVS algorithm requires choosing a small value of δ to obtain the same result as that obtained using the original HD-DVS algorithm in Algorithm 1. While Lemma 1 ensures the existence of such δ values, a guideline for the choice of δ should be established for practical purposes. From the fact that $D_\delta(\gamma)$ is obtained from $D(\gamma)$ by replacing $\mathbf{y}^{\mathrm{T}}\mathbf{P}_\gamma\mathbf{y}$ with $\mathbf{y}^{\mathrm{T}}\mathbf{P}_{\gamma,\delta}\mathbf{y}$, we find a valid value of δ by determining a mild condition for having $\mathbf{y}^{\mathrm{T}}\mathbf{P}_\gamma\mathbf{y} - \mathbf{y}^{\mathrm{T}}\mathbf{P}_{\gamma,\delta}\mathbf{y} = o_p(1)$ as $n \to \infty$.

Using the singular value decomposition, we decompose $\mathbf{X}_\gamma = \mathbf{U}_\gamma\mathbf{D}_\gamma\mathbf{V}_\gamma^{\mathrm{T}}$ such that $\mathbf{U}_\gamma^{\mathrm{T}}\mathbf{U}_\gamma = \mathbf{I}_{p_\gamma}$, $\mathbf{V}_\gamma^{\mathrm{T}}\mathbf{V}_\gamma = \mathbf{V}_\gamma\mathbf{V}_\gamma^{\mathrm{T}} = \mathbf{I}_{p_\gamma}$, and $\mathbf{D}_\gamma = \mathrm{Diag}(d_1, \ldots, d_{p_\gamma})$. Then, $\mathbf{P}_{\gamma,\delta}$ can be written as

$$\begin{aligned}\mathbf{P}_{\gamma,\delta} &= \mathbf{X}_\gamma(\mathbf{X}_\gamma^{\mathrm{T}}\mathbf{X}_\gamma + \delta\mathbf{I}_{p_\gamma})^{-1}\mathbf{X}_\gamma^{\mathrm{T}} \\ &= \mathbf{U}_\gamma\mathbf{D}_\gamma\mathbf{V}_\gamma^{\mathrm{T}}\{\mathbf{V}_\gamma(\mathbf{D}_\gamma^2 + \delta\mathbf{I}_{p_\gamma})\mathbf{V}_\gamma^{\mathrm{T}}\}^{-1}\mathbf{V}_\gamma\mathbf{D}_\gamma\mathbf{U}_\gamma^{\mathrm{T}} \\ &= \mathbf{U}_\gamma \mathrm{Diag}\left(\frac{d_1^2}{d_1^2+\delta}, \ldots, \frac{d_{p_\gamma}^2}{d_{p_\gamma}^2+\delta}\right)\mathbf{U}_\gamma^{\mathrm{T}}.\end{aligned}$$

From the fact that $\mathbf{P}_\gamma = \mathbf{U}_\gamma\mathbf{U}_\gamma^{\mathrm{T}}$, it follows that

$$\mathbf{y}^{\mathrm{T}}\mathbf{P}_{\gamma,\delta}\mathbf{y} = \sum_{j=1}^{p_\gamma}(\mathbf{y}^{\mathrm{T}}\mathbf{u}_j)^2\frac{d_j^2}{d_j^2+\delta} \le \sum_{j=1}^{p_\gamma}(\mathbf{y}^{\mathrm{T}}\mathbf{u}_j)^2 = \mathbf{y}^{\mathrm{T}}\mathbf{P}_\gamma\mathbf{y},$$

where $\mathbf{u}_j$ is the j-th column of $\mathbf{U}_\gamma$. This implies that

$$0 \le \mathbf{y}^{\mathrm{T}}\mathbf{P}_\gamma\mathbf{y} - \mathbf{y}^{\mathrm{T}}\mathbf{P}_{\gamma,\delta}\mathbf{y} = \sum_{j=1}^{p_\gamma}(\mathbf{y}^{\mathrm{T}}\mathbf{u}_j)^2\frac{\delta}{d_j^2+\delta} \le \frac{\|\mathbf{y}\|^2 p_\gamma}{d_{\min}^2}\delta,$$

where $d_{\min}^2 = \min_{1\le j\le p_\gamma} d_j^2$ and the last inequality is obtained from the Cauchy–Schwarz inequality. It is typical to assume that $\|\mathbf{y}\|^2/n = O_p(1)$ and $d_{\min}^2/n = O_p(1)$. Hence, a sufficient condition for having $\mathbf{y}^{\mathrm{T}}\mathbf{P}_\gamma\mathbf{y} - \mathbf{y}^{\mathrm{T}}\mathbf{P}_{\gamma,\delta}\mathbf{y} = o_p(1)$ is $\delta = o(p_\gamma^{-1})$. Since $p_\gamma \ll \max\{n,p\}$ in general, we propose to set $\delta = 1/\max\{n,p\}$ as a default choice.

5 Simulation study

In this section, we conduct a simulation study to investigate the finite-sample performance of our proposed method. We generate a dataset $\{(y_i, x_{i1}, \ldots, x_{ip}) : i = 1, \ldots, n\}$ from a high-dimensional sparse linear regression model,

$$y_i = \beta_0 + \sum_{j=1}^{p}\beta_j x_{ij} + \epsilon_i,$$

where $\epsilon_i \overset{iid}{\sim} N(0,3)$, $\beta_0 = 1$, $\beta_1 = \cdots = \beta_7 = 2$, $\beta_8 = \cdots = \beta_p = 0$, and $(x_{i1}, \ldots, x_{ip}) \overset{iid}{\sim} N_p(\mathbf{0}, \mathbf{\Sigma}_x)$ with $\mathbf{\Sigma}_x = (0.5^{|i-j|})_{p\times p}$. While we fix the sample size to be $n = 100$, we consider three values of $p \in \{1000, 3000, 5000\}$. For each value of p, we repeat 1000 Monte Carlo simulations. Among the p predictors available in the simulated data, the first seven are relevant and the rest of them are redundant.

For each simulated data set, we compare the following six deterministic methods for variable selection:

1. HD-DVS: The hybrid deterministic–deterministic variable selection method, which is proposed in Sect. 3 and implemented by Algorithm 1.
2. Fast HD-DVS: The fast implementation method of HD-DVS, which is proposed in Sect. 4.
3. Fast DVS: The fast deterministic variable selection method, which is performed by the fast local best model search step given in Sect. 4.1.
4. EMVS: The EM approach to Bayesian variable selection under the spike-and-slab normal mixture prior (Ročková and George 2014).
5. SCAD: The smoothly clipped absolute deviation method (Fan and Li 2001).
6. LASSO: The least absolute shrinkage and selection operator method (Tibshirani 1996).

The EMVS method is implemented using the `R` package `EMVS` with the default setting of Rocková and Moran (2021). The SCAD method is implemented using the `R` package `ncvreg`, where the tuning parameter is chosen by 10-fold cross-validation using the `cv.ncvreg` function. The LASSO method is performed by the `R` package `glmnet`, where the tuning parameter is determined by 10-fold cross-validation with the `cv.glmnet` function.

To evaluate the model selection performance, we compute the number of correctly selected active predictors (TPN, true positive number), the number of correctly excluded inactive predictors (TNN, true negative number), and the proportion of selecting the exact true model (CFR, correctly fitted ratio) over 1000 Monte Carlo replications. For example, if we always obtain $\hat{\gamma} = \{1, 2, \ldots, 8\}$ over 1000 replications in the case of $p = 1000$, then we have TPN $= 7$ and TNN $= 992$, whereas CFR $= 0$. The computational efficiency is assessed by measuring the execution time of each method in `R` using a Windows-based desktop computer with an Intel Core i5-7200 processor and 8 gigabytes of memory. In addition, we compute the predictive mean squared error (PMSE) to assess the performance for future prediction. In order to perform a fair comparison between Bayesian and penalized likelihood approaches, we define the PMSE in the following way based on the sparse Ordinary Least Squares (OLS) estimate, which consists of the OLS estimate for the set of the selected predictors and zeros for the excluded predictors:

$$\text{PMSE} = \frac{1}{n_{\text{new}}} \sum_{i=n+1}^{n+n_{\text{new}}} (y_i - \mathbf{x}_i^{\text{T}} \hat{\boldsymbol{\beta}})^2,$$

where $\{(y_i, \mathbf{x}_i) : i = n+1, \ldots, n+n_{\text{new}}\}$ is a newly generated dataset of size $n_{\text{new}} = 100$ and $\hat{\boldsymbol{\beta}}$ is the sparse OLS estimate obtained under each selected model. We use the OLS estimate in order to make a fair comparison between the Bayesian methods and the frequentist methods. Under the g-prior, the posterior mean of $\boldsymbol{\beta}_\gamma$ is obtained as

Table 1 Simulation results for model selection performance and computational efficiency over 1000 Monte Carlo replications

Case	Method	TPN	TNN	CFR	Time (sec)
p = 1000	HD-DVS	7	992.959	0.961	0.858
	Fast HD-DVS	7	992.959	0.961	0.050
	Fast DVS	6.977	992.960	0.956	0.043
	EMVS	6.990	992.995	0.988	0.426
	SCAD	7	981.943	0.006	0.259
	LASSO	7	989.127	0.285	0.119
p = 3000	HD-DVS	6.996	2992.968	0.970	2.865
	Fast HD-DVS	6.996	2992.968	0.970	0.451
	Fast DVS	6.913	2992.975	0.949	0.396
	EMVS	6.667	2992.996	0.792	1.422
	SCAD	6.980	2974.588	0.002	0.787
	LASSO	7	2988.509	0.226	0.265
p = 5000	HD-DVS	6.979	4992.970	0.967	5.207
	Fast HD-DVS	6.979	4992.970	0.967	1.212
	Fast DVS	6.880	4992.969	0.937	1.179
	EMVS	5.820	4992.999	0.415	2.789
	SCAD	6.960	4970.880	0	1.227
	LASSO	7	4988.229	0.198	0.423

$$E(\boldsymbol{\beta}_\gamma \mid \mathbf{y}, \gamma) = E\big\{E(\boldsymbol{\beta}_\gamma \mid \sigma^2, \mathbf{y}, \gamma) \mid \mathbf{y}, \gamma\big\} = \frac{g}{1+g}\hat{\boldsymbol{\beta}}_\gamma,$$

where $\hat{\boldsymbol{\beta}}_\gamma = (\mathbf{X}_\gamma^{\mathrm{T}}\mathbf{X}_\gamma)^{-1}\mathbf{X}_\gamma^{\mathrm{T}}\mathbf{y}$ is the sparse OLS estimate under γ. As we assume $g = n$, the marginal posterior mean is approximately equal to the OLS estimate, that is,

$$E(\boldsymbol{\beta} \mid \mathbf{y}, \gamma) \approx \hat{\boldsymbol{\beta}}_\gamma$$

for a sufficiently large n.

Table 1 displays the averages of TPN, TNN, CFR and the execution time (in sec) over 1000 replications for $p = 1000, 3000, 5000$. From the comparison between HD-DVS and Fast HD-DVS, it is evident that our proposed fast implementation strategy for HD-DVS indeed results in a dramatic reduction in computing time without any loss of performance. The results for Fast HD-DVS and Fast DVS clearly show that the introduction of the global optimality check step improves variable selection performance by providing a way to jump out of a local trap. It is also important to note that the global optimality check can be run in much less than 0.05 seconds on average. When the data are generated with $p = 1000$ predictors, EMVS, which is the most well-known deterministic approach to Bayesian variable selection, outperforms the others in variable selection with respect to TNN and CFR. However, as p increases, the performance of

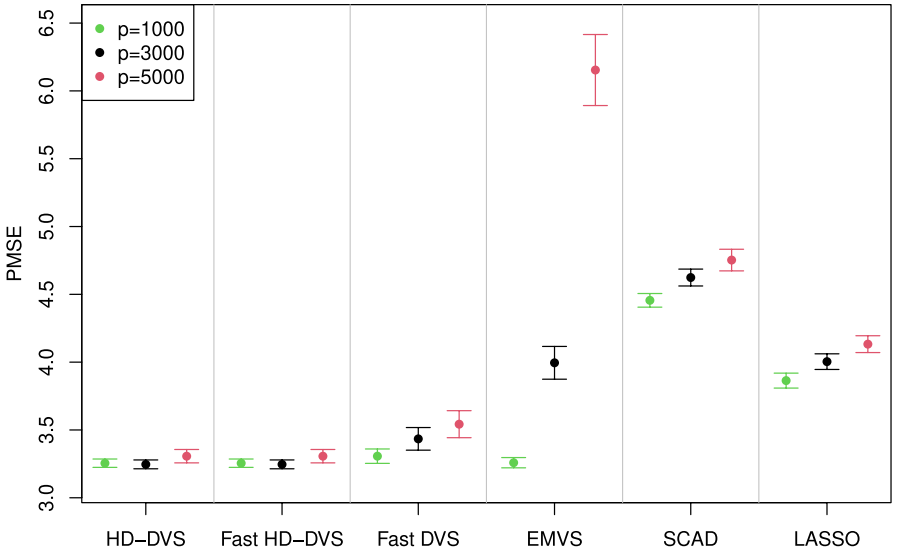

Fig. 1 Simulation results for predictive mean squared error (PMSE) over 1000 Monte Carlo replications, where each dot indicates the average PMSE and the error bar represents a 95% confidence interval

EMVS is remarkably degraded, especially with respect to TPN. In addition, the computational efficiency of EMVS is always worse than Fast HD-DVS, SCAD, and LASSO. Both penalized likelihood approaches, SCAD and LASSO, always provide poor performance in identifying the exact true model for all cases of p. It is worth noting that the execution time of Fast HD-DVS is always faster than that of SCAD.

Figure 1 shows 95% confidence intervals for PMSE from 1000 Monte Carlo experiments. Both HD-DVS and Fast HD-DVS provide identical results for future prediction. In addition, their prediction errors are low and relatively steady for increasing values of p. In contrast, as p increases, the prediction performance of EMVS deteriorates rapidly. It is also worth noting that SCAD and LASSO produce poor prediction performance even when $p = 1000$.

To summarize, our HD-DVS approach produces outstanding and reliable results in variable selection and prediction. In addition, the proposed fast implementation scheme permits us to run HD-DVS with low computational cost.

6 Real data application

In this section, we illustrate major drawbacks of a stochastic search approach for Bayesian variable selection by analyzing the Bardet–Biedl syndrome (BBS) gene expression data (Scheetz et al. 2006). In addition, we demonstrate that our Fast HD-DVS approach can serve as a great alternative to stochastic variable selection methods.

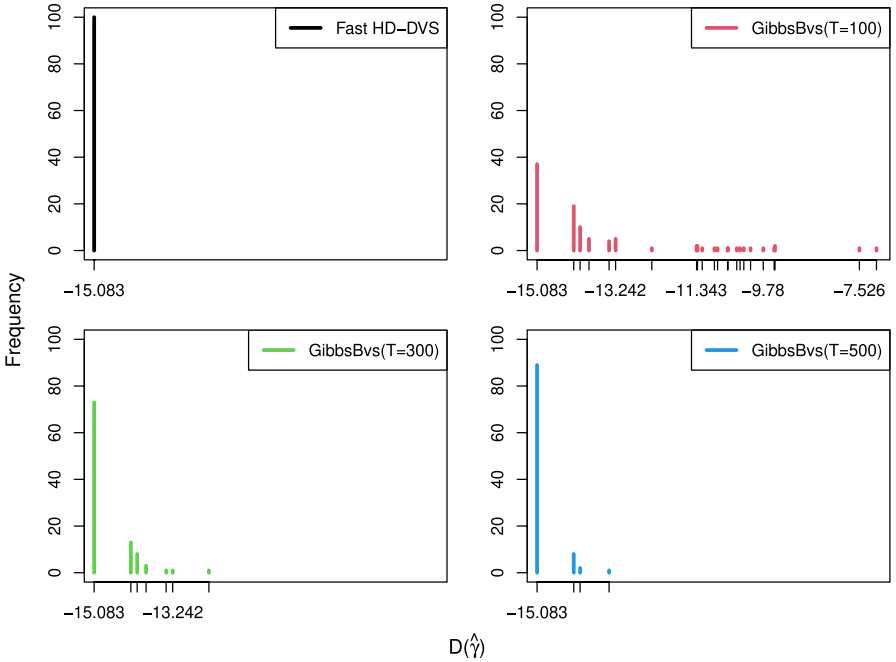

Fig. 2 A summary of BBS gene expression data analysis for HPM estimation over 100 replications, where the x-axis represents $D(\hat{\gamma})$ evaluated at the HPM estimate $\hat{\gamma}$ and the y-axis indicates the frequency of values of $D(\hat{\gamma})$

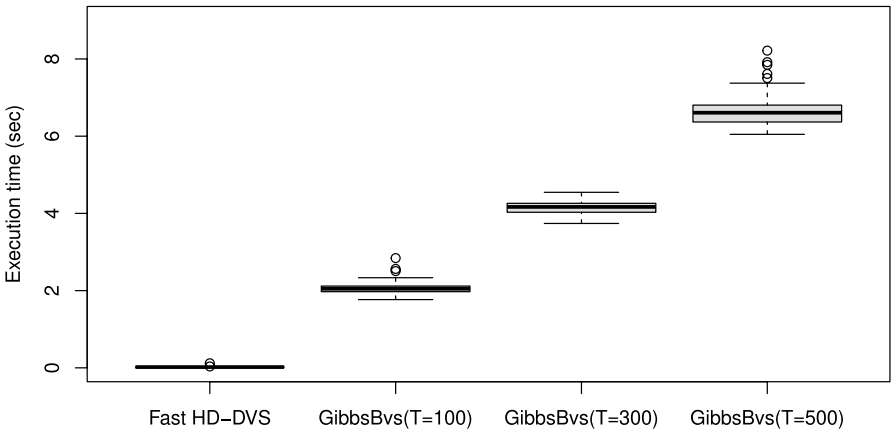

Fig. 3 A summary of BBS gene expression data analysis for execution times over 100 replications

The BBS gene expression data contain the expression levels of 501 genes, including the TRIM32 gene, collected from $n = 120$ mammalian eye tissue samples. The BBS gene expression data are available as `trim32` in the `R` package `abess`. The TRIM32 gene is well known to be linked to BBS. We are interested in

identifying significant genes that are related to the TRIM32 gene. To this end, we perform Bayesian variable selection in the framework of high-dimensional linear regression discussed in Sect. 2, where the TRIM32 gene serves as the response variable and the remaining $p = 500$ genes are the potential predictors. Recall that, in the Bayesian variable selection paradigm, the main goal is to estimate $\gamma_{\text{HPM}} = \arg\max_{\gamma} p(\gamma \mid \mathbf{y})$.

First, we consider a Gibbs sampling-based Bayesian variable selection method, called GibbsBvs. The GibbsBvs method performs a stochastic search over the model space via a Gibbs sampling algorithm. In our analysis, the GibbsBvs method is implemented using the `R` package `BayesVarSel`. Let T denote the number of Gibbs iterations performed after a burn-in period. Theoretically, it is not arguable that, as $T \to \infty$, the GibbsBvs algorithm always finds γ_{HPM}. In practice, however, the determination of T is an open question. To investigate the influence of T on GibbsBvs, we consider varying values of $T \in \{100, 300, 500\}$. For each value of T, we set the length of the burn-in period to be $0.2 \times T$. To account for the uncertainty associated with the stochastic search approach, we repeat the variable selection process of GibbsBvs 100 times for each value of T.

Second, we apply our Fast HD-DVS method to the BBS gene expression data for the sake of comparison. To measure the performance variability, the implementation of Fast HD-DVS is also repeated 100 times. Note that, while the computation time may vary with the 100 replicates, the result of Fast HD-DVS must be invariant to the repetition since it relies on a deterministic search scheme.

For both GibbsBvs and Fast HD-DVS, we use the same priors that are specified in Sect. 2. In the `BayesVarSel` package, the GibbsBvs method with the g-prior and the hierarchical uniform model prior can be implemented by using the `GibbsBvs` function with the options: `prior.model="ScottBerger"` and `prior.betas="gZellner"`.

To measure the performance of each method for finding γ_{HPM}, we compute $D(\hat{\gamma})$, which is equivalent to $-2 \log p(\hat{\gamma} \mid \mathbf{y})$, for each HPM estimate, $\hat{\gamma}$. Note that the model with the minimum value of D is preferred. The execution time is measured to evaluate the computational cost of each method. The analysis results are summarized in Figs. 2 and 3.

In Fig. 2, the first plot shows the D-values obtained from Fast HD-DVS over 100 replications. We always observe that $\hat{\gamma} = \{189, 209, 243\}$ is selected with $D(\hat{\gamma}) = -15.083$. The results of GibbsBvs for $T = 100, 300, 500$ are shown in the remaining three plots in Fig. 2. When $T = 100$ is used in GibbsBvs, only about 40% of 100 repetitions successfully find the smallest value of D. While the performance of GibbsBvs gets better as T increases, the chance of missing the minimum value of D is not trivial even when $T = 500$.

Figure 3 displays the boxplot of the execution times over 100 replications for each method. The average execution time of Fast HD-DVS is 0.021 seconds, which are approximately 10 times faster than the average speed of GibbsBvs with $T = 100$. The upward movement in the computational costs of GibbsBvs for increasing

T implies that the trade-off between efficiency and accuracy must be taken into consideration when we specify T. In contrast, the use of Fast HD-DVS enables us to improve both efficiency and accuracy simultaneously.

Figure 2 can also be used as an MCMC convergence diagnostic tool for the MCMC-based Bayesian variable selection method. It appears that the Gibbs sampler with $T = 500$ often fails to reach the convergence to the stationary distribution. This implies that $T = 500$ is not large enough for GibbsBvs to consistently identify the HPM. This convergence issue can be addressed by increasing T. However, in this case, GibbsBvs becomes less appealing as it is considerably more expensive when compared with our proposed Fast HD-DVS (Fig. 3).

The three genes (Probe Set ID: 1382223_at, 1388491_at, 1389910_at) that are selected in our analysis are known to be genetically important. For example, 1382223_at (called ZMYM4) has been identified as a novel MYB binding protein (Cibis et al. 2020). Hindmarch et al. (2008) reports that 1388491_at is similar to hypothetical protein FLJ20445. In Deng et al. (2016), 1389910_at is known to be one of transmembrane proteins. Hence, we believe that a further experiment with these genes will be helpful for estimating the connection to TRIM32.

7 Concluding remarks

In this paper, we have developed HD-DVS, a new deterministic approach for estimating the global mode of posterior model probabilities in the context of high-dimensional variable selection. To accelerate the speed of the HD-DVS algorithm dramatically, we have proposed a fast implementation strategy of HD-DVS using inverse matrix identities and matrix approximations. The simulation and real data analysis results suggest that our proposed HD-DVS approach holds great promise for high-dimensional Bayesian variable selection.

While our attention has been restricted to the use of Zellner's g-prior due to its long-standing popularity in Bayesian linear regression, our proposed idea is easily adapted to a variety of conjugate priors. In addition, the proposed method is immediately applicable to the use of different model priors for γ. An extension to binary response regression can also be achieved by using the data augmentation approach of Albert and Chib (1993), which is left for future research.

Despite its computational efficiency and reliability, the proposed HD-DVS approach poses some limitations. First, the success of HD-DVS relies on the large-sample theory. In other words, when the sample size is small, HD-DVS could fail to identify the global mode of the posterior model distribution. In such case, we can address the issue by incorporating the fast stochastic search-based global optimality check procedure of Jin and Goh (2021) into the HD-DVS algorithm. Second, the applicability of HD-DVS is limited to the task of selecting a single best model. As a result, HD-DVS is not applicable to the framework of Bayesian model averaging

(Raftery et al. 1997). Lastly, the execution time could grow when the initial model is chosen to be a large model since, in this case, our HD-DVS algorithm will start with backward elimination to remove irrelevant predictors. Nevertheless, the additional execution time is negligible due to the use of the fast implementation method described in Sect. 4.

Appendix A: Proof of Theorem 3.1

Let $\Sigma_{X_\gamma,y}$, $\Sigma_{X_\gamma,X_\gamma}$ and $\Sigma_{y,y}$ denote the probability limits of $n^{-1}\mathbf{X}_\gamma^{\mathrm{T}}\mathbf{y}$, $n^{-1}\mathbf{X}_\gamma^{\mathrm{T}}\mathbf{X}_\gamma$ and $n^{-1}\mathbf{y}^{\mathrm{T}}\mathbf{y}$, respectively. Note that

$$\frac{1}{n}\mathbf{y}^{\mathrm{T}}\mathbf{X}_\gamma(\mathbf{X}_\gamma^{\mathrm{T}}\mathbf{X}_\gamma)^{-1}\mathbf{X}_\gamma^{\mathrm{T}}\mathbf{y} = (n^{-1}\mathbf{y}^{\mathrm{T}}\mathbf{X}_\gamma)(n^{-1}\mathbf{X}_\gamma^{\mathrm{T}}\mathbf{X}_\gamma)^{-1}(n^{-1}\mathbf{X}_\gamma^{\mathrm{T}}\mathbf{y}),$$

which converges to $\Sigma_{X_\gamma,y}^{\mathrm{T}}\Sigma_{X_\gamma,X_\gamma}^{-1}\Sigma_{X_\gamma,y}$ in probability as $n \to \infty$. This implies that

$$\frac{1}{n}\mathbf{y}^{\mathrm{T}}\mathbf{P}_\gamma^{\perp}\mathbf{y} \to \Sigma_{y,y} - \Sigma_{X_\gamma,y}^{\mathrm{T}}\Sigma_{X_\gamma,X_\gamma}^{-1}\Sigma_{X_\gamma,y}$$

in probability as $n \to \infty$, where $\mathbf{P}_\gamma^{\perp} = \mathbf{I}_n - \mathbf{X}_\gamma(\mathbf{X}_\gamma^{\mathrm{T}}\mathbf{X}_\gamma)^{-1}\mathbf{X}_\gamma^{\mathrm{T}}$. Hence, we can write

$$\begin{aligned}\mathbf{y}^{\mathrm{T}}\mathbf{y} - \frac{n}{1+n}\mathbf{y}^{\mathrm{T}}\mathbf{X}_\gamma(\mathbf{X}_\gamma^{\mathrm{T}}\mathbf{X}_\gamma)^{-1}\mathbf{X}_\gamma^{\mathrm{T}}\mathbf{y} &= \mathbf{y}^{\mathrm{T}}\mathbf{P}_\gamma^{\perp}\mathbf{y} + \frac{1}{1+n}\mathbf{y}^{\mathrm{T}}\mathbf{X}_\gamma(\mathbf{X}_\gamma^{\mathrm{T}}\mathbf{X}_\gamma)^{-1}\mathbf{X}_\gamma^{\mathrm{T}}\mathbf{y} \\ &= \mathbf{y}^{\mathrm{T}}\mathbf{P}_\gamma^{\perp}\mathbf{y}\left(1 + \frac{1}{1+n}\frac{\mathbf{y}^{\mathrm{T}}\mathbf{X}_\gamma(\mathbf{X}_\gamma^{\mathrm{T}}\mathbf{X}_\gamma)^{-1}\mathbf{X}_\gamma^{\mathrm{T}}\mathbf{y}}{\mathbf{y}^{\mathrm{T}}\mathbf{P}_\gamma^{\perp}\mathbf{y}}\right) \\ &= \mathbf{y}^{\mathrm{T}}\mathbf{P}_\gamma^{\perp}\mathbf{y}\{1 + O_p(n^{-1})\}.\end{aligned}$$

It follows that

$$n\log\left\{\mathbf{y}^{\mathrm{T}}\mathbf{y} - \frac{n}{1+n}\mathbf{y}^{\mathrm{T}}\mathbf{X}_\gamma(\mathbf{X}_\gamma^{\mathrm{T}}\mathbf{X}_\gamma)^{-1}\mathbf{X}_\gamma^{\mathrm{T}}\mathbf{y}\right\} = n\log(\mathbf{y}^{\mathrm{T}}\mathbf{P}_\gamma^{\perp}\mathbf{y}) + O_p(1). \tag{A1}$$

Also, note that

$$p_\gamma\log(n+1) = p_\gamma\log n + p_\gamma\log\left(\frac{n+1}{n}\right) = p_\gamma\log n + o(1). \tag{A2}$$

Using (A1) and (A2), we can write $D(\gamma)$ as

$$\begin{aligned}D(\gamma) &= n\log\left(\mathbf{y}^{\mathrm{T}}\mathbf{P}_\gamma^{\perp}\mathbf{y}\right) + p_\gamma\log n + 2\log\binom{p}{p_\gamma} + O_p(1) \\ &= \left\{n\log\left(\mathbf{y}^{\mathrm{T}}\mathbf{P}_\gamma^{\perp}\mathbf{y}\right) + p_\gamma\log n + 2\log\binom{p}{p_\gamma}\right\}\{1 + o_p(1)\} \\ &= \mathrm{EBIC}(\gamma)\{1 + o_p(1)\},\end{aligned} \tag{A3}$$

where EBIC$(\cdot)$ denotes the extended Bayesian information criterion (Chen and Chen 2008). Note that Theorem 1 of Chen and Chen (2008) implies that

$$\mathrm{EBIC}(\gamma_{\mathrm{HPM}}) < \mathrm{EBIC}(\gamma_1) < \mathrm{EBIC}(\gamma_2)$$

in probability as $n \to \infty$ for any γ_1 and γ_2 such that (a) $\gamma_{\mathrm{HPM}} \subsetneq \gamma_1 \subsetneq \gamma_2$ or (b) $\gamma_{\mathrm{HPM}} \subset \gamma_1$ and $\gamma_{\mathrm{HPM}} \not\subset \gamma_2$. By the asymptotic equivalence in A3, we therefore obtain the results of our Theorem 1.

Appendix B: Proof of Theorem 3.2

Suppose that, in Step 1, the HD-DVS algorithm (Algorithm 1) visits $\tilde{\gamma}$ such that $\tilde{\gamma} \supset \gamma_{\mathrm{HPM}}$. Then, by Theorem 1(a), the probability that the HD-DVS algorithm converges to γ_{HPM} goes to one as $n \to \infty$.

Suppose that, in Step 1, the HD-DVS algorithm never visits $\tilde{\gamma}$ such that $\tilde{\gamma} \supset \gamma_{\mathrm{HPM}}$. In this case, by Theorem 2 of Wang (2009) and A3, the probability that Step 2 of the HD-DVS algorithm converges to $\tilde{\gamma}_+ (\supset \gamma_{\mathrm{HPM}})$ goes to one as $n \to \infty$. Then, the algorithm goes back to Step 1 with the initial value $\hat{\gamma} = \tilde{\gamma}_+$. Therefore, this time the algorithm converges to γ_{HPM} in probability. This completes our proof.

References

Albert JH, Chib S (1993) Bayesian analysis of binary and polychotomous response data. J Am Stat Assoc 88(422):669–679

Barbieri MM, Berger JO (2004) Optimal predictive model selection. Ann Stat 32(3):870–897

Bhattacharya A, Chakraborty A, Mallick BK (2016) Fast sampling with Gaussian scale mixture priors in high-dimensional regression. Biometrika 103:985–991

Carvalho CM, Polson NG, Scott JG (2009) Handling sparsity via the horseshoe. In: Artificial intelligence and statistics. PMLR, pp 73–80

Carvalho CM, Polson NG, Scott JG (2010) The horseshoe estimator for sparse signals. Biometrika 97(2):465–480

Casella G, Moreno E (2006) Objective Bayesian variable selection. J Am Stat Assoc 101(473):157–167

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

Cibis H, Biyanee A, Dörner W, Mootz HD, Klempnauer KH (2020) Characterization of the zinc finger proteins ZMYM2 and ZMYM4 as novel B-MYB binding proteins. Sci Rep 10(1):8390

Deng HX, Shi Y, Yang Y, Ahmeti KB, Miller N, Huang C, Cheng L, Zhai H, Deng S, Nuytemans K et al (2016) Identification of TMEM230 mutations in familial Parkinson's disease. Nat Genet 48(7):733–739

Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc 96(456):1348–1360

George EI, McCulloch RE (1993) Variable selection via Gibbs sampling. J Am Stat Assoc 88(423):881–889

Hans C, Dobra A, West M (2007) Shotgun stochastic search for large p regression. J Am Stat Assoc 102(478):507–516

Hindmarch C, Fry M, Yao ST, Smith PM, Murphy D, Ferguson AV (2008) Microarray analysis of the transcriptome of the subfornical organ in the rat: regulation by fluid and food deprivation. Am J Physiol Regul Integr Comp Physiol 295(6):R1914–R1920

Jin S, Goh G (2021) Bayesian selection of best subsets via hybrid search. Comput Stat 36(3):1991–2007

Johndrow J, Orenstein P, Bhattacharya A (2020) Scalable approximate MCMC algorithms for the horseshoe prior. J Mach Learn Res 21(73):1–61

Kass RE, Wasserman L (1995) A reference Bayesian test for nested hypotheses and its relationship to the Shwarz criterion. J Am Stat Assoc 90(431):928–934

Koslovsky M, Swartz MD, Leon-Novelo L, Chan W, Wilkinson AV (2018) Using the EM algorithm for Bayesian variable selection in logistic regression models with related covariates. J Stat Comput Simul 88(3):575–596

Lu TT, Shiou SH (2002) Inverses of 2× 2 block matrices. Comput Math Appl 43(1–2):119–129

Moreno E, Girón J, Casella G (2015) Posterior model consistency in variable selection as the model dimension grows. Stat Sci 30(2):228–241

Narisetty NN, Shen J, He X (2018) Skinny Gibbs: a consistent and scalable Gibbs sampler for model selection. J Am Stat Assoc 114(527):1205–1217

Park T, Casella G (2008) The Bayesian lasso. J Am Stat Assoc 103(482):681–686

Raftery AE, Madigan D, Hoeting JA (1997) Bayesian model averaging for linear regression models. J Am Stat Assoc 92(437):179–191

Ročková V, George EI (2014) EMVS: the EM approach to Bayesian variable selection. J Am Stat Assoc 109(506):828–846

Ročková V, George EI (2018) The spike-and-slab lasso. J Am Stat Assoc 113(521):431–444

Rocková V, Moran G (2021) EMVS Vignette

Scheetz TE, Kim KYA, Swiderski RE, Philp AR, Braun TA, Knudtson KL, Dorrance AM, DiBona GF, Huang J, Casavant TL, Sheffield VC, Stone EM (2006) Regulation of gene expression in the mammalian eye and its relevance to eye disease. Proc Natl Acad Sci 103(39):14429–14434

Scott JG, Berger JO (2010) Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. Ann Stat 38:2587–2619

Tadesse MG, Vannucci M (2021) Handbook of Bayesian variable selection. CRC Press, Boca Raton

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Roy Stat Soc Ser B (Methodol) 58(1):267–288

Wang H (2009) Forward regression for ultra-high dimensional variable screening. J Am Stat Assoc 104(488):1512–1524

Yang Y, Wainwright MJ, Jordan MI (2016) On the computational complexity of high-dimensional Bayesian variable selection. Ann Stat 44(6):2497–2532

Zellner A (1986) On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In: Goel PK, Zellner A (eds) Bayesian inference and decision techniques. Elsevier, New York, pp 233–243

Zhang Z (2014) The matrix ridge approximation: algorithms and applications. Mach Learn 97(3):227–258

Zhao K, Lian H (2016) The expectation–maximization approach for Bayesian quantile regression. Comput Stat Data Anal 96:1–11