Structured Variational Inference for Bayesian Variable Selection

Jinwoo Lee[1], Junghoon Lim[1], Seonghyun Jeong[2] and Taeyoung Park[2]

[1] Department of Statistics and Data Science, Yonsei University, {josh_99, meaningfull9502}@ yonsei.ac.kr
[2] Department of Applied Statistics, Yonsei University, test2@email.com

Abstract

Bayesian Variable Selection offers advantages in addressing sparse variable selection problems by incorporating model selection priors. However, exact computation of Bayesian Variable Selection is infeasible for large covariate spaces. Variational Inference provides computational efficiency but sacrifices accuracy. To overcome these challenges, we propose a novel approach that integrates Markov chain Monte Carlo within Structured Variational Inference for Bayesian Varaible Selection. We demonstrate the effectiveness of our approach in obtaining accurate estimates. Our experiments show improved performance, particularly for strongly correlated variables. Our findings contribute to the advancement of sparse model selection methods in high-dimensional and correlated covariate settings.

Keywords— *Bayesian Variable Selection, Variational Inference*

I. INTRODUCTION

In the era of big data, the analysis of complex datasets with a large number of variables has become a common challenge across diverse fields. Many datasets exhibit a high-dimensional nature, where the number of variables exceeds the number of observations, and collinearity among variables is prevalent. Identifying the subset of meaningful variables that contribute to explaining the data has motivated the development of sparse model selection methods. However, accurate and efficient variable selection remains a significant challenge.

In the frequentist framework, penalized regression techniques, such as the LASSO (least absolute shrinkage and selection operator), have been commonly used for inducing sparsity in models through penalty functions. Nevertheless, these methods have limitations, including the need for careful selection of penalty parameters and potential bias in the estimates.

Bayesian variable selection (BVS) offers several advantages in addressing the sparse variable selection problem. BVS provides a framework for estimating posterior distributions of all sub-models, allowing for comprehensive uncertainty quantification. Moreover, BVS effectively induces sparsity by incorporating model selection priors, such as spike-and-slab priors, which naturally capture the presence or absence of covariates, with the ”spike” component representing zero-effect variables and the ”slab” component representing non-zero effect variables. Moreover, the spike-and-slab prior enables the computation of the posterior probability of including a specific variable in the model.

However, exact posterior inference for all the variables of interest becomes intractable due to the lack of closed-form solutions. To address this challenge, Markov Chain Monte Carlo (MCMC) methods have been widely employed to explore the posterior space and obtain unbiased estimates. Nonetheless, the computational demands of MCMC can be substantial for models with large covariate spaces, requiring a considerable number of iterations to achieve reliable results.

Variational Inference (VI) has emerged as an alternative approach, known for its computational efficiency. VI approximates the posterior distribution by finding a tractable distribution that minimizes the Kullback-Leibler divergence with the true posterior. Mean-field VI (MFVI), a popular variant, assumes independence among latent variables, providing fast but often inaccurate estimates. Structured VI (SVI) relaxes the independence assumption, yielding more accurate estimates. However, SVI can lead to infeasible computations in certain scenarios [4].

While combining MCMC within SVI has proven successful for other models, such as Latent Dirichlet Allocation [8], to the best of our knowledge, no application of this approach to Bayesian variable selection has been reported. In this paper, we propose a novel approach that integrates MCMC within SVI for Bayesian variable selection.

We demonstrate the effectiveness of incorporating MCMC within SVI for BVS, showcasing improved estimation accuracy compared to pure VI schemes with a reduced number of MCMC iterations. Specifically, we

present the proposed SVI model and investigate its performance enhancements compared to traditional VI and deterministic. Our findings highlight that while SVI with MCMC may introduce a slight increase in computational time compared to Mean-Field VI, it offers substantial improvements, particularly for variables exhibiting strong correlation patterns, such as autoregressive (AR1), block-correlated, or pairwise-correlated structures.

II. Bayesian Variable Selection

Suppose that we model the continuous data $\mathbf{y} \in \mathbb{R}^n$ as linear regression with given $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_p]$, $n \times p$ design matrix of p candidate covariates. Under Bayesian framework with latent binary variables $\gamma = (\gamma_1, ..., \gamma_p)^\top$, the model can be written as

$$\mathbf{y} = \alpha \mathbf{1}_n + \mathbf{X}\Gamma\theta + \varepsilon,$$

where $\mathbf{1}_n$ is a $n \times 1$ vector of 1's, $\mathbf{I}_n$ is a $n \times n$ identity matrix, α is an intercept, $\theta = (\theta_1, ..., \theta_p)^\top$ is a vector of coefficients, $\Gamma = \text{diag}(\gamma)$, and $\varepsilon \sim N(0, \sigma^2 \mathbf{I}_n)$. Assuming that the model is sparse, we assign prior distributions to each γ_j and θ_j as follows:

$$\begin{aligned} \gamma_j &\sim \text{Bernoulli}(\rho), \\ \theta_j &\sim N(0, \sigma^2 \tau_j^2), \end{aligned}$$

where ρ is an expected proportion of included covariates and τ_j^2 is a corresponding variance of θ_j. To automatically control sparsity, we use an exponential distribution $\text{Exp}(\lambda^2/2)$ to model τ_j^2 and assign ρ a beta prior distribution $\text{Beta}(a_0, b_0)$. Note that by marginalizing out each τ_j^2 for θ_j, we obtain a spike and slab prior for $\tilde{\theta}_j = \gamma_j \theta_j$ such that the slab part follows a Laplace distribution

$$\tilde{\theta}_j \sim \gamma_j \text{Laplace}(\theta_j | 0, \lambda/\sigma) + (1 - \gamma_j)\delta_0,$$

where δ_0 is a point mass at 0. This Laplace slab is often advantageous over a Gaussian slab as it assign more probability mass to extreme values due to its exponential tail. We initially set $\lambda = 1$, and take $a_0 = 1$ and $b_0 = p$ since this is known to yield optimal performance in sparse settings [3, 10]. For σ^2, we assign the improper prior distribution $\pi(\sigma^2) \propto 1/\sigma^2$ as it is common to all possible models. To account for an intercept α, we center $\mathbf{y}$ and the columns of $\mathbf{X}$ hence they all have mean zero.

While the spike and slab prior is considered gold standard in sparse bayesian variable selection, exact posterior inference for the latent variables $\mathbf{z} = (\theta, \rho, \sigma^2, \gamma, \tau^2)$ is intractable due to the absence of closed-form solutions. MCMC methods can provide unbiased estimates, however, as the number of covariates increases, it becomes more difficult for a Markov chain to explore the model space. To address the challenges posed by high dimensional models, we turn to VI as an efficient alternative approach.

III. Variational Inference

VI is an optimization method that aims to find the best approximation of the posterior distribution within a class of tractable distributions. Given observable variables $\mathbf{o}$ and latent variables $\mathbf{z}$, optimal solution can be obtained by minimizing the Kullback-Leibler (KL) divergence between posterior distribution $\pi(\mathbf{z}|\mathbf{o})$ and variational distribution $q(\mathbf{z})$, which is equivalent to maximizing the lower bound of marginal likelihood (ELBO) :

$$\log \pi(\mathbf{o}) \geq L(q) = \mathbb{E}_q \left[\log \pi(\mathbf{z}, \mathbf{o}) - \log q(\mathbf{z})\right].$$

If we assume the variational distribution to be factorized as $q(\mathbf{z}) = \prod_k q(\mathbf{z}_k)$, then the optimal solution follows the form of

$$q^\star(\mathbf{z}_k) \propto \exp\left\{\mathbb{E}_{q(\mathbf{z}_{-k})}[\log p(\mathbf{z}_k | \mathbf{z}_{-k}, \mathbf{o})]\right\}, \qquad (1)$$

where $\mathbf{z}_{-k}$ denotes latent variables $\mathbf{z}$ without the kth partition of variables.

The accuracy of the VI approximation and its computational feasibility depends on the choice of variational family, i.e., choosing the way of factorizing the latent variables. Below, we describe two approaches, the one which factorizes the variational distributions which are all analytically tractable, and the other one which retrieve dependencies across some variables but become computationally infeasible.

A. Meam Field Variational Inference

Inspired by [9], we first present the factorization of the variables $\mathbf{z}$ as

$$\begin{aligned} q(\mathbf{z}) = q(\theta|\mu, \Sigma) q(\rho|a_\rho, b_\rho) q(\sigma^2|a_{\sigma^2}/2, b_{\sigma^2}/2) \\ \times \prod_{j=1}^p q(\tau_j^2|1/2, \lambda^2, \nu_j) \prod_{j=1}^p q(\gamma_j|w_j), \end{aligned}$$

where μ and Σ are Gaussian parameters denoting mean and covariance of coefficients; a_ρ and b_ρ are Beta parameters controlling overall inclusion probabilities; a_{σ^2} and b_{σ^2} are Inverse Gamma parameters; each ν_j is one of Generalized Inverse Gaussian parameter controlling the variance of each θ_j; each w_j is Bernoulli parameter denoting the probability of including jth covariate. Updating procedure for these distributions are summarized in Algorithm 1, and we denote this as mean field variational inference (MFVI).

Despite the careful selection of initial settings, Algorithm 1 often converges to suboptimal solutions when applied to large models. To address this issue, we propose the use of a deterministic annealing variant of the Variational algorithm, as proposed by [6]. Specifically, we introduce an annealing process for the entropy of the variational dis-

Algorithm 1: MFVI Algorithm

Input : y, X
repeat :
  Update $q(\theta|\mu,\Sigma)$ with parameters

$$\Sigma = \left[\mathbb{E}_q[1/\sigma^2]\left(\mathbf{X}^\top\mathbf{X}\odot\mathbb{E}_q[\gamma\gamma^\top]+\mathbf{V}\right)\right]^{-1}$$
$$\mu = \mathbb{E}_q[1/\sigma^2]\Sigma\,\mathrm{diag}(\mathbf{w})\mathbf{X}^\top\mathbf{y}$$

  where $\mathbf{V} = \mathrm{diag}\left(\mathbb{E}_q[\tau_1^{-2}],...,\mathbb{E}_q[\tau_p^{-2}]\right)$
  Update each $q(\tau_j^2|1/2,\lambda^2,\nu_j)$ related parameters

$$\nu_j = \mathbb{E}_q[\theta_j^2/\sigma^2]$$

  Update $q(\sigma^2|a_{\sigma^2}/2,b_{\sigma^2}/2)$ related parameters

$$a_{\sigma^2} = n+p$$
$$b_{\sigma^2} = \mathbf{y}^\top\mathbf{y} - 2\mathbf{y}^\top\mathbf{X}\,\mathrm{diag}(\mathbf{w})\mu + \mathrm{tr}\left\{\left(\mathbf{X}^\top\mathbf{X}\odot\mathbb{E}_q[\gamma\gamma^\top]+\mathbf{V}\right)\mathbb{E}_q[\theta\theta^\top]\right\}$$

  Update $q(\rho|a_\rho,b_\rho)$ related parameters

$$a_\rho = 1+\mathbf{1}^\top\mathbf{w}$$
$$b_\rho = 2p-\mathbf{1}^\top\mathbf{w}$$

  Update $q(\gamma_j|\mathbf{w}_j)$ related parameters

$$\xi_j = \eta + \mathbb{E}_q[1/\sigma^2]\Big(\mathbf{y}^\top\mathbf{X}_j\mu_j - \mathbf{X}_j^\top\mathbf{X}_{-j}\mathbf{W}_{-j}\mathbb{E}_q[\theta_j\theta_{-j}] - \frac{1}{2}\mathbb{E}_q[\theta_j^2]\Big)$$
$$w_j = \mathrm{expit}(\xi_j), \quad \mathbf{w} = (w_1,...,w_p)^\top$$

  where $\eta = \mathbb{E}_q[\mathrm{logit}(\rho)]$
until convergence

tribution of γ, resulting in a modified ELBO :

$$L(q;\xi) = \mathbb{E}_q[\log\pi(\mathbf{z},\emptyset)] + \mathbb{E}_q[-\log q(\mathbf{z}_{-\gamma})] + \frac{1}{\beta}\mathbb{E}_q[-\log q(\gamma)],$$

where $\frac{1}{\beta}$ serves as an analog to temperature. The parameter of each $q(\gamma_j|w_j(\beta))$ is then defined as:

$$w_j(\beta) = \mathrm{expit}(\beta\xi_j), \tag{2}$$

where ξ is defined in Algorithm 1. It is worth noting that at high temperatures, i.e., when β is close to 0, the probabilities $w_j(\beta)$ and $1-w_j(\beta)$ are nearly equal, irrespective of the covariate inclusion probability magnitude. In practice, we initiate the annealing process at a high temperature, such as $\beta = 1000$, and gradually decrease it towards one.

B. *Structured Variational Inference*

While the VI with mean field assumption allows tractable solutions, it can lead to a biased approximation of the actual posterior distribution, which may be a result of bad local optimum due to its restrictive assumption of complete independence. SVI, on the other hand, offers a more accurate approximation by restoring some dependencies and is less susceptible to the problem of multiple local optima. However its applicability is limited because the expectation contained in the ELBO must be computationally feasible with respect to the structured distribution.

Previous approaches to Bayesian Variable Selection in variational inference have typically assumed full independence among variables, including γ, σ^2, and θ [2, 5, 9, 10]. In order to improve the approximation, we propose restoring dependencies by factorizing the joint distribution as follows:

$$q(\theta,\sigma^2,\gamma) = q(\theta|\sigma^2,\gamma)q(\sigma^2|\gamma)q(\gamma).$$

We now provide the derivation of each variational distribution. Using Equation (1), the joint distribution $q(\theta,\sigma^2,\gamma)$ can be derived as:

$$q(\theta,\sigma^2,\gamma) \propto (\sigma^2)^{-\frac{n+p+2}{2}}\exp\Big\{\eta\mathbf{1}^\top\gamma - \frac{1}{2\sigma^2}\big(\mathbf{y}^\top\mathbf{y} - 2\mathbf{y}^\top\mathbf{X}\Gamma\theta + \theta^\top(\Gamma\mathbf{X}^\top\mathbf{X}\Gamma+\mathbf{V})\theta\big)\Big\}. \tag{3}$$

Then the distribution $q(\theta|\sigma^2,\gamma)$ follows normal distribution $N(\mu_\gamma,\Sigma_{\sigma^2,\gamma})$ with parameters

$$\Sigma_{\sigma^2,\gamma} = \sigma^2\mathbf{B}_\gamma^{-1}, \tag{4}$$
$$\mu_\gamma = \mathbf{B}_\gamma^{-1}\mathbf{b}_\gamma, \tag{5}$$

where $\mathbf{B}_\gamma = \Gamma\mathbf{X}^\top\mathbf{X}\Gamma+\mathbf{V}$ and $\mathbf{b}_\gamma = \Gamma\mathbf{X}^\top\mathbf{y}$.

To derive $q(\sigma^2|\gamma)$, we need to marginalize out the variable θ from the equation (3):

$$\begin{aligned} q(\sigma^2,\gamma) &\propto (\sigma^2)^{-\frac{n+p+2}{2}}\exp\Big(\eta\mathbf{1}^\top\gamma - \frac{1}{2\sigma^2}\mathbf{y}^\top\mathbf{y}\Big) \\ &\quad\times\int_\theta\exp\Big\{\frac{1}{2\sigma^2}\big(2\mathbf{b}_\gamma^\top - \theta^\top\mathbf{B}_\gamma\theta\big)\Big\}d\theta \\ &\propto |\mathbf{B}_\gamma|^{-\frac{1}{2}}(\sigma^2)^{-\frac{n+2}{2}}\exp\Big\{\eta\mathbf{1}^\top\gamma - \frac{1}{2\sigma^2}\big(\mathbf{y}^\top\mathbf{y} - \mathbf{b}_\gamma^\top\mathbf{B}_\gamma^{-1}\mathbf{b}_\gamma\big)\Big\}. \end{aligned} \tag{6}$$

Then $q(\sigma^2|\gamma)$ follows an Inverse Gamma distribution $IG\left(n/2, s_\gamma/2\right)$, where

$$s_\gamma = \mathbf{y}^\top\mathbf{y} - \mathbf{b}_\gamma^\top\mathbf{B}_\gamma^{-1}\mathbf{b}_\gamma \tag{7}$$

To derive $q(\gamma)$, we need to marginalize out the variable σ^2 from the equation (6) as follows:

$$\begin{aligned} q(\gamma) &\propto |\mathbf{B}_\gamma|^{-\frac{1}{2}}e^{\eta\mathbf{1}^\top\gamma}\int_{\sigma^2}(\sigma^2)^{-\frac{n+2}{2}}\exp\Big\{-\frac{s_\gamma}{2\sigma^2}\Big\}d\sigma^2 \\ &\propto |\mathbf{B}_\gamma|^{-\frac{1}{2}}e^{\eta\mathbf{1}^\top\gamma}\left(s_\gamma/2\right)^{-n/2} \overset{let}{=} Q(\gamma) \end{aligned}$$

Then $q(\gamma)$ has the distributional form of $Q(\gamma)/Z$ where $Z = \sum_\gamma Q(\gamma)$. Note that other latent variables τ^2 and ρ are factorized in the same manner as MFVI.

In order to update variational distributions $q(\tau^2)$ and $q(\rho)$, computation of expectations with respect to $q(\gamma)$ is required, which is challenging even for a moderate size of p. We utilize MCMC to approximate expectations, which allows us to sample from $q(\gamma)$ using $Q(\gamma)$. Specifically, once we have sampled $\{\gamma^s\}_{s=1}^S \sim q(\gamma)$, we approximate each r_j and ω_j as follows:

$$\nu_j \approx \frac{1}{S}\sum_{s=1}^{S}\left\{\mathbf{B}^{-1}_{\gamma^s,j,j} + \phi_{\gamma^s}\mu^2_{\gamma^s,j}\right\}$$
$$w_j \approx \frac{1}{S}\sum_{s=1}^{S}\mathbb{1}(\gamma^s_j = 1),$$

where $\phi_{\gamma^s} = \mathbb{E}_q[1/\sigma^2|\gamma^s]$.

In our simulation, we employed the Gibbs Sampler (GS) to obtain $\{\gamma^s\}_{s=1}^S$ for each VI iteration. It is important to note that other MCMC methods, such as Metropolis-Hastings within Gibbs Sampler, can be applied for more efficient sampling. However, in our simulation setting, the Gibbs Sampler was sufficient to demonstrate good performance. Algorithm 2 summarizes the updating procedure for PSVI.

Algorithm 2: PSVI Algorithm

Input : y, X
repeat :
  Update $q(\theta, \sigma^2, \gamma)$ related parameters $\mu_\gamma, \Sigma_{\sigma^2,\gamma}, Q(\gamma)$ according to (4),(5),(7), and (8)
  Sample $\{\gamma^s\}_{s=1}^S \sim q(\gamma)$
  Update $q(\rho) = Beta(a_\rho, b_\rho)$ by

$$a_\rho = 1 + \mathbf{1}^\top \mathbf{w}$$
$$b_\rho = 2p - \mathbf{1}^\top \mathbf{w}$$

  Update each $q(\tau_j^2|1/2, \lambda^2, \nu_j)$ related parameters

$$\nu_j = \mathbb{E}_q[\theta_j^2/\sigma^2]$$

until convergence

C. Convergence criterion

To ensure the convergence of the VI, ELBO is commonly utilized as a stopping rule. For the MFVI, the explicit derivation of the ELBO follows a similar approach as described by [9]. For the SVI, the ELBO is given by

$$\begin{aligned} L_{\text{PSVI}}(q) &= \mathbb{E}_q[\log \pi(\mathbf{z},\mathbf{y}) - \log q(\mathbf{z})] \\ &\propto -\frac{\mathbb{E}_q[1/\sigma^2]}{2}\mathbf{y}^\top\mathbf{y} + \mathbf{y}^\top\mathbf{X}\mathbb{E}_q[\tilde{\theta}/\sigma^2] \\ &\quad - \frac{1}{2}\text{tr}\left\{(\mathbf{X}^\top\mathbf{X} + \mathbf{V}) \odot \mathbb{E}_q[\tilde{\theta}\tilde{\theta}^\top/\sigma^2]\right\} \\ &\quad + p\mathbb{E}_q[\log(1-\rho)] + \log Z \\ &\quad + \sum_j^p \frac{\log 2\lambda^2}{2} r_j + \log K_{r_j}\left(\sqrt{\frac{\lambda^2}{2}}\right) \\ &\quad + \left(\frac{1}{2} - \nu_j\right)\mathbb{E}_q[\log \tau_j^2] + \left(\frac{1}{4} - \frac{\lambda^2}{2}\right)\mathbb{E}_q[\tau_j^2] \\ &\quad + \frac{\lambda^2}{2}\mathbf{V}_{j,j}. \end{aligned}$$

Due to the presence of expectation terms with respect to $q(\gamma)$ and the normalizing constant of $q(\gamma)$, the ELBO should be approximated with MCMC. We denote this approximation as $\hat{L}_S(q)$. To account for the stochastic fluctuations arising from the MCMC sampling, we adopt a long-term convergence criterion, given by

$$\left|\frac{\{\hat{L}_S(q_i) + \hat{L}_S(q_{i-1})\} - \{\hat{L}_S(q_{i-2}) + \hat{L}_S(q_{i-3})\}}{\hat{L}_S(q_{i-2}) + \hat{L}_S(q_{i-3})}\right| < \varepsilon$$

IV. Simulation studies

We conduct a simulation study to evaluate the effect of structured variational distribution with three different experiments. To begin, we compare MFVI and SVI using synthetic datasets with small number of covariates, specifically 12. For SVI, we can update the parameters exactly by computing the expectation with all combinatorial summations. The results can be used to evaluate the accuracy of the MCMC approximation. Next, we investigate the performance of these methods when dealing with larger number of covariates. In this case, we can only rely on the GS approximation for SVI (SVI+G). However, we demonstrate that these methods provide superior results in a reasonably fast time when compared to MFVI and other deterministic methods. Finally, we apply SVI to real data.

In order to evaluate the performance of the algorithms, two metrics are employed. We first use the F_1 score for assessing the accuracy of the model selection, defined as below,

$$F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

with

$$\text{precision} = \frac{TP}{TP+FP}, \quad \text{Recall} = \frac{TP}{TP+FN},$$

where TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives.

Structure	corr	MFVI	SVI	SVI+G
Non	-	$0.87_{\pm 0.14}$	$0.91_{\pm 0.12}$	$0.91_{\pm 0.12}$
AR(1)	0.4	$0.86_{\pm 0.15}$	$0.90_{\pm 0.14}$	$0.90_{\pm 0.14}$
	0.8	$0.71_{\pm 0.21}$	$0.82_{\pm 0.18}$	$0.79_{\pm 0.19}$
Block	0.4	$0.86_{\pm 0.14}$	$0.91_{\pm 0.13}$	$0.90_{\pm 0.14}$
	0.8	$0.72_{\pm 0.20}$	$0.81_{\pm 0.17}$	$0.81_{\pm 0.17}$

Table 1. F1-score for low dimensional case (mean $\pm$ standard deviation).

To measure the prediction quality, rooted mean square error (RMSE) is adopted, that is,

$$\text{RMSE} = \sqrt{\frac{1}{n}\|\mathbf{X}\theta_0 - \mathbf{X}\hat{\theta}\|_2^2},$$

where θ_0 is the underlying true coefficients and $\hat{\theta}$ is the estimate of $\tilde{\theta}$ of each algorithm.

A. *Simulation in low dimensional case*

In this simulation study, we simulate synthetic data consisting of $n = 20$ observations and $p = 12$ covariates, where only 3 covariates have non-zero effects and the rest are considered unimportant. The true coefficients vector is set to be $\theta_0 = (1,0,0,0,1.5,0,0,0,2,0,0,0)$. To construct the design matrix, we assume that each $\mathbf{x}_i$ follows $N(0,\Sigma)$. We consider three correlation structures: the first assumes no correlation, the second is an AR(1) structure denoted by Σ_ρ^{ar} with $\Sigma_{\rho,jk}^{\text{ar}} = \rho^{|j-k|}$, and the third is a block diagonal correlation structure $\Sigma_\rho^{\text{blk}} = \text{bdiag}(\Sigma_\rho, \Sigma_\rho, \Sigma_\rho)$ where $\Sigma_\rho = (\sigma_{jk})_{j,k=1}^{4,4}$ such that $\sigma_{jk} = \rho$ for $j \neq k$ and $\sigma_{jj} = 1$ otherwise. We considered two correlation coefficient values, $\rho = 0.4$ and $\rho = 0.8$, resulting in a total of 5 simulation settings. The response variable $\mathbf{y}$ was generated according to the linear model $\mathbf{y} = \mathbf{X}\theta_0 + \varepsilon$, where $\varepsilon \sim N(0, \mathbf{I}_n)$.

The algorithms compared in this study are MFVI, SVI, and SVI+G. For each iteration of the SVI+G, twenty simulated samples were generated after discarding the first ten percent of initial samples. No thinning was applied.

The simulation results, presented in Table 1, display the average $F1$-scores across 100 applications. Notably, we observed that SVI consistently outperformed MFVI, exhibiting higher $F1$-scores. Moreover, SVI demonstrated greater robustness in highly correlated structures, as indicated by the smaller decrease in $F1$-scores compared to MFVI. While the results from SVI+G displayed some stochastic variability relative to SVI, they still outperformed MFVI with a comparable standard deviation.

B. *Simulation in high dimensional case*

We now present the details of high dimensional simulation study, which encompasses two distinct cases characterized by varying sample sizes, numbers of covariates, and coefficients :

- Case 1 : In this case, we set the sample size to $n = 100$ and the number of covariates to $p = 200$. We consider a total of 20 coefficients, evenly spaced between 0.1 and 3 in terms of their absolute values. These coefficients exhibit alternating signs.
- Case 2 : In this case, we set the sample size and number of covariates as $(n, p) = (200, 1000)$. Similar to Case 1, we consider 40 coefficients with the same properties.

For each case, we explore three different correlation structures. Two of these structures have been previously mentioned and include an AR(1) structure and a Block structure. Additionally, we introduce a new correlation structure called Pairwise correlated, where every pair of covariates is correlated. For each correlation structure, we consider two correlation coefficients: 0.4 and 0.8. This results in a total of 12 different experiment sets. Note that the variance of the noise term ε is fixed at 1 across all experiment sets.

In addition to comparing the SVI+G and MFVI methods, we include four other variable selection methods: EMVS [11], VARBVS [2], SSLASSO [12], and SPVB [10]. Each of these methods has its own dedicated R-package, allowing for a comprehensive assessment of their performance. For EMVS, we set $v_0 \in \{0.01, 0.02, ..., 0.1\}$, $v_1 = 1$, and $a = 1, b = p$, with other settings set as default. Among ten different choices of v_0, we select the model with the smallest RMSE. Regarding the VARBVS method, we utilize the default settings and computed the RMSE using the predict function provided by the package. For the SSLASSO method, we set $\lambda_1 = 0.1$, λ_0 an arithmetic series between λ_1 and p with p elements, variance option as 'unknown', penalty option as 'adaptive', $a = 1$, $b = p$, and theta=$1/p$. We choose the model based on the stabilization of the regularization path. The SPVB method was configured with the "laplace" option for the slab, setting prior scale to 1. The remaining options were left at their default values. Finally, for the pure MCMC approach, we employed the GS scheme as the following to perform posterior inference :

$$\begin{aligned}
\gamma^{s+1} &\sim \pi(\gamma \mid \mathbf{T}^s, \rho^s) \\
\rho^{s+1} &\sim \pi(\rho \mid \gamma^{s+1}) \\
(\sigma^2)^{s+1} &\sim \pi(\sigma^2 \mid \gamma^{s+1}) \\
\theta^{s+1} &\sim \pi(\theta \mid (\sigma^2)^{s+1}, \gamma^{s+1}) \\
(\tau^2)^{s+1} &\sim \pi(\tau^2 \mid \theta^{s+1}, (\sigma^2)^{s+1})
\end{aligned}$$

Geweke's diagnostic is employed to determine the convergence of the MCMC : we terminate the MCMC when all the Geweke's statistics of coefficients are within the predetermined interval. We used the geweke.diag function from the coda package. The default settings for frac1 and frac2 were used, where frac1 was set to 0.1 and frac2 to 0.5. A two-sided test with a significance level of 0.01 was employed to reject the null hypothesis.

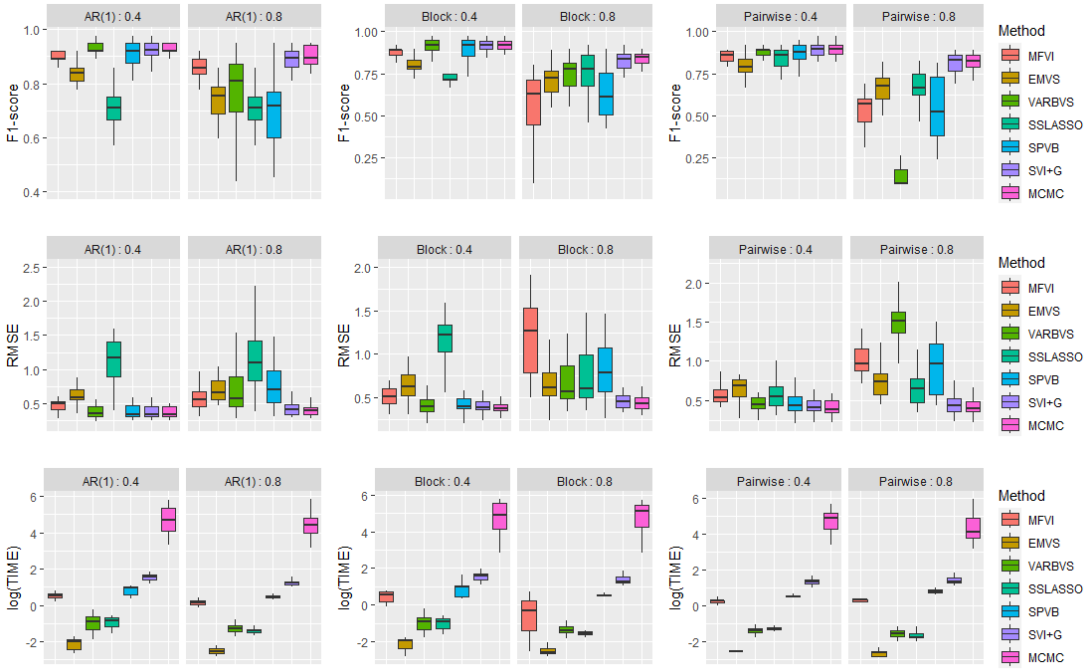

Fig. 1. Summary of F1-score, RMSE, and log of time for Case 1.

Summary of the results of the performance are plotted in Figure 1 and Figure 2. In both two cases, we can see that SVIG consistently gives good results on both F1-score and RMSE among other methods regardless of both the correlation structures and the strength of correlation. Both MFVI and SPVB performances gets worse for larger correlation for all correlation structures. While VARBVS gives good results on small correlation situations, its performance deteriorates on Pairwise structure with high correlation. Although the time records for SVI+G are higher than other methods, they are considerably lower than MCMC, while showing similar performance to MCMC and high performance to other methods. This shows high efficiency on the performance with respect to time.

C. *Application to Real data set*

We investigate the performance of SVIG on two real-world datasets: the McDonald dataset (MD) by [7], which includes mortality and air pollution data, and the TopGear dataset (TG) from [1], derived from the popular British TV show Top Gear. In our analysis of both MD and TG datasets, we account for potential non-linear relationships by incorporating interaction terms among numerical variables and dummy variables representing categorical variables. Additionally, we include squared terms for the numerical variables. Specifically, for the TG dataset, we apply a log transformation to the response variable to address skewness, and we consider 10 numerical variables and 15 categorical variables. For the MD dataset, we utilize all available covariates. Consequently, the MD dataset comprises $n = 60$ observations and $p = 135$ covariates, while the TG dataset consists of $n = 242$ observations and $p = 751$ covariates. Missing values in the data are excluded from the analysis.

To assess the performance of SVI+G, we compare it with the methods employed in the high-dimensional simulation settings. We randomly split the data into training and test sets, utilizing 80% of the observations for training and reserving the remaining 20% for testing. We repeated this procedure 100 times and computed the mean and standard deviation of the test error and the number of selected variables.

Table 2 presents the results, indicating that SVI+G achieves competitive test error rates. For the MD dataset, both SVI+G and SSLASSO demonstrate parsimonious variable selection compared to other methods, with SVI+G showcasing superior predictive performance. On the other hand, for the TG dataset, SVI+G yields less conservative variable selection while maintaining good predictive accuracy.

V. Conclusion

In this paper, we proposed the SVI approach for BVS, which enhances the accuracy of model selection and prediction by incorporating dependencies across variables. Our method employs MCMC to approximate intractable expectations. In the future, we plan to explore the theoretical properties of SVI and investigate alternative faster approximations that can replace the reliance on MCMC.

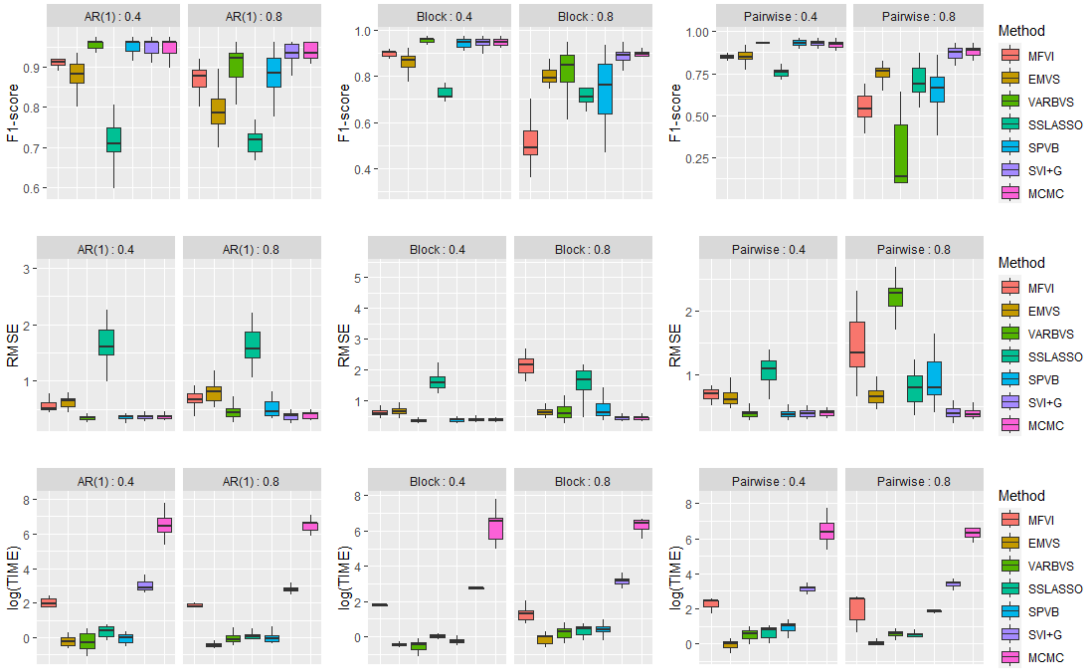

Fig. 2. Summary of F1-score, RMSE, and log of time for Case 2.

	MD ($n = 48, p = 135$)			TG ($n = 194, p = 751$)		
Model	RMSE	Selected variables	time	RMSE$\times$100	Selected variables	time
MFVI	$51.0_{\pm 10.4}$	$3.14_{\pm 1.25}$	$0.18_{\pm 0.11}$	$17.7_{\pm 3.79}$	$10.6_{\pm 1.44}$	$7.94_{\pm 2.08}$
EMVS	$111_{\pm 197}$	$39.0_{\pm 3.00}$	$0.03_{\pm 0.01}$	$21.8_{\pm 5.37}$	$4.98_{\pm 0.94}$	$0.42_{\pm 0.09}$
VARBVS	$51.5_{\pm 19.40}$	$2.27_{\pm 0.80}$	$0.13_{\pm 0.07}$	$19.2_{\pm 4.40}$	$8.04_{\pm 0.93}$	$0.59_{\pm 0.16}$
SSLASSO	$50.9_{\pm 9.04}$	$1.06_{\pm 1.16}$	$0.00_{\pm 0.01}$	$19.3_{\pm 4.64}$	$8.78_{\pm 1.62}$	$8.79_{\pm 2.14}$
SPVB	$57848_{\pm 216307}$	$6.00_{\pm 16.7}$	$0.61_{\pm 0.19}$	$21.2_{\pm 5.15}$	$10.7_{\pm 3.05}$	$0.86_{\pm 0.19}$
SVI+G	$47.3_{\pm 8.77}$	$1.77_{\pm 0.79}$	$0.51_{\pm 0.22}$	$17.1_{\pm 4.20}$	$10.7_{\pm 1.80}$	$9.21_{\pm 1.30}$
MCMC	$47.1_{\pm 9.10}$	$3.42_{\pm 1.63}$	$3.54_{\pm 3.04}$	$21.3_{\pm 7.72}$	$14.2_{\pm 3.25}$	$576_{\pm 121}$

Table 2. Results on Real data (mean $\pm$ standard deviation).

REFERENCES

[1] Andreas Alfons. robusthd: Robust methods for high-dimensional data. URL: https://CRAN.R-project.org/package=robustHD.

[2] Peter Carbonetto and Matthew Stephens. Scalable variational inference for bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis*, 7(1):73–108, 2012.

[3] I Castillo, J Schmidt-Hieber, and A van Der Vaart. Bayesian linear regression with sparse priors. *Annals of Statistics*, 43(5):1986–2018, 2015.

[4] Matthew D Hoffman and David M Blei. Structured stochastic variational inference. In *Artificial Intelligence and Statistics*, volume 38, pages 361–369, San Diego, CA, USA, 2015.

[5] Xichen Huang, Jin Wang, and Feng Liang. A variational algorithm for bayesian variable selection. *arXiv preprint arXiv:1602.07640*, 2016.

[6] Kentaro Katahira, Kazuho Watanabe, and Masato Okada. Deterministic annealing variant of variational bayes method. In *Journal of Physics: Conference Series*, volume 95, page 012015. IOP Publishing, 2008.

[7] Gary C McDonald and Richard C Schwing. Instabilities of regression estimates relating air pollution to mortality. *Technometrics*, 15(3):463–481, 1973.

[8] David Mimno, Matthew D Hoffman, and David M Blei. Sparse stochastic inference for latent dirichlet allocation. In *ICML'12: Proceedings of the 29th International Conference on Machine Learning*, 2012.

[9] John T Ormerod, Chong You, and Samuel Müller. A variational bayes approach to variable selection. *Electronic Journal of Statistics*, 11:3549–3594, 2017.

[10] Kolyan Ray and Botond Szabó. Variational bayes for high-dimensional linear regression with sparse priors. *Journal of the American Statistical Association*, 117(539):1270–1281, 2022.

[11] Veronika Ročková and Edward I George. Emvs: The em approach to bayesian variable selection. *Journal of the American Statistical Association*, 109(506):828–846, 2014.

[12] Veronika Ročková and Edward I George. The spike-and-slab lasso. *Journal of the American Statistical Association*, 113(521):431–444, 2018.

SUMMARY OF THIS PAPER

A. Problem Setup

The analysis of complex datasets with a large number of variables has become a common challenge in various fields due to the era of big data. Sparse model selection methods have been developed to identify meaningful variables, but accurate and efficient variable selection remains challenging. Bayesian variable selection (BVS) offers advantages by estimating posterior distributions and incorporating spike-and-slab priors for sparsity. However, exact posterior inference is intractable, leading to the use of computationally demanding Markov Chain Monte Carlo (MCMC) methods.

B. Novelty

C. Algorithms

D. Experiments