

## BIVAS: A Scalable Bayesian Method for Bi-Level Variable Selection With Applications

Mingxuan Cai, Mingwei Dai, Jingsi Ming, Heng Peng, Jin Liu & Can Yang

To cite this article: Mingxuan Cai, Mingwei Dai, Jingsi Ming, Heng Peng, Jin Liu & Can Yang (2020) BIVAS: A Scalable Bayesian Method for Bi-Level Variable Selection With Applications, Journal of Computational and Graphical Statistics, 29:1, 40-52, DOI: [10.1080/10618600.2019.1624365](https://doi.org/10.1080/10618600.2019.1624365)

To link to this article: <https://doi.org/10.1080/10618600.2019.1624365>



View supplementary material [↗](#)



Published online: 19 Jul 2019.



Submit your article to this journal [↗](#)



Article views: 474



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 4 View citing articles [↗](#)



# BIVAS: A Scalable Bayesian Method for Bi-Level Variable Selection With Applications

Mingxuan Cai<sup>a</sup>, Mingwei Dai<sup>b</sup>, Jingsi Ming<sup>a</sup>, Heng Peng<sup>c</sup>, Jin Liu<sup>d</sup>, and Can Yang<sup>a</sup>

<sup>a</sup>Department of Mathematics, The Hong Kong University of Science and Technology, Kowloon, Hong Kong; <sup>b</sup>Center of Statistical Research and School of Statistics, Southwestern University of Finance and Economics, Chengdu, China; <sup>c</sup>Department of Mathematics, Hong Kong Baptist University, Kowloon, Hong Kong; <sup>d</sup>Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore

## ABSTRACT

In this article, we consider a Bayesian bi-level variable selection problem in high-dimensional regressions. In many practical situations, it is natural to assign group membership to each predictor. Examples include that genetic variants can be grouped at the gene level and a covariate from different tasks naturally forms a group. Thus, it is of interest to select important groups as well as important members from those groups. The existing Markov chain Monte Carlo methods are often computationally intensive and not scalable to large datasets. To address this problem, we consider variational inference for bi-level variable selection. In contrast to the commonly used mean-field approximation, we propose a hierarchical factorization to approximate the posterior distribution, by using the structure of bi-level variable selection. Moreover, we develop a computationally efficient and fully parallelizable algorithm based on this variational approximation. We further extend the developed method to model datasets from multitask learning. The comprehensive numerical results from both simulation studies and real data analysis demonstrate the advantages of BIVAS for variable selection, parameter estimation, and computational efficiency over existing methods. The method is implemented in R package “bivas” available at <https://github.com/mxcai/bivas>. Supplementary materials for this article are available online.

## ARTICLE HISTORY

Received March 2018  
Revised May 2019

## KEYWORDS

Bayesian variable selection;  
Group sparsity; Parallel  
computing; Variational  
inference



## 1. Introduction

Variable selection plays an important role in modern data analysis with the ever-increasing number of variables, where it is often assumed that only a small proportion of variables are relevant to the response variable (Hastie, Tibshirani, and Wainwright 2015). In many real applications, this sparse pattern could be more complicated. In this article, we consider a class of regression problems in which the grouping structure of the variables naturally exists. Examples include, but not limited to, the categorical predictors that are often represented by a group of indicators and continuous predictors that can be expressed by a group of basis functions. We assume that only a proportion of groups are relevant to the response variable and within each relevant group, only a subset of variables is relevant. Hence, we consider a bi-level variable selection problem, that is, variable selection at both the individual and group levels (Breheny and Huang 2009).


There have been rich literatures on variable selection (Tibshirani 1996; Fan and Li 2001; Zhang 2010; Yuan and Lin 2006), but majority of them focus on variable selection at the individual level, including penalized methods, such as Lasso (Tibshirani 1996), SCAD (Fan and Li 2001), and MCP (Zhang 2010), and Bayesian variable selection methods based on sparsity-promoting priors, such as Laplace-like priors (Figueiredo 2003; Mallick and Bae 2004; Yuan and Lin 2005; Park and Casella 2008) and spike-slab priors (Mitchell and Beauchamp 1988; George and McCulloch 1993; Madigan and Raftery 1994;

George and McCulloch 1997; Ročková and George 2014; Ormerod, You, and Müller 2017; Zhang, Xu, and Zhang 2019; Ročková 2018). To perform variable selection at the group level, the group Lasso (Yuan and Lin 2006) introduced the  $L_1$ – $L_2$  norm penalty to group variables and perform group selection using the  $L_1$  norm. CAP (Zhao, Rocha, and Yu 2009) generalized this idea to be the  $L_1$ – $L_\gamma$  norm, where  $\gamma \in [1, +\infty)$ . Under the Bayesian framework, this is achieved by specifying the prior over a whole group of variables (Kyung et al. 2010; Raman et al. 2009; Xu and Ghosh 2015).

The group variable selection methods usually act like Lasso at the group level and variables are selected in the “all-in or all-out” manner. However, these methods does not yield sparsity within a group, that is, if a group is selected, all variables within that group will be nonzero. To conduct variable selection at both the individual and group levels, various methods have been proposed for bi-level selection from different perspectives including both penalized and Bayesian methods. Penalized methods often consider a composition of two penalties. The group bridge (Huang et al. 2009) adopts a bridge penalty on the group level and the  $L_1$  penalty on the variable level. Hierarchical Lasso (Zhou and Zhu 2010) can be viewed as a special case of group bridge with bridge index fixed at 0.5. Under certain regularity conditions, the global group bridge solution is proved to be group selection consistent. However, the singularity nature of these penalties at 0 potentially complicates the optimization in practice. The composite MCP (cMCP) (Breheny and Huang

**CONTACT** Can Yang  [macyang@ust.hk](mailto:macyang@ust.hk)  Department of Mathematics, The Hong Kong University of Science and Technology, Room 3432, Clearwater Bay, Kowloon, Hong Kong.

Color versions of one or more of the figures in the article can be found online at [www.tandfonline.com/r/JCGS](http://www.tandfonline.com/r/JCGS).

 Supplementary materials for this article are available online. Please go to [www.tandfonline.com/r/JCGS](http://www.tandfonline.com/r/JCGS).

© 2019 American Statistical Association, Institute of Mathematical Statistics, and Interface Foundation of North America

2009) and group exponential Lasso (GEL) (Breheny 2015) proposed to apply their penalty at both levels in a manner that puts less penalization as the absolute value of a coefficient becomes larger. On the other hand, Bayesian methods usually assume a spike-slab prior on variables at both the individual and group levels to promote bi-level sparsity. Despite the convenience of using Bayesian methods to depict hierarchical structure among variables, such posteriors are usually intractable. Hence, current literatures mainly rely on sampling methods to approximate the posterior distribution, such as Markov chain Monte Carlo (MCMC) (Xu and Ghosh 2015; Chen et al. 2016). The computational costs of these methods become very expensive in the presence of a large number of variables.

In this article, we propose a scalable Bayesian method for bi-level variable selection (BIVAS). Instead of using MCMC, we adopt variational inference, which greatly reduces computational cost and makes our algorithm scalable. In contrast to standard mean-field variational approximation, we propose a hierarchically factorizable approximation, making use of the special structure of bi-level variable selection. A computationally efficient variational expectation-maximization (EM) algorithm is developed to handle large datasets. Moreover, we extend our approach to handle a class of multitask learning. We further use comprehensive simulation studies to demonstrate that BIVAS can significantly outperform its alternatives in term of variable selection, prediction accuracy and computational efficiency.

The remainder of this article is organized as follows. In Section 2, we describe both model settings and algorithms. In particular, we show the rationale to improve the computational efficiency. We further discuss the way of extending our approach to multitask learning. In Section 3, we evaluated the performance of BIVAS based on comprehensive simulation studies, especially checked the cases variational assumptions are violated. The experimental results show that BIVAS can stably outperform its alternatives in various settings. Then we applied BIVAS to three real data examples. We conclude the article with a short discussion in Section 4.

## 2. Methods

### 2.1. Regression With BIVAS

#### 2.1.1. Model Setting

Suppose we have collected dataset  $\{\mathbf{y}, \mathbf{Z}, \mathbf{X}\}$  with sample size  $n$ , where  $\mathbf{y} \in \mathbb{R}^n$  is the vector of response variable,  $\mathbf{Z} \in \mathbb{R}^{n \times r}$  is the design matrix of  $r$  columns including an intercept and a few covariates ( $r < n$ ), and  $\mathbf{X} \in \mathbb{R}^{n \times p}$  is the design matrix of  $p$  predictors. Besides, each of the  $p$  variables in  $\mathbf{X}$  is labeled with one of  $K$  known groups, where the number of variables in group  $k$  is denoted by  $l_k$  and  $\sum_{k=1}^K l_k = p$ . We consider the following linear model that links  $\mathbf{y}$  to  $\mathbf{Z}$  and  $\mathbf{X}$

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\omega} + \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \quad (1)$$

where  $\boldsymbol{\omega} \in \mathbb{R}^r$  is a vector of fixed effects,  $\boldsymbol{\beta} \in \mathbb{R}^p$  is a vector of random effects, and  $\mathbf{e} \in \mathbb{R}^n$  is a vector of independent noise. We assume  $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma_e^2 \mathbf{I}_n)$ , where  $\mathbf{I}_n$  is the  $n$ -by- $n$  identity matrix. Under this model, the bi-level selection aims to identify nonzero entries of  $\boldsymbol{\beta}$  at both the group and individual-variable levels. For this reason, we introduce two binary variables:  $\eta_k$  indicates

whether group  $k$  is active ( $\eta_k = 1$ ) or not ( $\eta_k = 0$ ); and  $\gamma_{jk}$  indicates whether the  $j$ th variable in group  $k$  is zero ( $\gamma_{jk} = 0$ ) or not ( $\gamma_{jk} = 1$ ). Hence, we introduce a bi-level spike-slab prior on  $\boldsymbol{\beta}$

$$\beta_{jk} | \eta_k, \gamma_{jk}; \sigma_\beta^2 \sim \begin{cases} \mathcal{N}(\beta_{jk} | 0, \sigma_\beta^2) & \text{if } \eta_k = 1, \gamma_{jk} = 1, \\ \delta_0(\beta_{jk}) & \text{otherwise,} \end{cases} \quad (2)$$

where  $\mathcal{N}(\beta_{jk} | 0, \sigma_\beta^2)$  denotes the Gaussian distribution with mean 0 and variance  $\sigma_\beta^2$  and  $\delta_0(\beta_{jk})$  denotes a Dirac function at zero. This bi-level structure means that  $\beta_{jk}$  is drawn from  $\mathcal{N}(0, \sigma_\beta^2)$  if and only if both the  $k$ th group and its  $j$ th variable are included in the model. Let  $\Pr(\eta_k = 1) = \pi$  and  $\Pr(\gamma_{jk} = 1) = \alpha$  be the prior inclusion probability of groups and variables, respectively.

The presence of Dirac function may introduce additional troubles in algorithm derivation. To get rid of the Dirac function, we reparameterize the model as following

$$\begin{aligned} \beta_{jk} | \sigma_\beta^2 &\sim \mathcal{N}(0, \sigma_\beta^2), \quad \gamma_{jk} | \alpha \sim \alpha^{\gamma_{jk}} (1 - \alpha)^{1 - \gamma_{jk}}, \\ \eta_k | \pi &\sim \pi^{\eta_k} (1 - \pi)^{1 - \eta_k}. \end{aligned} \quad (3)$$

Consequently, the prior of  $\beta_{jk}$  does not depend on  $\gamma_{jk}$  and  $\eta_k$  any more, and the product  $\eta_k \gamma_{jk} \beta_{jk}$  form a new random variable exactly distributed as given in (2). After reparameterization, model (1) becomes

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\omega} + \sum_{k=1}^K \sum_{j=1}^{l_k} \eta_k \gamma_{jk} \beta_{jk} \mathbf{x}_{jk} + \mathbf{e}.$$

We shall use this reparameterized version through the article.

Let  $\boldsymbol{\theta} = \{\alpha, \pi, \sigma_\beta^2, \sigma_e^2, \boldsymbol{\omega}\}$  be the collection of model parameters and  $\{\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\eta}\}$  be the set of latent variables. The joint probabilistic model is

$$\begin{aligned} \Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta} | \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}) &= \Pr(\mathbf{y} | \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}) \Pr(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta} | \boldsymbol{\theta}) \\ &= \mathcal{N}(\mathbf{y} | \mathbf{Z}\boldsymbol{\omega} + \sum_{k=1}^K \sum_{j=1}^{l_k} \eta_k \gamma_{jk} \beta_{jk} \mathbf{x}_{jk}, \sigma_e^2) \prod_{k=1}^K \pi^{\eta_k} (1 - \pi)^{1 - \eta_k} \\ &\quad \times \prod_{j=1}^{l_k} \mathcal{N}(0, \sigma_\beta^2) \alpha^{\gamma_{jk}} (1 - \alpha)^{1 - \gamma_{jk}}, \end{aligned} \quad (4)$$

where  $\mathbf{x}_{jk}$  is a column of  $\mathbf{X}$  corresponding to the  $j$ th variable in the  $k$ th group. Instead of applying a full Bayesian approach by introducing priors to  $\boldsymbol{\theta}$ , we treat  $\boldsymbol{\theta}$  as fixed values and adopt the empirical Bayes framework (Efron 2012; Bishop 2006; Carbonetto and Stephens 2012; MacKay 2003). Specifically, we first estimate the unknown parameters  $\boldsymbol{\theta}$  from the data and then evaluate the posterior distributions of  $\boldsymbol{\beta}$ ,  $\boldsymbol{\gamma}$ , and  $\boldsymbol{\eta}$  given the estimate  $\hat{\boldsymbol{\theta}}$ . To estimate  $\boldsymbol{\theta}$ , we first optimize the marginal likelihood

$$\log \Pr(\mathbf{y} | \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}) = \log \sum_{\boldsymbol{\gamma}} \sum_{\boldsymbol{\eta}} \int_{\boldsymbol{\beta}} \Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta} | \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}) d\boldsymbol{\beta}. \quad (5)$$

Given  $\hat{\boldsymbol{\theta}}$ , the posterior is given as

$$\Pr(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta} | \mathbf{y}, \mathbf{X}, \mathbf{Z}; \hat{\boldsymbol{\theta}}) = \frac{\Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta} | \mathbf{X}, \mathbf{Z}; \hat{\boldsymbol{\theta}})}{\Pr(\mathbf{y} | \mathbf{X}, \mathbf{Z}; \hat{\boldsymbol{\theta}})}. \quad (6)$$

### 2.1.2. Algorithm

Conventionally, the model involving latent variables is often solved by the EM algorithm. However, the standard EM algorithm cannot be applied here due to the difficulty of the E-step caused by the combinatorial nature of  $\gamma$  and  $\eta$ . Alternatively, we propose a variational EM algorithm via approximate Bayesian inference (Bishop 2006).

To apply variational approximation, we first define  $q(\eta, \gamma, \beta)$  as an approximated distribution of posterior  $\Pr(\eta, \gamma, \beta | \mathbf{y}, \mathbf{X}, \mathbf{Z}; \theta)$ . Then we can obtain the lower bound of log-marginal likelihood by Jensen's inequality

$$\begin{aligned} \log p(\mathbf{y} | \mathbf{X}, \mathbf{Z}; \theta) &= \log \sum_{\gamma} \sum_{\eta} \int_{\beta} \Pr(\mathbf{y}, \eta, \gamma, \beta | \mathbf{X}, \mathbf{Z}; \theta) d\beta \\ &\geq \sum_{\gamma} \sum_{\eta} \int_{\beta} q(\eta, \gamma, \beta) \log \frac{\Pr(\mathbf{y}, \eta, \gamma, \beta | \mathbf{X}, \mathbf{Z}; \theta)}{q(\eta, \gamma, \beta)} d\beta \quad (7) \\ &= \mathbb{E}_q[\log \Pr(\mathbf{y}, \eta, \gamma, \beta | \mathbf{X}, \mathbf{Z}; \theta) - \log q(\eta, \gamma, \beta)] \\ &= \mathcal{L}(q), \end{aligned}$$

where the equality holds if and only if  $q(\eta, \gamma, \beta) = \Pr(\eta, \gamma, \beta | \mathbf{y}, \mathbf{X}, \mathbf{Z}; \theta)$ . Then, we can iteratively maximize  $\mathcal{L}(q)$  instead of working with the marginal likelihood directly. Conventionally,  $q$  is often assumed to be fully factorizable based on the mean-field theory (Bishop 2006). As there is hierarchical structure between the group level and the variable level, here we propose a novel variational distribution to accommodate the bi-level variable selection. Specifically, we consider the following hierarchically structured distribution as an approximation to posterior  $\Pr(\eta, \gamma, \beta | \mathbf{y}, \mathbf{X}, \mathbf{Z})$

$$q(\eta, \gamma, \beta) = \prod_k^K \left( q(\eta_k) \prod_j^{l_k} (q(\beta_{jk} | \eta_k, \gamma_{jk}) q(\gamma_{jk})) \right). \quad (8)$$

Without any other assumptions, we can show (with details in supplementary materials) that the optimal solution of  $q$  is given as

$$\begin{aligned} q(\eta, \gamma, \beta) &= \prod_k^K \left( \pi_k^{\eta_k} (1 - \pi_k)^{1 - \eta_k} \prod_j^{l_k} \right. \\ &\quad \times \left. \left( \alpha_{jk}^{\gamma_{jk}} (1 - \alpha_{jk})^{1 - \gamma_{jk}} \mathcal{N}(\mu_{jk}, s_{jk}^2)^{\eta_k \gamma_{jk}} \mathcal{N}(0, \sigma_{\beta}^2)^{1 - \eta_k \gamma_{jk}} \right) \right), \quad (9) \end{aligned}$$

where

$$\begin{aligned} s_{jk}^2 &= \frac{\sigma_e^2}{\mathbf{x}_{jk}^T \mathbf{x}_{jk} + \frac{\sigma_e^2}{\sigma_{\beta}^2}}, \\ \mu_{jk} &= \left( \mathbf{x}_{jk}^T (\mathbf{y} - \mathbf{Z}\omega) - \sum_{k' \neq k}^K \sum_j^{l_{k'}} \mathbb{E}_{j'k'} [\eta_{k'} \gamma_{j'k'} \beta_{j'k'}] \mathbf{x}_{j'k'}^T \mathbf{x}_{jk} \right. \\ &\quad \left. - \sum_{j' \neq j}^{l_k} \mathbb{E}[\gamma_{j'k} \beta_{j'k}] \mathbf{x}_{j'k}^T \mathbf{x}_{jk} \right) / \left( \mathbf{x}_{jk}^T \mathbf{x}_{jk} + \frac{\sigma_e^2}{\sigma_{\beta}^2} \right) \quad (10) \end{aligned}$$

$$\begin{aligned} \pi_k &= \frac{1}{1 + \exp(-u_k)}, \\ \text{with } u_k &= \log \frac{\pi}{1 - \pi} + \frac{1}{2} \sum_j^{l_k} \alpha_{jk} \left( \log \frac{s_{jk}^2}{\sigma_{\beta}^2} + \frac{\mu_{jk}^2}{s_{jk}^2} \right), \quad (11) \end{aligned}$$

$$\begin{aligned} \alpha_{jk} &= \frac{1}{1 + \exp(-v_{jk})}, \\ \text{with } v_{jk} &= \log \frac{\alpha}{1 - \alpha} + \frac{1}{2} \pi_k \left( \log \frac{s_{jk}^2}{\sigma_{\beta}^2} + \frac{\mu_{jk}^2}{s_{jk}^2} \right). \quad (12) \end{aligned}$$

By inspections of Equations (8) and (9), we have  $q(\eta_k = 1) = \pi_k$  and  $q(\gamma_{jk} = 1) = \alpha_{jk}$ , which can be viewed as approximations to the posterior distributions  $\Pr(\eta_k = 1 | \mathbf{y}, \mathbf{X}, \mathbf{Z}; \theta)$  and  $\Pr(\gamma_{jk} = 1 | \mathbf{y}, \mathbf{X}, \mathbf{Z}; \theta)$ , respectively. Similarly,  $q(\beta_{jk} | \eta_k \gamma_{jk} = 1) = \mathcal{N}(\mu_{jk}, s_{jk}^2)$  can be interpreted as the variational approximation to  $\Pr(\beta_{jk} | \eta_k \gamma_{jk} = 1, \mathbf{y}, \mathbf{X}, \mathbf{Z}; \theta)$ , which is the conditional posterior distribution of  $\beta_{jk}$  given it is selected in both the group level and the variable level. Accordingly,  $q(\beta_{jk} | \eta_k \gamma_{jk} = 0) = \mathcal{N}(0, \sigma_{\beta}^2)$  approximates  $\Pr(\beta_{jk} | \eta_k \gamma_{jk} = 0, \mathbf{y}, \mathbf{X}, \mathbf{Z}; \theta)$ , corresponding to the case when  $\beta_{jk}$  is irrelevant in either of the two levels.

Note that the form of variational parameters provides an intuitive interpretation. Group-level posterior inclusion probability  $\pi_k$  and variable-level posterior inclusion probability  $\alpha_{jk}$  can be viewed as their prior inclusion probability ( $\pi, \alpha$ ) updated by data-driven information. Furthermore,  $\pi_k$  and  $\alpha_{jk}$  are inter-dependent. On one hand, if more and more  $\alpha_{jk}$  within the  $k$ th group become closer to one, then  $\pi_k$  will be closer to one, as seen in Equation (11). On the other hand, if  $\pi_k$  increases, then the variables in the  $k$ th group are more likely to be selected, see Equation (12).

With Equation (9), the lower bound  $\mathcal{L}(q)$  can be evaluated analytically. By setting the derivative of  $\mathcal{L}(q)$  with respect to  $\theta$  to be zero, we have the updating equations for parameter estimation

$$\begin{aligned} \sigma_e^2 &= \frac{\|\mathbf{y} - \mathbf{Z}\omega - \sum_k^K \sum_j^{l_k} \pi_k \alpha_{jk} \mu_{jk} \mathbf{x}_{jk}\|^2}{n} \\ &\quad + \frac{\sum_k^K \sum_j^{l_k} [\pi_k \alpha_{jk} (s_{jk}^2 + \mu_{jk}^2) - (\pi_k \alpha_{jk} \mu_{jk})^2] \mathbf{x}_{jk}^T \mathbf{x}_{jk}}{n} \\ &\quad + \frac{\sum_k^K (\pi_k - \pi_k^2) [\sum_j^{l_k} \sum_{j'}^{l_k} \alpha_{j'k} \mu_{j'k} \alpha_{jk} \mu_{jk}] \mathbf{x}_{j'k}^T \mathbf{x}_{jk}}{n}, \\ \sigma_{\beta}^2 &= \frac{\sum_k^K \sum_j^{l_k} \pi_k \alpha_{jk} (s_{jk}^2 + \mu_{jk}^2)}{\sum_k^K \sum_j^{l_k} \pi_k \alpha_{jk}}, \\ \alpha &= \frac{1}{p} \sum_k^K \sum_j^{l_k} \alpha_{jk}, \\ \pi &= \frac{1}{K} \sum_k^K \pi_k, \\ \omega &= (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T (\mathbf{y} - \sum_k^K \sum_j^{l_k} \pi_k \alpha_{jk} \mu_{jk} \mathbf{x}_{jk}). \quad (13) \end{aligned}$$

To summarize, the algorithm can be regarded as a variational extension of the EM algorithm. At E-step, the lower bound

$\mathcal{L}(q)$  is obtained by evaluating the expectation w.r.t variational posterior  $q$ . At M-step, the current  $\mathcal{L}(q)$  is optimized w.r.t model parameters in  $\theta$ . As a result, the lower bound increases at each iteration and the convergence is guaranteed.

## 2.2. Multitask Learning With BIVAS

In this section, we consider bi-level variable selection in multitask learning. In real applications, some related regression tasks may have similar patterns in the effects of predictor variables. A joint model that analyze all such related tasks simultaneously can efficiently increase statistical power, which is called multitask learning (Caruana 1997). As we shall see later, a class of multitask regression problem can be naturally solved by BIVAS with proper adjustment for the likelihood. To avoid ambiguity, we refer to the model described in Section 2.1 as “group BIVAS” and the one discussed in this section as “multitask BIVAS.”

Suppose we have collected dataset  $\{\mathbf{y}, \mathbf{Z}, \mathbf{X}\} = \{\mathbf{y}_j, \mathbf{Z}_j, \mathbf{X}_j\}_{j=1}^L$  from  $L$  related regression tasks, each of which has sample size  $n_j$ . In practice,  $\mathbf{y}_j \in \mathbb{R}^{n_j}$  is the response vector of  $j$ th task from  $n_j$  individuals;  $\mathbf{Z}_j \in \mathbb{R}^{n_j \times r}$  includes an intercept and a few shared covariates;  $\mathbf{X}_j \in \mathbb{R}^{n_j \times K}$  is the design matrix of  $K$  shared predictors. We relate  $\mathbf{y}_j$  to  $\mathbf{X}_j$  and  $\mathbf{Z}_j$  using the following linear mixed model

$$\mathbf{y}_j = \mathbf{Z}_j \boldsymbol{\omega}_j + \mathbf{X}_j \boldsymbol{\beta}_j + \mathbf{e}_j, \quad j = 1, \dots, L, \quad (14)$$

where  $\boldsymbol{\omega}_j \in \mathbb{R}^r$  is the vector of fixed effects,  $\boldsymbol{\beta}_j \in \mathbb{R}^K$  is the vector of random effects, and  $\mathbf{e}_j \in \mathbb{R}^{n_j}$  is the vector of independent noise with  $\mathbf{e}_j \sim \mathcal{N}(\mathbf{0}, \sigma_{e_j}^2 \mathbf{I}_{n_j})$ . For convenience, we denote  $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_L] \in \mathbb{R}^{K \times L}$  and  $\beta_{jk}$  be  $k$ th entry in  $\boldsymbol{\beta}_j$ . Clearly, it is not reasonable to assume that all shared predictors are relevant to all responses, especially when  $K$  is large. A more reasonable assumption is that the majority of predictors are irrelevant to all the responses and only a few of them are relevant with many responses. With this assumption, it is natural to treat each shared predictor as a group across different task  $l$ , which corresponds to a row of  $\boldsymbol{\beta}$ . Then the group-level selection aims at excluding variables which are irrelevant to all responses and the individual-level selection further identifies fine-grained relevance between variables and response of specific task. For this purpose, we introduce two binary variables:  $\eta_k$  indicates whether the  $k$ th row of  $\boldsymbol{\beta}$  is active or not and  $\gamma_{jk}$  indicates whether  $\beta_{jk}$  is zero or not. Then the bi-level spike-slab prior on  $\boldsymbol{\beta}$  is introduced by

$$\beta_{jk} | \eta_k, \gamma_{jk}; \sigma_{\beta_j}^2 \sim \begin{cases} \mathcal{N}(\beta_{jk} | 0, \sigma_{\beta_j}^2) & \text{if } \eta_k = 1, \gamma_{jk} = 1, \\ \delta_0(\beta_{jk}) & \text{otherwise,} \end{cases} \quad (15)$$

where prior inclusion probabilities are defined as  $\Pr(\eta_k = 1) = \pi$  and  $\Pr(\gamma_{jk} = 1) = \alpha$ .

Again we reparameterize the model to remove the Dirac function

$$\begin{aligned} \beta_{jk} | \sigma_{\beta_j}^2 &\sim \mathcal{N}(0, \sigma_{\beta_j}^2), \quad \gamma_{jk} | \alpha \sim \alpha^{\gamma_{jk}} (1 - \alpha)^{1 - \gamma_{jk}}, \\ \eta_k | \pi &\sim \pi^{\eta_k} (1 - \pi)^{1 - \eta_k}. \end{aligned} \quad (16)$$

Let  $\theta = \{\alpha, \pi, \sigma_{\beta_j}^2, \sigma_{e_j}^2, \boldsymbol{\omega}_j\}_{j=1}^L$  be the collection of parameters under the multitask model. Our goal is to maximize the

marginal likelihood, which is of the same form as Equation (5), and evaluate the posterior distribution of  $\beta_{jk}$ . The variational EM algorithm of multitask BIVAS is straightforward following the similar procedure in Section 2.1.2. We leave the detailed derivations in Section 2 of the supplementary materials.

## 2.3. Implementation Details

After the convergence of algorithm, we can approximate the posterior inclusion probabilities by the variational approximation. For group BIVAS, the approximations are given by

$$\Pr(\eta_k = 1 | \mathbf{y}, \mathbf{X}, \mathbf{Z}; \hat{\theta}) \approx q(\eta_k = 1 | \hat{\theta}) = \pi_k,$$

$$\Pr(\gamma_{jk} = 1 | \mathbf{y}, \mathbf{X}, \mathbf{Z}; \hat{\theta}) \approx q(\gamma_{jk} = 1 | \hat{\theta}) = \alpha_{jk}.$$

These evaluations are based on parameter estimates  $\hat{\theta}$ . However, as there is no guarantee of global optimal for the EM algorithm, the choice of initial value  $\theta^0$  is critical. A bad initial value will lead to a poor  $\hat{\theta}$ . In our model, due to the existence of multiple latent variables ( $\boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\gamma}$ ), choosing a good initial value could be challenging. Here we consider the importance sampling suggested by varbvs (Carbonetto and Stephens 2012): we further introduce a prior over  $\theta$  and integrate over the value of  $\theta$  to obtain the final evaluations. In contrast to varbvs, we introduce prior only on the group sparsity parameter  $\pi$ . We first select  $h$  values of  $\pi$  ( $\{\pi(i)\}_{i=1}^h$ ) such that  $\log_{10}$  odds of  $\pi(i)$  is uniformly distributed on  $[-\log_{10}(K), 0]$  which encourages group sparsity. With this additional setting, the new collection of parameters is then defined as  $\theta' = \{\theta'_i\}_{i=1}^h$  with  $\theta'_i = \{\alpha, \pi(i), \sigma_{\beta_j}^2, \sigma_{e_j}^2, \boldsymbol{\omega}_j\}$ ; and the posterior inclusion probability can be approximated as follows

$$\begin{aligned} \Pr(\eta_k = 1 | \mathbf{y}, \mathbf{X}, \mathbf{Z}) &\approx \int q(\eta_k = 1 | \theta') \Pr(\theta' | \mathbf{y}, \mathbf{X}, \mathbf{Z}) d\theta' \\ &\approx \frac{\sum_{i=1}^h q(\eta_k = 1 | \theta'_i) w(\theta'_i)}{\sum_{i=1}^h w(\theta'_i)}, \\ \Pr(\gamma_{jk} = 1 | \mathbf{y}, \mathbf{X}, \mathbf{Z}) &\approx \int q(\gamma_{jk} = 1 | \theta') \Pr(\theta' | \mathbf{y}, \mathbf{X}, \mathbf{Z}) d\theta' \\ &\approx \frac{\sum_{i=1}^h q(\gamma_{jk} = 1 | \theta'_i) w(\theta'_i)}{\sum_{i=1}^h w(\theta'_i)}, \end{aligned} \quad (17)$$

where  $w(\theta'_i)$  is the unnormalized importance weight for  $i$ th component. For each of the two equations in (20), the first approximation is due to the variational inference; the second approximation is due to the importance sampling. Besides,  $w(\theta'_i)$  can be approximated by exponential of  $\mathcal{L}(q)$  given  $\theta'_i$  since  $\mathcal{L}(q)$  takes similar shape to  $\log \Pr(\mathbf{y} | \mathbf{X}, \mathbf{Z}; \theta)$  when the marginal likelihood is relatively large (Carbonetto and Stephens 2012). Hence, we can derive the final evaluation of posteriors

$$\begin{aligned} \Pr(\eta_k = 1 | \mathbf{y}, \mathbf{X}, \mathbf{Z}) &\approx \sum_{i=1}^h \pi_k(i) \cdot \tilde{w}(i) \equiv \tilde{\pi}_k, \\ \Pr(\gamma_{jk} = 1 | \mathbf{y}, \mathbf{X}, \mathbf{Z}) &\approx \sum_{i=1}^h \alpha_{jk}(i) \cdot \tilde{w}(i) \equiv \tilde{\alpha}_{jk}, \\ \mathbb{E}(\beta_{jk} | \eta_k \gamma_{jk} = 1, \mathbf{y}, \mathbf{X}, \mathbf{Z}) &\approx \sum_{i=1}^h \mu_{jk}(i) \cdot \tilde{w}(i) \equiv \tilde{\mu}_{jk}, \\ \mathbb{E}(\eta_k \gamma_{jk} \beta_{jk} | \mathbf{y}, \mathbf{X}, \mathbf{Z}) &\approx \tilde{\pi}_k \tilde{\alpha}_{jk} \tilde{\mu}_{jk}, \end{aligned} \quad (18)$$



where

$$\tilde{w}(i) = \exp(w(\theta'_i) - m), \quad m = \max(w(\theta'_i)).$$

Here we handle the normalization inside the exponential so that the calculation is numerically stable. The same weighting evaluation applies to the parameters  $\theta'$ . We can derive the same procedure for multitask BIVAS as the one for group BIVAS. Although we need to run EM algorithm  $h$  times in this procedure, each EM algorithm becomes more stable and converges in less iterations. In practice,  $h = 20\text{--}40$  is often good enough for large scale datasets. Furthermore, taking the advantage of independence among  $\pi(i)$ 's, the  $h$  procedures can be fully parallelized. Common solutions to parallelization are based on APIs such as OpenMP. These solutions, however, usually require the tasks to be allocated beforehand. In our model, this restriction may lead to inefficiency because the time of convergence for each procedure can be very different. Thus, we adopt a dynamic threading technique that can immediately allocate a new task to a thread once it has finished an old task. This technique greatly improves the efficiency of parallelization compared to OpenMP.

Given the prefixed values of  $\pi$ , we only need to consider the initialization of  $\alpha$ , which is less sensitive. In practice, we recommend to initialize  $\alpha$  as  $\frac{1}{\pi \log(p+1)}$  that prefers a sparse model. Through comprehensive simulations in Section 3, we showed this setting has good performance in identifying variables and groups as well as providing stable posterior inclusion probabilities  $\gamma_{jk}$  and  $\eta_k$  under various sparsity conditions. We summarize the variational EM algorithm with importance sampling in Algorithm 1.

#### 2.4. Variable Selection and Prediction

With the results obtained by importance sampling, we extract information from our model for the purpose of variable selection and prediction. Using the approximation of the posterior inclusion probability in (21), we can approximate local false discovery rate (fdr) of group  $k$  by  $\text{fdr}_k = 1 - \tilde{\pi}_k$  and fdr of  $j$ th variable in  $k$ th group by  $\text{fdr}_{jk} = 1 - \tilde{\alpha}_{jk}$  (Efron 2005). The local fdr is also known as “posterior exclusion probability.” To control the global false discovery rate (FDR), we first sort the variables (groups) by fdr in increasing order and then identify the  $j$ th reordered variable (group) as active if

$$\text{FDR}_{(j)} = \frac{\sum_{i=1}^j \text{fdr}_{(i)}}{j} \leq \tau, \quad (19)$$

where  $\text{fdr}_{(i)}$  is the  $i$ th ordered fdr,  $\text{FDR}_{(j)}$  is the corresponding global FDR, and  $\tau$  is the selected threshold that the global FDR is controlled at (e.g., 0.1). One can either control the local fdr (e.g.,  $\text{fdr}_{jk} \leq 0.1$ ) or the global FDR with  $\tau = 0.1$ . Although the parameter estimates may not be accurate due to the variational approximation, the posterior means of latent variables appear to be accurate. We will verify this result later in the simulation.

In addition to variable selection, we can also predict  $\hat{y}$  (or  $\hat{y}_j$  for multitask learning) with a new data  $\{\mathbf{Z}^{\text{new}}, \mathbf{X}^{\text{new}}\}$ . Since  $\mathbb{E}_q(\eta_k \gamma_{jk} \beta_{jk}) \approx \tilde{\pi}_k \tilde{\alpha}_{jk} \tilde{\mu}_{jk}$  gives the estimate of effect size for the  $jk$ th random effect, the predicted value is simply obtained by

---

#### Algorithm 1 Variational EM algorithm with importance sampling for group BIVAS

---

**Inputs:**  $\mathbf{Z}, \mathbf{X}, \mathbf{y}$ , and  $\boldsymbol{\pi} = \{\pi(i)\}_{i=1}^h$ .

**Outputs:**  $\tilde{\pi}_k, \tilde{\alpha}_{jk}, \tilde{\mu}_{jk}$

**Parameter Initialization:**  $\boldsymbol{\omega} = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{y}$ ,  $\sigma_e^2 = \sigma_\beta^2 = \text{var}(\mathbf{y})/2$ ,  $\mu_{jk} = 0$ .

**for**  $i = 1, \dots, h$  **do**

Set  $\pi = \pi(i)$ . Initialize  $\alpha = 1/\pi(\log(p+1))$ ,  $\pi_k = \pi$ ,  $\alpha_{jk} = \alpha$ , for all  $k$ . Let  $\mathbf{y}^b = \sum_k^K \sum_j^{l_k} \pi_k \alpha_{jk} \mu_{jk} \mathbf{x}_{jk}$ .

**repeat**

**E-step:**

**for**  $k = 1, \dots, K$  **do**

Let  $\mathbf{y}^w = \sum_j^{l_k} \alpha_{jk} \mu_{jk} \mathbf{x}_{jk}$ ;  $\mathbf{y}_k^b = \mathbf{y}^b - \pi_k \mathbf{y}^w$ .

**for**  $j = 1, \dots, l_k$  **do**

Let  $\mathbf{y}_{jk}^w = \mathbf{y}^w - \alpha_{jk} \mu_{jk} \mathbf{x}_{jk}$ .

Update variable-level variational parameters:

$$s_{jk}^2 = \frac{\sigma_e^2}{\mathbf{x}_{jk}^T \mathbf{x}_{jk} + \frac{\sigma_e^2}{\sigma_\beta^2}}, \quad \mu_{jk} = \frac{\mathbf{x}_{jk}^T (\mathbf{y} - \mathbf{Z}\boldsymbol{\omega} - \mathbf{y}_k^b - \mathbf{y}_{jk}^w)}{\mathbf{x}_{jk}^T \mathbf{x}_{jk} + \frac{\sigma_e^2}{\sigma_\beta^2}}$$

$$\alpha_{jk} = \frac{1}{1 + \exp(-v_{jk})} \quad \text{with}$$

$$v_{jk} = \log \frac{\alpha}{1 - \alpha} + \frac{1}{2} \pi_k \left( \log \frac{s_{jk}^2}{\sigma_\beta^2} + \frac{\mu_{jk}^2}{s_{jk}^2} \right)$$

Set  $\mathbf{y}^w = \mathbf{y}_{jk}^w + \alpha_{jk} \mu_{jk} \mathbf{x}_{jk}$ .

**end for**

Update group-level variational parameters:

$$\pi_k = \frac{1}{1 + \exp(-u_k)} \quad \text{with}$$

$$u_k = \log \frac{\pi}{1 - \pi} + \frac{1}{2} \sum_j^{l_k} \alpha_{jk} \left( \log \frac{s_{jk}^2}{\sigma_\beta^2} + \frac{\mu_{jk}^2}{s_{jk}^2} \right).$$

Set  $\mathbf{y}^b = \mathbf{y}_k^b + \pi_k \mathbf{y}^w$ .

**end for**

**M-step:** Update parameters  $\sigma_e^2, \sigma_\beta^2, \alpha$  and  $\boldsymbol{\omega}$  according to (13).

**until** The lower bound stops increasing or the maximum iteration is reached.

**end for**

**Importance sampling** Reweight the obtained parameter estimates according to (21).

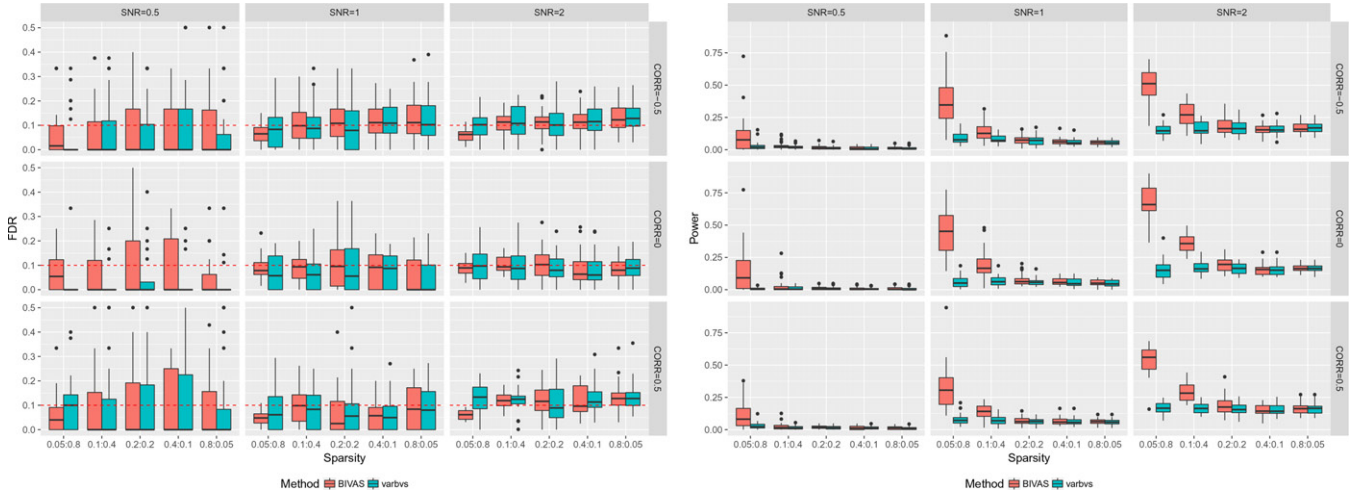
---

$$\hat{y} = \sum_r \tilde{\omega}_r z_r^{\text{new}} + \sum_k \sum_j \tilde{\pi}_k \tilde{\alpha}_{jk} \tilde{\mu}_{jk} x_{jk}^{\text{new}} \quad (\text{in multitask learning})$$

$$\hat{y}_j = \sum_r \tilde{\omega}_{jr} z_{jr}^{\text{new}} + \sum_k \tilde{\pi}_k \tilde{\alpha}_{jk} \tilde{\mu}_{jk} x_{jk}^{\text{new}} \quad (\text{for } j\text{th task}).$$

### 3. Numerical Examples

In this section, we gauged the performance of BIVAS in comparison with alternative methods using both simulation and real data analysis. The experiments and analyses were conducted using R software (R Core Team 2018) and figures were generated using R packages “ggplot2” (Wickham 2016), “qqman” (Turner 2018), and “wordcloud” (Fellows 2018). In the spirit of



**Figure 1.** Comparison of BIVAS and varbvs for individual variable selection. We controlled the global FDR at the nominal level 0.1 (the red dashed line in the left panel) to evaluate the FDR and power. Left: Boxplots of the true FDR. Right: Boxplots of the power. Each inner panel corresponds to a specific simulation setting. BIVAS outperforms varbvs when group-level sparsity dominates.

reproducibility, all the simulation codes are made publicly available at <https://github.com/mxcai/sim-bivas>.

### 3.1. Simulation Study

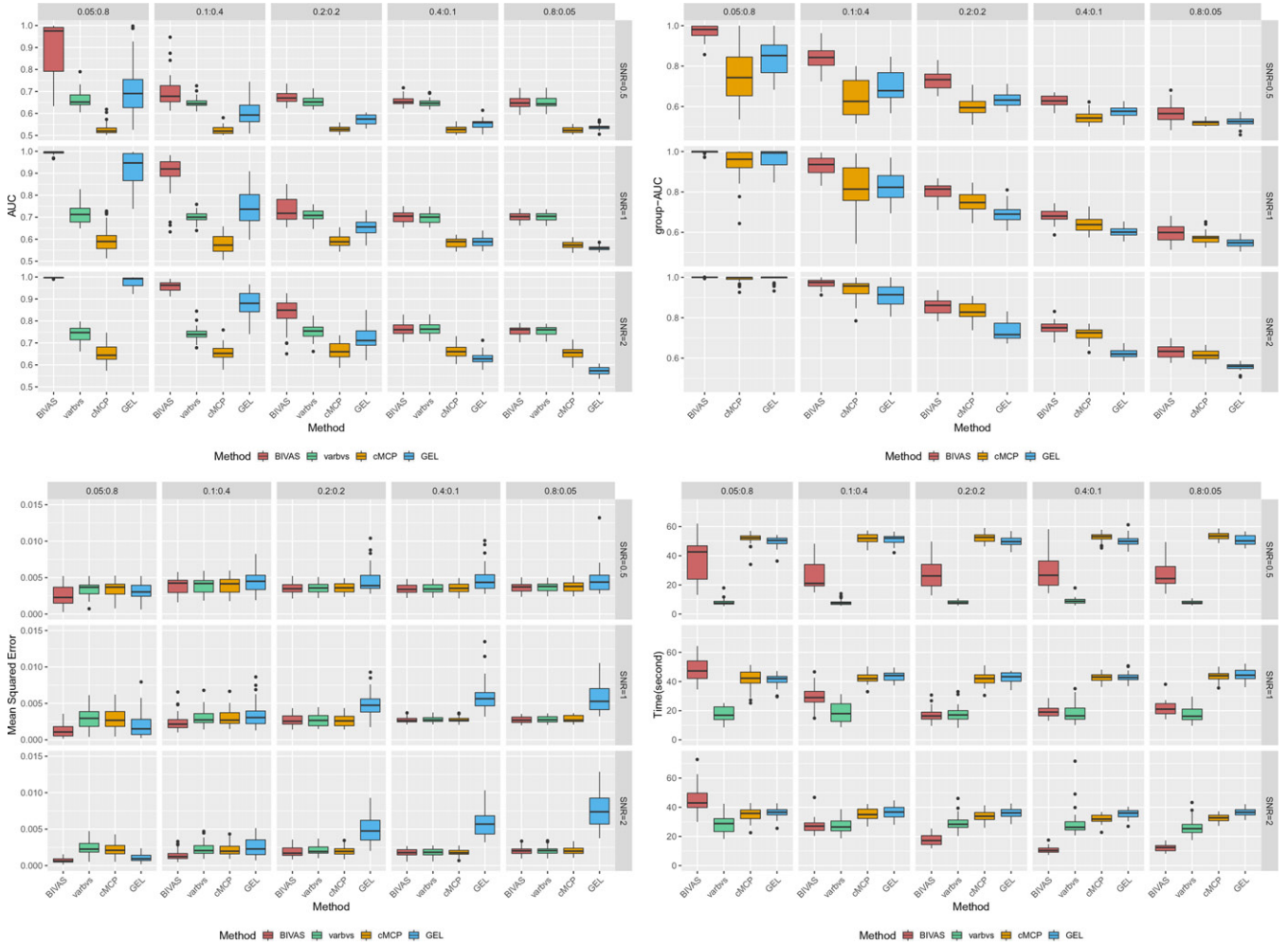
For group BIVAS, we compared it with varbvs (Carbonetto and Stephens 2012), cMCP (Breheny and Huang 2009), and GEL (Breheny 2015). The simulation datasets were generated as follows. For all settings, we fixed  $n = 1000$ ,  $p = 5,000$ ,  $K = 250$  with 20 variables in each group. The design matrix  $\mathbf{X}$  was generated from normal distribution  $\mathcal{N}(\mathbf{0}, \Sigma(\rho))$ , where  $\Sigma_{jj'} = \rho^{|j-j'|}$  if variable  $j$  and  $j'$  are in the same group and  $\Sigma_{jj'} = 0$  for  $j$  and  $j'$  from different groups. Therefore,  $\rho$  quantifies the within-group autoregressive correlation. As the variational approximation assumes a hierarchically factorizable distribution, we selected  $\rho \in \{-0.5, 0, 0.5\}$  to evaluate the influence of violation of this assumption (extreme cases with  $\rho > 0.9$  are provided in Section 4 of the supplementary materials). Next, we generated  $\eta_k$  and  $\gamma_{jk}$  from Bernoulli distributions with different settings of Bernoulli parameters:  $(\pi, \alpha) \in \{(0.05, 0.8), (0.1, 0.4), (0.2, 0.2), (0.4, 0.1), (0.8, 0.05)\}$ . Note that the total sparsity was fixed at  $\alpha \cdot \pi = 0.04$  for different combinations of  $\pi$  and  $\alpha$ . Given the true status  $\eta_k$  and  $\gamma_{jk}$ , we then simulated true  $\beta$  using generative model (1). Finally, we obtained the response vector  $\mathbf{y}$  by controlling the signal-to-noise ratio (SNR) at  $\text{SNR} = \text{var}(\mathbf{X}\beta)/\sigma_e^2 \in \{0.5, 1, 2\}$ . All results were summarized from 30 replications.

To make comparisons between different methods, we first define the metrics to be evaluated. The true FDR is measured as  $\text{FDR} = \frac{\text{No. of false discoveries}}{\text{total no. of discoveries}}$ . The statistical power measures the ability to identify active variables, which is evaluated as  $\text{power} = \frac{\text{No. of true discoveries}}{\text{total no. of active variables}}$ . As cMCP and GEL did not provide fdr estimates, we are unable to report their performances based on FDR. Therefore, we first compared BIVAS with varbvs based on the measures of FDR and power. To evaluate these metrics, we first obtained the global FDR from local fdr through the procedure in Equation (19), and then identified active variables

by controlling the global FDR at nominal level 0.1. Figure 1 shows the performance of FDR control and statistical power for individual variable selection obtained by BIVAS and varbvs. When the SNR is small, both methods are underpowered. However, BIVAS gains more power as the group sparsity dominates and further enlarges the gap as SNR increases. As  $\rho$  moves away from zero, empirical FDRs of both methods are slightly inflated.

Figure 2 shows the comparison of BIVAS with varbvs, cMCP, and GEL in terms of bi-level variable selection, estimation accuracy, and computational efficiency when  $\rho = 0$  (complete results with other settings of  $\rho$  are provided in Supplementary Figures 1–4). To evaluate the AUC of cMCP and GEL, the relative importance of each variable was approximated using relative effect sizes (i.e., the absolute values of effect sizes divided by maximum of the absolute values). As varbvs only selects individual variables, we are unable to evaluate its group selection performance. Therefore, we treat it as a base line for comparisons of BIVAS with other two alternatives. In the bottom left panel, estimation errors of all the three methods decrease steadily as SNR increases when the sparsity-in-group dominates. BIVAS has similar performance with cMCP and GEL when SNR is moderate ( $\text{SNR} = 0.5$ ) but outperforms them when SNR is relatively large ( $\text{SNR} = 1, 2$ ). When sparsity-in-variable dominates, the estimation performances of BIVAS and cMCP are close to varbvs, but the estimation error of GEL is inflated.

To evaluate the performance of variable selection, we primarily focus on the measure of area under the receiver operating characteristic (ROC) curve (AUC) both at the group and individual levels. The top left panel of Figure 2 shows the AUC of the four methods. Compared to varbvs, BIVAS has much higher AUC when the group sparsity dominates. As the variable level sparsity becomes dominant, BIVAS correctly detects that most of the groups are active (i.e.,  $\eta_k = 1$ ) and attributes the sparsity to variable level. In fact, BIVAS reduces to varbvs when all  $\eta_k = 1$  and only  $\gamma_{jk}$  plays a role in selecting variables. Therefore, we can observe that BIVAS gradually converges to varbvs as more groups become active. The performance of variable selection for



**Figure 2.** Boxplots for comparing BIVAS, varbvs, cMCP and GEL (coupling parameter  $\tau = 1/3$ ). Top left: AUC for individual variable selection. Top right: AUC for group selection. Bottom left: Mean squared error (MSE) of coefficient estimates. Bottom right: Computational time.

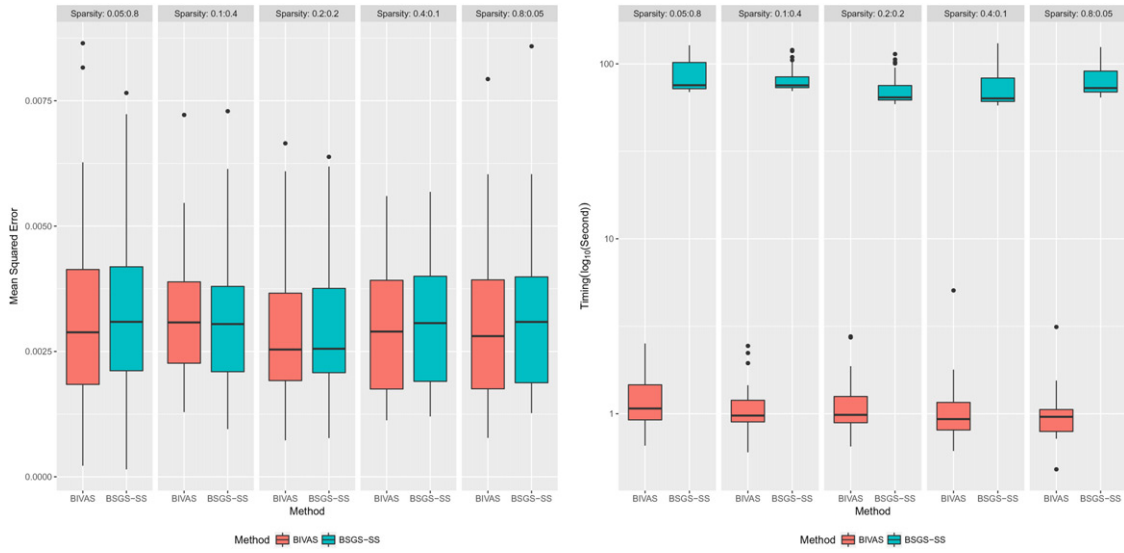
GEL is comparable with BIVAS when SNR is large. When the signal is weak (SNR = 0.5), the AUC of BIVAS becomes much larger than that of GEL. Moreover, BIVAS outperforms GEL when many groups are active. This pattern is consistent with that we observe in the measurement of estimation error. In all settings we considered, the performance of cMCP is poor. The top right panel in Figure 2 shows the performance of variable selection at the group level (group-AUC). The pattern of group-AUC is similar to the individual level AUC. The bottom right panel in Figure 2 illustrates the computational efficiency of the four methods. With multithread computation, the speed of BIVAS is comparable to other methods and faster than cMCP and GEL in most cases.

In addition, we also made comparisons of the estimation accuracy and computational efficiency between BIVAS and Bayesian methods adopting MCMC. Here we focused on BSGS (Chen et al. 2016) and BSGS-SS (Xu and Ghosh 2015) implemented in the R-packages “BSGS” and “MBSGS,” respectively. Since BSGS has similar prior specification with BIVAS, we first treat it as a baseline and compare the posterior means of  $\beta$  obtained by BIVAS and BSGS using a simple example. We chose  $n = 50$ ,  $p = 100$ , and  $K = 10$  with 10 variables in each group. Then, we made group 1, 2, 5, 8 active and set the corresponding

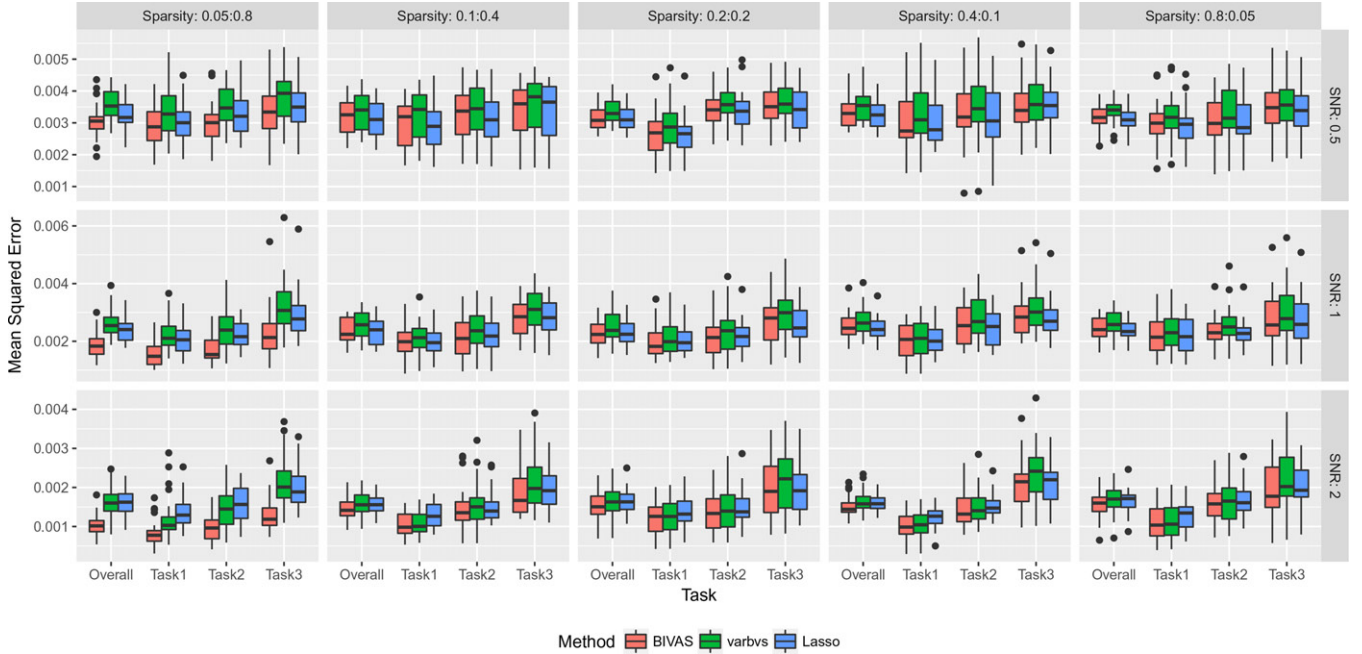
nonzero effects  $\beta_{7,1} = \beta_{8,1} = \beta_{9,1} = 3.2$ ,  $\beta_{1,2} = \beta_{2,2} = 1.5$ ,  $\beta_{3,5} = -1.5$ , and  $\beta_{7,8} = -2$ , where  $\beta_{j,k}$  represents the effect size of the  $j$ th variable in the  $k$ th group. The total number of iterations of the BSGS Gibbs sampler was set at 2000 and the fixed parameters were set as  $\tau_{jk}^2 = 5$ ,  $\rho_k = \theta_{jk} = 0.5$ . We used the same values to initialize BIVAS. It took 192.648 sec for BSGS and only 0.198 sec for BIVAS to fit the model. The resulting posterior mean estimates of the two approaches are given in Supplementary Figure 14. We can observe that BIVAS produces accurate posterior means of  $\beta$ . Next, we compare BIVAS with BSGS-SS with moderate-sized simulation dataset. We set  $n = 200$ ,  $p = 1000$ ,  $K = 100$  with 10 variables in each group and  $\rho = 0.5$ , SNR = 1. The total iteration for Gibbs sampler is set at 500 with 100 burn-in iterations. As illustrated in Figure 3, BIVAS achieves almost the same estimation accuracy as BSGS-SS but uses only around 1% of its computational time.

While the above outcomes were obtained by assuming there is no fixed effect, the performance of BIVAS is not affected by incorporating the covariates  $\mathbf{Z}$  with fixed effects. We demonstrated this result by introducing additional fixed effects  $\omega = [1, 2, 3, 4, 5]^T$  with the details provided in Section 5 of the supplementary materials.





**Figure 3.** Comparison of BIVAS and BSGS-SS. Left: Boxplots of mean squared error of coefficient estimates. Right: Boxplots of time. BIVAS achieves similar estimation accuracy with much less computational time compared to BSGS-SS.



**Figure 4.** Boxplots of mean squared error for comparing BIVAS, varbvs, Ridge, and Lasso in multitask learning. Each panel corresponds to a specific simulation setting. Within panel, both overall MSE and task-specific MSE are evaluated. BIVAS outperforms other methods when group-sparsity dominates.

For multitask BIVAS, we compared with varbvs and Lasso that are applied separately to each task. We simulated  $L = 3$  tasks with sample sizes  $n_1 = 600$ ,  $n_2 = 500$ ,  $n_3 = 400$ . Number of variables  $K = 2000$  was used throughout. We followed the settings in group BIVAS for the sparsity pattern and SNR. The estimation error was evaluated on both overall scale and individual-task scale, as shown in Figure 4.

As one can observe, BIVAS outperforms varbvs and Lasso when the group sparsity is predominant and the difference increases as the signal becomes stronger. Even when the proportion of group sparsity decreases, BIVAS is still comparable with the other two alternatives. In addition, when a strong group-sparsity pattern exists (leftmost column), BIVAS has its biggest

gain on Task 3, which has the smallest sample size. This is because BIVAS takes the advantage of shared sparsity pattern in different tasks.

### 3.2. Real Data Analysis

To examine the performance of BIVAS in large scale data, we provide three real examples: we first apply the regression model to the GWAS data from the Wellcome Trust Case Control Consortium (WTCCC) (Wellcome Trust Case Control Consortium 2007) and the Northern Finland Birth Cohort (NFBC) (Sabatti et al. 2009); then we analyze a movie review dataset from IMDb.com (Maas et al. 2011) using the multitask model.

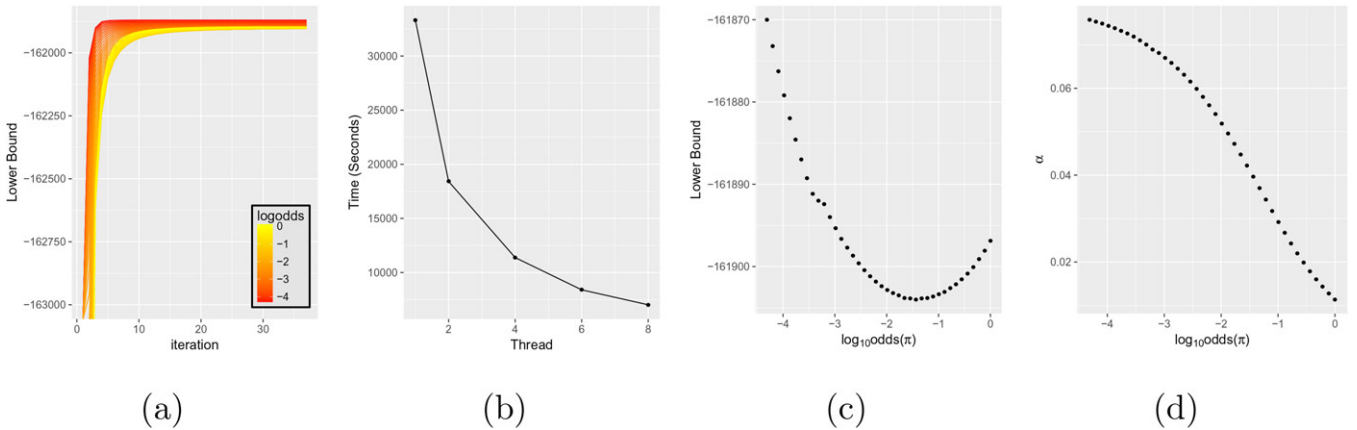
### 3.2.1. GWAS Data

In the GWAS datasets, we conducted quality control based on PLINK (Purcell et al. 2007) and GCTA (Yang et al. 2011): individuals with  $>2\%$  missing genotypes were first removed; we also removed the SNPs with minor allele frequency  $<0.05$ , missingness  $>1\%$ , or  $p$ -value  $<0.001$  in Hardy–Weinberg equilibrium test, excluding individuals with genetic relatedness greater than 0.025.

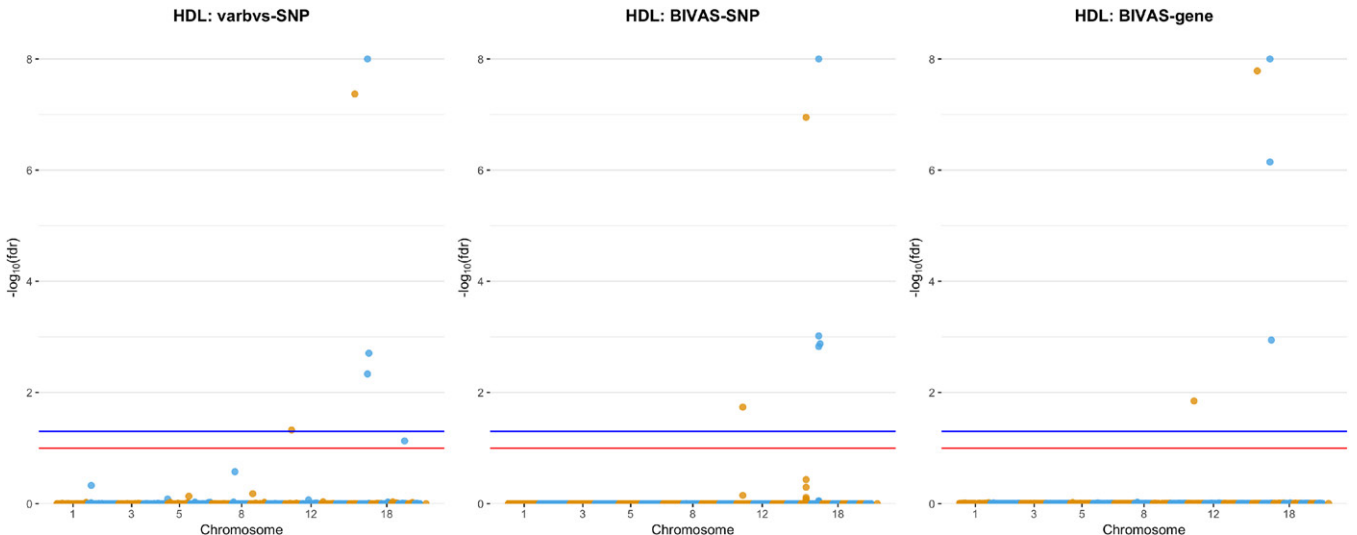
We first considered the high-density lipoprotein (HDL) from the NFBC data, which was obtained from the database of Genotypes and Phenotypes (dbGaP) with accession number phs000276.v1.p1. This data set was composed of 5123 individuals and 319,147 SNPs. In our analysis, the SNPs were first annotated with their corresponding gene region using ANNOVAR (Hakonarson, Li, and Wang 2010), which leads to 318,686 SNPs in 20,493 genes without overlap. Treating the genes as groups, we applied both BIVAS and varbvs to the data. Figure 5(a) shows the convergence of each EM procedure for BIVAS. One can observe that the EM algorithm converges faster for smaller values of  $\pi$ , suggesting the evidence of group sparsity. Computational times for different numbers of threads are presented in Figure 5(b).

When  $h = 40$ , BIVAS took around 3.2 hr to converge using 4 threads and only took 1.6 hr using 8 threads, which indicates that the developed algorithm achieved almost perfect efficiency in parallelization. Estimates of lower bound and parameter  $\alpha$  are shown in Figures 5(c) and (d), suggesting the effectiveness of leveraging group structure using BIVAS. After the convergence, we identified the SNPs and genes based on  $\text{fdr} < 0.05$ . Five risk variants (rs2167079, rs1532085, rs3764261, rs7499892, rs255052) were identified by varbvs. BIVAS discovered one more variant: rs1532624. For the group level selection, BIVAS identified five associated genes, among which CETP contained two risk SNPs: rs7499892 was also identified by varbvs but rs1532624 was a new one. The above results are visualized in the Manhattan plots (Figure 6).

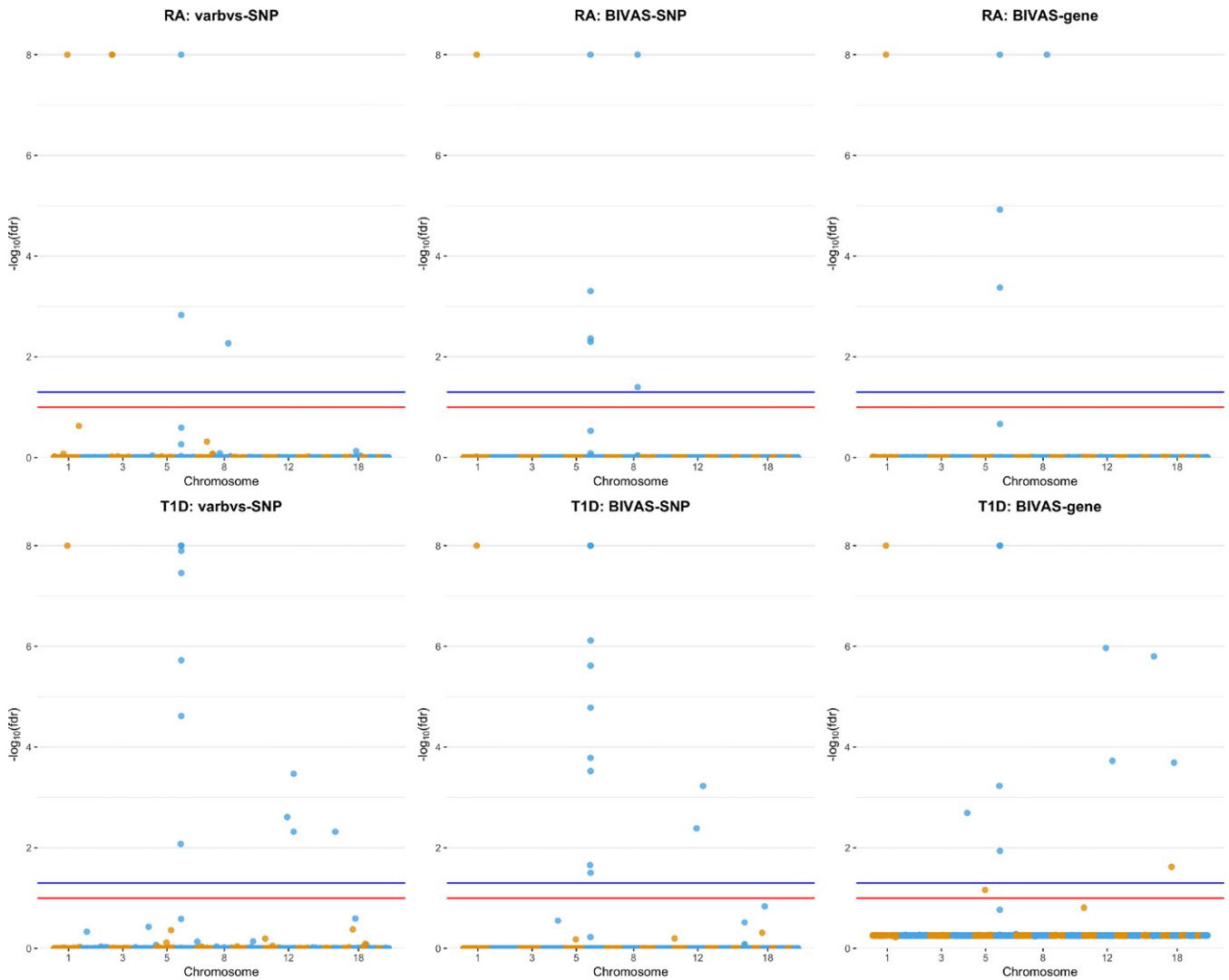
In the second example, we analyzed rheumatoid arthritis (RA) and type 1 diabetes (T1D) in the WTCCC data. These datasets were from WTCCC websites [https://www.wtccc.org.uk/info/access\\_to\\_data\\_samples.html](https://www.wtccc.org.uk/info/access_to_data_samples.html). After quality control, we had 4494 individuals and 307,089 SNPs for RA, and 4986 individuals and 307,357 SNPs for T1D. The SNPs were then matched with corresponding genes using HapMap3 as reference, leading to



**Figure 5.** BIVAS in fitting HDL. (a) Convergence of lower bound for  $h = 40$  EM procedure. (b) Computational times using 1, 2, 4, 6, 8 threads. (c) Lower bound for the 40 settings procedure after convergence. (d)  $\alpha$  for the 40 settings after convergence.



**Figure 6.** Manhattan plots of high-density lipoprotein (HDL). The  $-\log_{10}(\text{fdr})$  of each SNP/gene is plotted according to its position in the genome. Red line represents  $\text{fdr} = 0.1$  and blue line represents  $\text{fdr} = 0.05$ . BIVAS identified one more variant than varbvs.



**Figure 7.** Manhattan plots of rheumatoid arthritis (RA) and Type 1 Diabetes (T1D). The  $-\log_{10}(\text{fdr})$  of each SNP/gene is plotted according to its position in the genome. Red line represents  $\text{fdr} = 0.1$  and blue line represents  $\text{fdr} = 0.05$ . BIVAS identified 3 genes in T1D that contains no association in the SNP level.

242,597 SNPs with 16,789 genes for RA and 242,824 SNPs with 16,815 genes for T1D. Manhattan plots are shown in Figure 7. At the SNP level, the identification results of BIVAS and varbvs are similar but BIVAS further interrogated signals at the gene level making the results more interpretable. For example, in T1D, genes *ADA1*, *LINC00469*, and *LOC100996324* were identified as associated by BIVAS, but these genes contain no single associated SNP either identified by varbvs or BIVAS. This suggests that the associations are weak at the SNP level, but they aggregatively improve power as a group and hence identified by BIVAS at the gene level.

We also applied penalized methods cMCP and GEL to the two example datasets. The numbers of identified genes and variables are summarized in Supplementary Table 1. The corresponding Manhattan plots are given in Supplementary Figures 15–17, where we used the relative effect size as a proxy of  $\text{fdr}$ . As we can observe, a large number of SNPs and genes were identified by the penalized methods. However, it is hard to claim significance of the findings because they do not provide FDR control. Consequently, there are difficulties in interpreting the results obtained by these penalized

methods. The Bayesian approaches based on MCMC, such as BSGS and BSGS-SS, are excluded in the real data analysis because they are computationally infeasible for these large datasets.

### 3.2.2. IMDB Movie Data

In the third example, we analyzed IMDB dataset (Maas et al. 2011) based on multitask BIVAS. The IMDB data set was publicly available at [IMDb.com](http://IMDb.com). The original data were extracted from movie reviews from IMDb.com. It contained 50K movie reviews that were equally split into a training set and a test set. Each review was marked with a rating ranging from 0 to 10, only the polarized reviews were retained (rating  $> 7$  or rating  $< 4$ ). The dataset was comprised of equal number of positive reviews and negative reviews. A bag of representative words was concluded from the whole review. Based on the bag of words, we adopted binary representation to indicate presence of the words. This led to  $K = 27,743$  features (words) with the rating being the response variable. We used six genres of movies as our tasks: drama, comedy, horror, action, thriller, and romance.

Only the reviews of movies that had exactly one genre were used. This led to the sample sizes 3354 for drama, 2235 for comedy, 1175 for horror, 346 for action, 258 for thriller, and 139 for romance. We compared BIVAS against Ridge, Lasso, and varbvs.

Table 1 shows the testing errors of the four methods. For the categories of horror, action, thriller, and romance, BIVAS has better performance than the other three methods. Note that these genres have smaller sample sizes compared to comedy and action. This result is consistent with what we obtain in the simulation study.

The words selected by BIVAS and varbvs are presented in Figures 8 and 9 using “wordcloud” package in R (Fellows 2018). The words in blue and yellow represent the negative and positive effects, respectively. The size of words represents the corresponding effect size. As shown in Figure 8, small number of words were identified by varbvs to be associated with Action, Horror or Thriller, which are genres with smallest sample sizes. However, as shown in Figure 9, BIVAS greatly enriches the effective words in these tasks by borrowing information from the large samples (drama, comedy, and horror). Many associated words that were overwhelmed by noise are now revealed. This can be viewed as a consequence of bi-level selection which selects the important variables and, at the same time, allows sparsity pattern to be shared within group (or through tasks in multitask learning). Hence, many useful words shared through tasks, such as “worst,” “awful,” and “amazing,” can be revealed for small sample and some

particular predictors, like “scariest” in thriller and horror, are maintained task-specific. On the other hand, varbvs (as well as Ridge and Lasso) does not account for the bi-level sparsity structure, so it is unable to capture the shared information through tasks.

## 4. Discussion

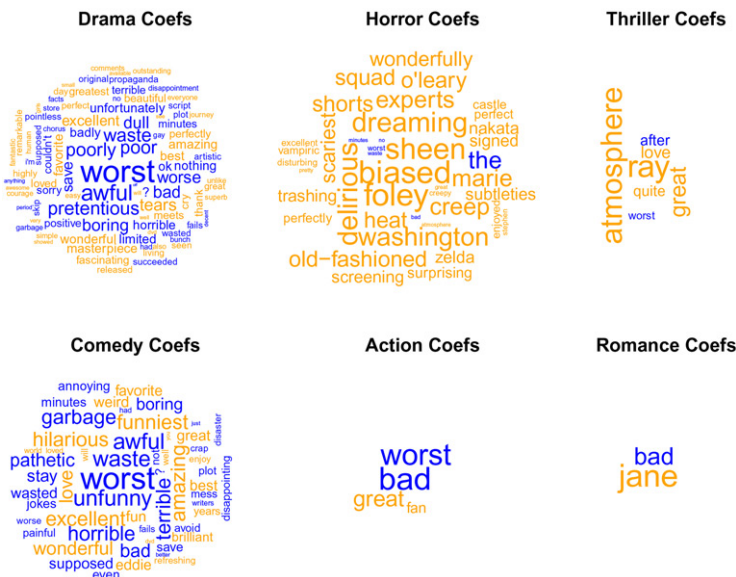
The bi-level variable selection aims at capturing the sparsity at both the individual variable level and the group level to better interrogate the structural information that can assist parameter estimation and variable selection. Bayesian bi-level selection methods are free of parameter tuning and able to obtain the posterior distributions of random effects. Based on the posterior distributions, variables can be selected at both levels by controlling  $\text{fdr}$ . Despite the convenience, existing Bayesian bi-level variable selection methods are often computationally inefficient and unscalable to large datasets due to the intractable posterior.

In this article, we propose a hierarchically factorizable formulation to approximate the posterior distribution, by using the structure of bi-level variable selection. Under the variational assumption, a computationally efficient algorithm is developed based on the variational EM algorithm and importance sampling. The convergence of algorithm is promised and the accurate approximation for the posterior mean can be obtained. The proposed algorithm is efficient, stable, and scalable. Our software is fast and capable of parallel computing. After convergence, variable selection at both levels can be conducted by controlling the  $\text{fdr}$ , prediction can be made based on posterior means. Through the simulation study we showed that our method is no worse than alternative methods given the same computational cost and outperforms some methods in many cases. We also applied BIVAS to real world data and verified its scalability and capability of bi-level selection.

**Table 1.** IMDb testing error.

	Overall	Drama	Comedy	Horror	Action	Thriller	Romance
Ridge	9.58	9.01	10.55	9.01	10.99	10.14	6.76
Lasso	6.67	<b>6.13</b>	<b>6.67</b>	7.20	8.65	9.27	6.77
varbvs	7.14	6.20	6.90	8.94	8.37	11.48	6.91
BIVAS	<b>6.66</b>	6.32	7.01	<b>6.76</b>	<b>6.89</b>	<b>7.44</b>	<b>5.39</b>

NOTE: The bold value represents the smallest testing error within a genre.



**Figure 8.** IMDb wordcloud generated by varbvs. Each panel represents a movie genre. The words in blue and yellow represent the negative and positive effects, respectively. The size of words represents the scale of effect size.





**Figure 9.** IMDb wordcloud generated by BIVAS. Both shared effects and genre-specific effects are shown in the figure. BIVAS identified more informative words in small sample tasks.

## Supplementary Materials

**Supp\_final.pdf** The supplementary materials include the detailed derivation for both regression and multitask learning. (PDF)

**R-package for BIVAS:** R-package “bivas.” The package contains the functions used in fitting BIVAS and making statistical inference. (zipped tar file)

**sim-bivas** The file contains the R scripts for generating numerical results in Section 3. (zipped file for R scripts)

## Funding

This work was supported in part by the National Science Funding of China [61501389]; the Hong Kong Research Grant Council [22302815, 12316116, 12301417, and 16307818]; The Hong Kong University of Science and Technology [startup grant R9405 and IGN175C02]; Duke-NUS Medical School WBS [R-913-200-098-263]; Ministry of Education, Singapore. AcRF Tier 2 [MOE2016-T2-2-029, MOE2018-T2-1-046, and MOE2018-T2-2-006].

## References

- Bishop, C. M. (2006), *Pattern Recognition and Machine Learning*, Information Science and Statistics, New York: Springer. [41,42]
- Breheny, P. (2015), “The Group Exponential Lasso for Bi-Level Variable Selection,” *Biometrics*, 71, 731–740. [41,45]
- Breheny, P., and Huang, J. (2009), “Penalized Methods for Bi-Level Variable Selection,” *Statistics and Its Interface*, 2, 369–380. [40,41,45]
- Carbonetto, P., and Stephens, M. (2012), “Scalable Variational Inference for Bayesian Variable Selection in Regression, and Its Accuracy in Genetic Association Studies,” *Bayesian Analysis*, 7, 73–108. [41,43,45]

- Caruana, R. (1997), “Multitask Learning,” *Machine Learning*, 28, 41–75. [43]
- Chen, R.-B., Chu, C.-H., Yuan, S., and Wu, Y. N. (2016), “Bayesian Sparse Group Selection,” *Journal of Computational and Graphical Statistics*, 25, 665–683. [41,46]
- Efron, B., and Hastie, T. (2016). *Computer Age Statistical Inference*, Vol. 5, Cambridge: Cambridge University Press. [44]
- (2012), *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction* (Vol. 1), Cambridge: Cambridge University Press. [41]
- Fan, J., and Li, R. (2001), “Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties,” *Journal of the American statistical Association*, 96, 1348–1360. [40]
- Fellows, I. (2018), “wordcloud: Word Clouds,” R Package Version 2.6. [44,50]
- Figueiredo, M. A. (2003), “Adaptive Sparseness for Supervised Learning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25, 1150–1159. [40]
- George, E. I., and McCulloch, R. E. (1993), “Variable Selection via Gibbs Sampling,” *Journal of the American Statistical Association*, 88, 881–889. [40]
- (1997), “Approaches for Bayesian Variable Selection,” *Statistica Sinica*, 7, 339–373. [40]
- Hakonarson, H., Li, M., and Wang, K. (2010, 07), “ANNOVAR: Functional Annotation of Genetic Variants From High-Throughput Sequencing Data,” *Nucleic Acids Research*, 38, e164–e164. [48]
- Hastie, T., Tibshirani, R., and Wainwright, M. (2015), *Statistical Learning With Sparsity: The Lasso and Generalizations*, Boca Raton, FL: CRC Press. [40]
- Huang, J., Ma, S., Xie, H., and Zhang, C.-H. (2009), “A Group Bridge Approach for Variable Selection,” *Biometrika*, 96, 339–355. [40]

- Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010), "Penalized Regression, Standard Errors, and Bayesian Lassos," *Bayesian Analysis*, 5, 369–411. [40]
- Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011), "Learning Word Vectors for Sentiment Analysis," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1), Association for Computational Linguistics, pp. 142–150. [47,49]
- MacKay, D. J. (2003), *Information Theory, Inference and Learning Algorithms*, Cambridge: Cambridge University Press. [41]
- Madigan, D., and Raftery, A. E. (1994), "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window," *Journal of the American Statistical Association*, 89, 1535–1546. [40]
- Mallick, B. K., and Bae, K. (2004), "Gene Selection Using a Two-Level Hierarchical Bayesian Model," *Bioinformatics*, 20, 3423–3430. [40]
- Mitchell, T. J., and Beauchamp, J. J. (1988), "Bayesian Variable Selection in Linear Regression," *Journal of the American Statistical Association*, 83, 1023–1032. [40]
- Ormerod, J. T., You, C., and Müller, S. (2017), "A Variational Bayes Approach to Variable Selection," *Electronic Journal of Statistics*, 11, 3549–3594. [40]
- Park, T., and Casella, G. (2008), "The Bayesian Lasso," *Journal of the American Statistical Association*, 103, 681–686. [40]
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., and Sham, P. C. (2007), "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses," *The American Journal of Human Genetics*, 81, 559–575. [48]
- R Core Team (2018), *R: A Language and Environment for Statistical Computing*, available at <https://www.R-project.org/>. [44]
- Raman, S., Fuchs, T. J., Wild, P. J., Dahl, E., and Roth, V. (2009), "The Bayesian Group-Lasso for Analyzing Contingency Tables," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ACM, pp. 881–888. [40]
- Ročková, V. (2018), "Particle EM for Variable Selection," *Journal of the American Statistical Association*, 113, 1684–1697. [40]
- Ročková, V., and George, E. I. (2014), "EMVS: The EM Approach to Bayesian Variable Selection," *Journal of the American Statistical Association*, 109, 828–846. [40]
- Sabatti, C., Hartikainen, A.-L., Pouta, A., Ripatti, S., Brodsky, J., Jones, C. G., Zaitlen, N. A., Varilo, T., Kaakinen, M., and Sovio, U. (2009), "Genome-Wide Association Analysis of Metabolic Traits in a Birth Cohort From a Founder Population," *Nature Genetics*, 41, 35–46. [47]
- Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Series B*, 58, 267–288. [40]
- Turner, S. D. (2018), "qqman: An R Package for Visualizing GWAS Results Using Q-Q and Manhattan Plots," *Journal of Open Source Software*, 3, 731. [44]
- Wellcome Trust Case Control Consortium (2007), "Genome-Wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls," *Nature*, 447, 661–678. [47]
- Wickham, H. (2016), *ggplot2: Elegant Graphics for Data Analysis*, New York: Springer-Verlag. [44]
- Xu, X., and Ghosh, M. (2015), "Bayesian Variable Selection and Estimation for Group Lasso," *Bayesian Analysis*, 10, 909–936. [40,41,46]
- Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011), "GCTA: A Tool for Genome-Wide Complex Trait Analysis," *The American Journal of Human Genetics*, 88, 76–82. [48]
- Yuan, M., and Lin, Y. (2005), "Efficient Empirical Bayes Variable Selection and Estimation in Linear Models," *Journal of the American Statistical Association*, 100, 1215–1225. [40]
- (2006), "Model Selection and Estimation in Regression With Grouped Variables," *Journal of the Royal Statistical Society, Series B*, 68, 49–67. [40]
- Zhang, C.-H. (2010), "Nearly Unbiased Variable Selection Under Minimax Concave Penalty," *The Annals of Statistics*, 38, 894–942. [40]
- Zhang, C.-X., Xu, S., and Zhang, J.-S. (2019), "A Novel Variational Bayesian Method for Variable Selection in Logistic Regression Models," *Computational Statistics & Data Analysis*, 133, 1–19. [40]
- Zhao, P., Rocha, G., and Yu, B. (2009), "The Composite Absolute Penalties Family for Grouped and Hierarchical Variable Selection," *Annals of Statistics*, 37, 3468–3497. [40]
- Zhou, N., and Zhu, J. (2010), "Group Variable Selection via a Hierarchical Lasso and Its Oracle Property," *Statistics and Its Interface*, 3, 557–574. [40]