MASTER'S THESIS

BIVAS: a scalable Bayesian method for bi-level variable selection

Cai, Mingxuan

Date of Award:
2018

HONG KONG BAPTIST UNIVERSITY

Master of Philosophy

THESIS ACCEPTANCE

DATE: May 17, 2018

STUDENT'S NAME: CAI Mingxuan

THESIS TITLE: BIVAS: A Scalable Bayesian Method for Bi-level Variable Selection

This is to certify that the above student's thesis has been examined by the following panel members and has received full approval for acceptance in partial fulfillment of the requirements for the degree of Master of Philosophy.

Chairman:	Dr. Tam Hon Wah Associate Professor, Department of Computer Science, HKBU (Designated by Dean of Faculty of Science)
Internal Members:	Dr. Liu Hongyu Associate Professor, Department of Mathematics, HKBU (Designated by Head of Department of Mathematics)
	Dr. Peng Heng Associate Professor, Department of Mathematics, HKBU
External Members:	Dr. Yao Yuan Associate Professor Department of Mathematics The Hong Kong University of Science and Technology

Issued by Graduate School, HKBU

BIVAS: A Scalable Bayesian Method for Bi-level Variable Selection

CAI Mingxuan

A thesis submitted in partial fulfillment of the requirements
for the degree of
Master of Philosophy

Principal Supervisor:
Dr. PENG Heng (Hong Kong Baptist University)

May 2018

DECLARATION

I hereby declare that this thesis represents my own work which has been done after registration for the degree of MPhil (or PhD as appropriate) at Hong Kong Baptist University, and has not been previously included in a thesis or dissertation submitted to this or any other institution for a degree, diploma or other qualifications.

I have read the University's current research ethics guidelines, and accept responsibility for the conduct of the procedures in accordance with the University's Committee on the Use of Human & Animal Subjects in Teaching and Research (HASC). I have attempted to identify all the risks related to this research that may arise in conducting this research, obtained the relevant ethical and/or safety approval (where applicable), and acknowledged my obligations and the rights of the participants.

Signature: 蔡銘軒

Date: May 2018

Abstract

In this thesis, 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 methods based on Markov Chain Monte Carlo (MCMC) are often computationally intensive and not scalable to large data sets. To address this problem, we consider variational inference for bi-level variable selection (BIVAS). In contrast to the commonly used mean-field approximation, we propose a hierarchical factorization to approximate the posterior distribution, by utilizing 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 data sets from multi-task 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 BIVAS software with support of parallelization is implemented in R package 'bivas' available at https://github.com/mxcai/bivas.

Keywords: Bayesian variable selection; Variational inference; Group sparsity; Parallel computing.

Acknowledgements

First and foremost, I would like to express my deep gratitude to my supervisor Dr. PENG Heng and Dr YANG Can for their inspiring guidance and enthusiastic support. Their superb intuition, broad knowledge, and continuous encouragement have been indispensable throughout my MPhil study and are extraordinarily beneficial in my future research career. I have learned various things from them in the fields of statistics during these years. It is my great honor to be their students.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Variable selection in regression 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 et al., 2015]. In many real applications, this sparse pattern could be more complicated. In this thesis, 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, i.e., variable selection at both the individual and group levels [Breheny and Huang, 2009].

There have been rich literatures on variable selection [Fan and Li, 2001; Tibshirani, 1996; Yuan and Lin, 2006; Zhang, 2010], 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 priors [Bae and Mallick, 2004; Figueiredo, 2003; Park and Casella, 2008; Yuan and Lin, 2005] and spike-slab priors [George and McCulloch, 1993, 1997; Madigan and Raftery, 1994;

Mitchell and Beauchamp, 1988]. To perform variable selection at the group level, the group Lasso [Yuan and Lin, 2006] introduced the L_1-L_2 norm penalty to grouped variables and perform group selection using the L_1 norm. CAP [Zhao et al., 2009] generalized this idea to be the L_1-L_γ 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 et al., 2015].

1.2 Bi-level variable selection

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, i.e. if a group is selected, all variables within that group will be non-zero. 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 methods 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, 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 struture 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) [Chen et al., 2016; Xu et al., 2015]. The computational costs of these methods become large in the presence of a large number

of variables.

In this thesis, 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-maximiztion (EM) algorithm is developed to handle large data sets. Moreover, we extend our approach to handle a class of multi-task 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.

1.3 Outline of the thesis

The remainder of this thesis is organized as follows. In Chapter 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 multi-task learning.

In Chapert 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 thesis with a short discussion in Section 4.

Chapter 2

Statistical Models and Algorithms

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 $\mathbf{X}$:

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\omega} + \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{2.1}$$

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 non-zero entries of $\boldsymbol{\beta}$ at both the group and individual-variable levels. For this reason, we introduce two binary variables: η_k indicates whether group k is active ($\eta_k = 1$) or not ($\eta_k = 0$); and γ_{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 $\boldsymbol{\beta}$:

$$\beta_{jk}|\eta_k, \gamma_{jk}; \sigma_\beta^2 \sim \begin{cases} \mathcal{N}(\beta_{jk}|0, \sigma_\beta^2) & \text{if } \eta_k = 1,\ \gamma_{jk} = 1, \\ \delta_0(\beta_{jk}) & \text{otherwise,} \end{cases} \tag{2.2}$$

where $\mathcal{N}(\beta_{jk}|0, \sigma_\beta^2)$ denotes the Gaussian distribution with mean 0 and variance σ_β^2 and $\delta_0(\beta_{jk})$ denotes a Dirac function at zero. This bi-level structure means that β_{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 re-parameterize the model as following:

$$\beta_{jk}|\sigma_\beta^2 \sim \mathcal{N}(0, \sigma_\beta^2), \quad \gamma_{jk}|\alpha \sim \alpha^{\gamma_{jk}}(1-\alpha)^{1-\gamma_{jk}}, \quad \eta_k|\pi \sim \pi^{\eta_k}(1-\pi)^{1-\eta_k}. \tag{2.3}$$

Consequently, the prior of β_{jk} does not depend on γ_{jk} and η_k any more, and the product $\eta_k\gamma_{jk}\beta_{jk}$ form a new random variable exactly distributed as given in (2). We shall use the re-parameterized version through the thesis.

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

$$\begin{aligned} &\Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}|\mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}) = \Pr(\mathbf{y}|\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\theta})\Pr(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}|\boldsymbol{\theta}) \\ =&\mathcal{N}(\mathbf{y}|\mathbf{Z}\boldsymbol{\omega} + \sum_k^K \sum_j^{l_k} \eta_k\gamma_{jk}\beta_{jk}\mathbf{x}_{jk}, \sigma_e^2) \prod_{k=1}^K \pi^{\eta_k}(1-\pi)^{1-\eta_k} \prod_{j=1}^{l_k} \mathcal{N}(0, \sigma_\beta^2)\alpha^{\gamma_{jk}}(1-\alpha)^{1-\gamma_{jk}}, \end{aligned} \tag{2.4}$$

where $\mathbf{x}_{jk}$ is a column of $\mathbf{X}$ corresponding to the j-th variable in the k-th group. The goal is to obtain the estimate of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$, by optimizing the marginal likelihood

$$\log\ \Pr(\mathbf{y}|\mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}) = \log \sum_{\boldsymbol{\gamma}} \sum_{\boldsymbol{\eta}} \int_{\boldsymbol{\beta}} \Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}|\mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}) d\boldsymbol{\beta}, \tag{2.5}$$

and evaluate the posterior

$$\Pr(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \mathbf{Z}; \hat{\boldsymbol{\theta}}) = \frac{\Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}|\mathbf{X}, \mathbf{Z}; \hat{\boldsymbol{\theta}})}{\Pr(\mathbf{y}|\mathbf{X}, \mathbf{Z}; \hat{\boldsymbol{\theta}})}. \tag{2.6}$$

2.1.2 Algorithm

Conventionally, the model involving latent variables is often solved by the Expectation-Maximization (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 $\boldsymbol{\gamma}$ and $\boldsymbol{\eta}$. Alternatively, we propose a variational EM algorithm via approximate Bayesian inference [Bishop, 2006].

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

$$\begin{aligned}
\log\ p(\mathbf{y}|\mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}) &= \log \sum_{\boldsymbol{\gamma}} \sum_{\boldsymbol{\eta}} \int_{\boldsymbol{\beta}} \Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}|\mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}) d\boldsymbol{\beta} \\
&\geq \sum_{\boldsymbol{\gamma}} \sum_{\boldsymbol{\eta}} \int_{\boldsymbol{\beta}} q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}) \log \frac{\Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}|\mathbf{X}, \mathbf{Z}; \boldsymbol{\theta})}{q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta})} d\boldsymbol{\beta} \\
&= \mathbb{E}_q[\log \Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}|\mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}) - \log q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta})] \\
&\equiv \mathcal{L}(q),
\end{aligned} \tag{2.7}$$

where the equality holds if and only if $q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}) = \Pr(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta})$. 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 distribition to accommodate the bi-level variable selection. Specifically, we consider the the following hierarchically structured distribution as an approximation to posterior $\Pr(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \mathbf{Z})$:

$$q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}) = \prod_{k}^{K} \left(q(\eta_k) \prod_{j}^{l_k} (q(\beta_{jk}|\eta_k, \gamma_{jk}) q(\gamma_{jk})) \right), \tag{2.8}$$

where we have assumed that groups are independent; and given a group, the factors

inside are also independent. With this assumption, we can rewrite the ELBO as:

$$\mathcal{L}(q) = \mathbb{E}_{q(\eta)}\left[\mathbb{E}_{q(\gamma,\beta|\eta)}\left[\log \Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}) - \log q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta})\right]\right]. \tag{2.9}$$

Without any other assumptions, we can show (with details in Appendix) that the optimal solution of q is obtained by:

$$\log q^*(\beta_{jk}|\eta_k = 1, \gamma_{jk} = 1) = \mathbb{E}_{j'\neq j|k}\left[\mathbb{E}_{k'\neq k} \log \Pr(\mathbf{y}, \boldsymbol{\eta}_{-k}, \eta_k = 1, \boldsymbol{\gamma}_{-jk}, \gamma_{jk} = 1, \boldsymbol{\beta})\right], \tag{2.10}$$

hence the joint variational posterior is

$$q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}) = \prod_k^K \left(\pi_k^{\eta_k}(1-\pi_k)^{1-\eta_k} \prod_j^{l_k} \left(\alpha_{jk}^{\gamma_{jk}}(1-\alpha_{jk})^{1-\gamma_{jk}} \mathcal{N}(\mu_{jk}, s_{jk}^2)^{\eta_k\gamma_{jk}} \mathcal{N}(0, \sigma_\beta^2)^{1-\eta_k\gamma_{jk}} \right) \right), \tag{2.11}$$

where

$$\begin{aligned} s_{jk}^2 &= \frac{\sigma_e^2}{\mathbf{x}_{jk}^T\mathbf{x}_{jk} + \frac{\sigma_e^2}{\sigma_\beta^2}}, \\ \mu_{jk} &= \frac{\mathbf{x}_{jk}^T(\mathbf{y} - \mathbf{Z}\boldsymbol{\omega}) - \sum_{k'\neq k}^K \sum_j^{l_k} \mathbb{E}_{jk'}[\eta_{k'}\gamma_{jk'}\beta_{jk'}]\mathbf{x}_{jk'}^T\mathbf{x}_{jk} - \sum_{j'\neq j}^{l_k} \mathbb{E}[\gamma_{j'k}\beta_{j'k}]\mathbf{x}_{j'k}^T\mathbf{x}_{jk}}{\mathbf{x}_{jk}^T\mathbf{x}_{jk} + \frac{\sigma_e^2}{\sigma_\beta^2}}, \end{aligned} \tag{2.12}$$

$$\pi_k = \frac{1}{1+\exp(-u_k)}, \text{ with } u_k = \log\frac{\pi}{1-\pi} + \frac{1}{2}\sum_j^{l_k} \alpha_{jk}\left(\log\frac{s_{jk}^2}{\sigma_\beta^2} + \frac{\mu_{jk}^2}{s_{jk}^2}\right), \tag{2.13}$$

$$\alpha_{jk} = \frac{1}{1+\exp(-v_{jk})}, \text{ with } v_{jk} = \log\frac{\alpha}{1-\alpha} + \frac{1}{2}\pi_k\left(\log\frac{s_{jk}^2}{\sigma_\beta^2} + \frac{\mu_{jk}^2}{s_{jk}^2}\right). \tag{2.14}$$

By inspections of Equations (2.8) and (2.11), we have $q(\eta_k = 1) = \pi_k$ and $q(\gamma_{jk} = 1) = \alpha_{jk}$, which can be viewed as approximations to the posterior distibutions $\Pr(\eta_k = 1|\mathbf{y}, \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta})$ and $\Pr(\gamma_{jk} = 1|\mathbf{y}, \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta})$, repectively. Similarly, $q(\beta_{jk}|\eta_k\gamma_{jk} = 1) = \mathcal{N}(\mu_{jk}, s_{jk}^2)$ can be interpreted as the variational approximation to $\Pr(\beta_{jk}|\eta_k\gamma_{jk} = 1, \mathbf{y}, \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta})$, which is the conditional posterior distribution of β_{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 $\Pr(\beta_{jk}|\eta_k\gamma_{jk} = 0, \mathbf{y}, \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta})$, corresponding to the case when β_{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 π_k and variable-level posterior inclusion probability α_{jk} can be viewed as their prior inclusion probability (π, α) updated by data-driven information. Furthermore, π_k and α_{jk} are interdependent. On one hand, if more and more α_{jk} within the k-th group become closer to one, then π_k will be closer to one, as seen in Equation (2.13). On the other hand, if π_k increases, then the variables in the k-th group are more likely to be selected, see Equation (2.14).

With Equation (2.11), the lower bound $\mathcal{L}(q)$ can be evaluated analytically. By setting the derivative of $\mathcal{L}(q)$ with respect to $\boldsymbol{\theta}$ to be zero, we have the updating equations for parameter estimation:

$$
\begin{aligned}
\sigma_e^2 =& \frac{||\mathbf{y} - \mathbf{Z}\boldsymbol{\omega} - \sum_k^K \sum_j^{l_k} \pi_k \alpha_{jk} \mu_{jk} \mathbf{x}_{jk}||^2}{n} \\
&+ \frac{\sum_k^K \sum_j^{l_k} [\pi_k \alpha_{jk}(s_{jk}^2 + \mu_{jk}^2) - (\pi_k \alpha_{jk} \mu_{jk})^2] \mathbf{x}_{jk}^T \mathbf{x}_{jk}}{n} \\
&+ \frac{\sum_k^K (\pi_k - \pi_k^2) [\sum_j^{l_k} \sum_{j'}^{l_k} \alpha_{j'k} \mu_{j'k} \alpha_{jk} \mu_{jk}] \mathbf{x}_{j'k}^T \mathbf{x}_{jk}}{n}, \\
\sigma_\beta^2 =& \frac{\sum_k^K \sum_j^{l_k} \pi_k \alpha_{jk}(s_{jk}^2 + \mu_{jk}^2)}{\sum_k^K \sum_j^{l_k} \pi_k \alpha_{jk}}, \\
\alpha =& \frac{1}{p} \sum_k^K \sum_j^{l_k} \alpha_{jk}, \\
\pi =& \frac{1}{K} \sum_k^K \pi_k, \\
\boldsymbol{\omega} =& (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T (\mathbf{y} - \sum_k^K \sum_j^{l_k} \pi_k \alpha_{jk} \mu_{jk} \mathbf{x}_{jk}).
\end{aligned}
\tag{2.15}
$$

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 $\boldsymbol{\theta}$. As a result, the lower bound increases at each iteration and the convergence is guaranteed.

2.2 Multi-task learning with BIVAS

2.2.1 Model setting

In this section, we consider bi-level variable selection in multi-task 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 multi-task learning [Caruana, 1998]. As we shall see later, a class of multi-task 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 'multi-task 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 the reponse 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:

$$\mathbf{y}_j = \mathbf{Z}_j \boldsymbol{\omega}_j + \mathbf{X}_j \boldsymbol{\beta}_j + \mathbf{e}_j, \ \ j = 1, \ldots, L, \tag{2.16}$$

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 $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_L] \in \mathbb{R}^{K \times L}$ and β_{jk} be k-th entry in $\boldsymbol{\beta}_j$. Clearly, it is not reasonable to assume that all shared predictors are relevant to all reponses, especially when K is large. A more reasonable assumpotion 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 a specific task. For this purpose, we introduce two binary variables: η_k indicates whether the k-th row of $\boldsymbol{\beta}$ is active or not and γ_{jk} indicates whether $\boldsymbol{\beta}_{jk}$

is zero or not. Then the bi-level spike-slab prior on $\boldsymbol{\beta}$ is introduced by:

$$\beta_{jk}|\eta_k, \gamma_{jk}; \sigma^2_{\beta_j} \sim \begin{cases} \mathcal{N}(\beta_{jk}|0, \sigma^2_{\beta_j}) & \text{if } \eta_k = 1,\ \gamma_{jk} = 1, \\ \delta_0(\beta_{jk}) & \text{otherwise,} \end{cases} \tag{2.17}$$

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

Again we re-parameterize the model to remove the Dirac function:

$$\beta_{jk}|\sigma^2_{\beta_j} \sim \mathcal{N}(0, \sigma^2_{\beta_j}), \quad \gamma_{jk}|\alpha \sim \alpha^{\gamma_{jk}}(1-\alpha)^{1-\gamma_{jk}}, \quad \eta_k|\pi \sim \pi^{\eta_k}(1-\pi)^{1-\eta_k}. \tag{2.18}$$

Let $\boldsymbol{\theta} = \{\alpha, \pi, \sigma^2_{\beta_j}, \sigma^2_{e_j}, \boldsymbol{\omega}_j\}_{j=1}^L$ be the collection of parameters under the multi-task model. The joint probabilistic model is

$$\begin{aligned} &\Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}|\mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}) = \Pr(\mathbf{y}|\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}) \Pr(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}|\boldsymbol{\theta}) \\ =&\prod_{j=1}^{L} \mathcal{N}(\mathbf{y}_j|\mathbf{Z}_j\boldsymbol{\omega}_j + \sum_{k}^{K} \eta_k\gamma_{jk}\beta_{jk}\mathbf{x}_{jk}, \sigma^2_{e_j}) \prod_{k=1}^{K} \pi^{\eta_k}(1-\pi)^{1-\eta_k} \prod_{j=1}^{L} \mathcal{N}(0, \sigma^2_{\beta_j})\alpha^{\gamma_{jk}}(1-\alpha)^{1-\gamma_{jk}}, \end{aligned} \tag{2.19}$$

where $\mathbf{x}_{jk}$ is the k-th column of $\mathbf{X}_j$, corresponding to the k-th variabe in the j-th task. Our goal is to maximize the marginal likelihood, which is of the same form as Equation (2.5), and evaluate the posterior distribution of β_{jk}.

2.2.2 Algorithm

The variational EM algorithm of multi-task BIVAS is straightforward following the similar procedure in 2.1.2. We leave the details in the Appendices. In summary, we

have

$$
\begin{aligned}
s_{jk}^2 &= \frac{\sigma_{e_j}^2}{\mathbf{x}_{jk}^T\mathbf{x}_{jk} + \frac{\sigma_{e_j}^2}{\sigma_{\beta_j}^2}} \\
\mu_{jk} &= \frac{\mathbf{x}_{jk}^T(\mathbf{y}_j - \mathbf{Z}_j\boldsymbol{\omega}_j - \tilde{\mathbf{y}}_{jk})}{\mathbf{x}_{jk}^T\mathbf{x}_{jk} + \frac{\sigma_{e_j}^2}{\sigma_{\beta_j}^2}} \\
\pi_k &= \frac{1}{1+\exp(-u_k)}, \text{ where } u_k = \log\frac{\pi}{1-\pi} + \frac{1}{2}\sum_j^L \alpha_{jk}\left(\log\frac{s_{jk}^2}{\sigma_{\beta_j}^2} + \frac{\mu_{jk}^2}{s_{jk}^2}\right) \\
\alpha_{jk} &= \frac{1}{1+\exp(-v_k)}, \text{ where } v_k = \log\frac{\alpha}{1-\alpha} + \frac{1}{2}\pi_k\left(\log\frac{s_{jk}^2}{\sigma_{\beta_j}^2} + \frac{\mu_{jk}^2}{s_{jk}^2}\right)
\end{aligned}
\tag{2.20}
$$

for E-step; and

$$
\begin{aligned}
\sigma_{e_j}^2 =& \frac{||\mathbf{y}_j - \mathbf{Z}_j\boldsymbol{\omega}_j - \sum_k^K \pi_k\alpha_{jk}\mu_{jk}\mathbf{x}_{jk}||^2}{N_j} \\
&+ \frac{\sum_k^K[\pi_k\alpha_{jk}(s_{jk}^2 + \mu_{jk}^2) - (\pi_k\alpha_{jk}\mu_{jk})^2]\mathbf{x}_{jk}^T\mathbf{x}_{jk}}{N_j}, \\
\sigma_{\beta_j}^2 =& \frac{\sum_k^K \pi_k\alpha_{jk}(s_{jk}^2 + \mu_{jk}^2)}{\sum_k^K \pi_k\alpha_{jk}}, \\
\alpha =& \frac{1}{p}\sum_k^K\sum_j^L \alpha_{jk}, \\
\pi =& \frac{1}{K}\sum_k^K \pi_k, \\
\boldsymbol{\omega}_j =& (\mathbf{Z}_j^T\mathbf{Z}_j)^{-1}\mathbf{Z}_j^T(\mathbf{y}_j - \sum_k^K \pi_k\alpha_{jk}\mu_{jk}\mathbf{x}_{jk}),
\end{aligned}
\tag{2.21}
$$

for M-step.

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

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

These evaluations are based on parameter estimates $\hat{\boldsymbol{\theta}}$. However, as there is no guarantee of global optimality for the EM algorithm, the choice of initial value $\boldsymbol{\theta}^0$ is critical. A bad initial value will lead to a poor $\hat{\boldsymbol{\theta}}$. In our model, due to the existence of multiple latent variables ($\boldsymbol{\beta}$, $\boldsymbol{\eta}$, $\boldsymbol{\gamma}$), choosing a good initial value can be challenging. Here we consider the importance sampling suggested by varbvs [Carbonetto et al., 2012]: we further introduce a prior over $\boldsymbol{\theta}$ and integrate over the value of $\boldsymbol{\theta}$ to obtain the final evaluations. In contrast to varbvs, we introduce prior only on the group sparsity parameter π. We first select h values of π ($\{\pi(i)\}_{i=1}^h$) 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 $\boldsymbol{\theta}' = \{\boldsymbol{\theta}'_i\}_{i=1}^h$ with $\boldsymbol{\theta}'_i = \{\alpha, \pi(i), \sigma_\beta^2, \sigma_e^2, \boldsymbol{\omega}\}$; and the posterior inclusion probability can be approximated as follows:

$$\begin{aligned}
\Pr(\eta_k = 1|\mathbf{y}, \mathbf{X}, \mathbf{Z}) &\approx \int q(\eta_k = 1|\boldsymbol{\theta}') \Pr(\boldsymbol{\theta}'|\mathbf{y}, \mathbf{X}, \mathbf{Z}) d\boldsymbol{\theta}' \approx \frac{\sum_{i=1}^h q(\eta_k = 1|\boldsymbol{\theta}'_i) w(\boldsymbol{\theta}'_i)}{\sum_{i=1}^h w(\boldsymbol{\theta}'_i)}, \\
\Pr(\gamma_{jk} = 1|\mathbf{y}, \mathbf{X}, \mathbf{Z}) &\approx \int q(\gamma_{jk} = 1|\boldsymbol{\theta}') \Pr(\boldsymbol{\theta}'|\mathbf{y}, \mathbf{X}, \mathbf{Z}) d\boldsymbol{\theta}' \approx \frac{\sum_{i=1}^h q(\gamma_{jk} = 1|\boldsymbol{\theta}'_i) w(\boldsymbol{\theta}'_i)}{\sum_{i=1}^h w(\boldsymbol{\theta}'_i)},
\end{aligned} \tag{2.22}$$

where $w(\boldsymbol{\theta}'_i)$ 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, $w(\boldsymbol{\theta}'_i)$ can be approximated by exponential of $\mathcal{L}(q)$ given $\boldsymbol{\theta}'_i$ since $\mathcal{L}(q)$ takes similar shape to $\log \Pr(\mathbf{y}|\mathbf{X}, \mathbf{Z}; \boldsymbol{\theta})$ when the marginal likelihood is relatively large [Carbonetto et al.,

2012]. Hence, we can derive the final evaluation of posteriors:

$$\begin{aligned}
\Pr(\eta_k = 1|\mathbf{y}, \mathbf{X}, \mathbf{Z}) &\approx \sum_{i=1}^{h} \pi_k(i) \cdot \tilde{w}(i) \equiv \tilde{\pi}_k, \\
\Pr(\gamma_{jk} = 1|\mathbf{y}, \mathbf{X}, \mathbf{Z}) &\approx \sum_{i=1}^{h} \alpha_{jk}(i) \cdot \tilde{w}(i) \equiv \tilde{\gamma}_{jk}, \\
\mathbb{E}(\beta_{jk}|\eta_k\gamma_{jk} = 1, \mathbf{y}, \mathbf{X}, \mathbf{Z}) &\approx \sum_{i=1}^{h} \mu_{jk}(i) \cdot \tilde{w}(i) \equiv \tilde{\mu}_{jk}, \\
\mathbb{E}(\eta_k\gamma_{jk}\beta_{jk}|\mathbf{y}, \mathbf{X}, \mathbf{Z}) &\approx \tilde{\pi}_k\tilde{\gamma}_{jk}\tilde{\mu}_{jk},
\end{aligned} \tag{2.23}$$

where

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

Here we handle the normalization inside the exponantial so that the calculation is numerically stable. The same weighting evaluation applies to the parameters $\boldsymbol{\theta}'$. The evaluation of importance sampling is the same for multi-task 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, $h = 20 \sim 40$ is often good enough for large scale data sets. 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 jobs 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 allocates a new task to a thread once it has finished an old task. This technique greatly improves the efficiency of parallelization compared to OpenMP.

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 $fdr_k = 1 - \tilde{\pi}_k$ and fdr of j-th variable in k-th group by

$fdr_{jk} = 1 - \tilde{\alpha}_{jk}$. Hence, by setting a reasonable threshold (e.g $fdr < 0.05$), variables and groups with high posterior inclusion probability can be identified as relevant. 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 multi-task learning) with a new data $\{\mathbf{Z}^{new}, \mathbf{X}^{new}\}$. 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 $\hat{y} = \sum_r \tilde{\omega}_r z_r^{new} + \sum_k \sum_j \tilde{\pi}_k \tilde{\alpha}_{jk} \tilde{\mu}_{jk} x_{jk}^{new}$ (in multi-task learning $\hat{y}_j = \sum_r \tilde{\omega}_{jr} z_{jr}^{new} + \sum_k \tilde{\pi}_k \tilde{\alpha}_{jk} \tilde{\mu}_{jk} x_{jk}^{new}$ for j-th task).

2.5 Appendices

2.5.1 Variational EM Algorithm: Regression with BIVAS

E-Step

Let $\boldsymbol{\theta} = \{\alpha, \pi, \sigma_\beta^2, \sigma_e^2, \boldsymbol{\omega}\}$ be the collection of model parameters in the main text. The joint probabilistic model is

$$\begin{aligned} &\Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta} | \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}) = \Pr(\mathbf{y} | \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}) \Pr(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta} | \boldsymbol{\theta}) \\ =&\mathcal{N}(\mathbf{y} | \mathbf{Z}\boldsymbol{\omega} + \sum_k^K \sum_j^{l_k} \eta_k \gamma_{jk} \beta jk \mathbf{x}_{jk}) \prod_{k=1}^K \pi^{\eta_k} (1-\pi)^{1-\eta_k} \prod_{j=1}^{l_k} \mathcal{N}(0, \sigma_\beta^2) \alpha^{\gamma_{jk}} (1-\alpha)^{1-\gamma_{jk}}. \end{aligned} \tag{2.24}$$

The logarithm of the marginal likelihood is

$$\begin{aligned} \log\ p(\mathbf{y} | \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}) &= \log \sum_{\boldsymbol{\gamma}} \sum_{\boldsymbol{\eta}} \int_{\boldsymbol{\beta}} \Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta} | \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}) d\boldsymbol{\beta} \\ &\geq \sum_{\boldsymbol{\gamma}} \sum_{\boldsymbol{\eta}} \int_{\boldsymbol{\beta}} q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}) \log \frac{\Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta} | \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta})}{q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta})} d\boldsymbol{\beta} \\ &= \mathbb{E}_q[\log \Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta} | \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}) - \log q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta})] \\ &\equiv \mathcal{L}(q), \end{aligned} \tag{2.25}$$

where we have adopted Jensen's inequality to obtain the lower bound $\mathcal{L}(q)$. Next step is to iteratively maximize $\mathcal{L}(q)$ instead of working with the marginal likelihood di-

rectly. As in the main text, we use the following hierarchically factorized distribution to approximate the true posterior:

$$q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}) = \prod_{k}^{K} \left(q(\eta_k) \prod_{j}^{l_k} (q(\beta_{jk}|\eta_k, \gamma_{jk}) q(\gamma_{jk})) \right), \tag{2.26}$$

where we have assumed that groups are independent; and given a group, the factors inside are also independent. With this assumption, we first rewrite the ELBO as:

$$\mathcal{L}(q) = \mathbb{E}_{q(\eta)} \left[\mathbb{E}_{q(\gamma,\beta|\eta)} \left[\log \Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}) - \log q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}) \right] \right]. \tag{2.27}$$

Let $q(\boldsymbol{\gamma}_k) = \prod_j^{l_k} q(\gamma_{jk})$, $q(\boldsymbol{\beta}_k|\eta_k, \boldsymbol{\gamma}_k) = \prod_j^{l_k} (q(\beta_{jk}|\eta_k, \gamma_{jk}) q(\gamma_{jk}))$ and $q(\eta_k, \boldsymbol{\gamma}_k, \boldsymbol{\beta}_k) = q(\eta_k) \prod_j^{l_k} q(\beta_{jk}|\eta_k, \gamma_{jk}) q(\gamma_{jk})$, the lower bound can be written in the following form:

$$\begin{aligned}
&\mathcal{L}(q) \\
=&\sum_{\boldsymbol{\eta}} \prod_{k}^{K} q(\eta_k) \sum_{\boldsymbol{\gamma}} \prod_{k}^{K} q(\boldsymbol{\gamma}_k) \int_{\boldsymbol{\beta}} \left(\log \Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}) - \sum_{k}^{K} \log q(\eta_k, \boldsymbol{\gamma}_k, \boldsymbol{\beta}_k) \right) \prod_{k}^{K} q(\boldsymbol{\beta}_k|\eta_k, \boldsymbol{\gamma}_{jk}) d\boldsymbol{\beta} \\
=&\sum_{\eta_k} q(\eta_k) \sum_{\boldsymbol{\gamma}_k} \prod_{j}^{l_k} q(\gamma_{jk}) \int \prod_{j}^{l_k} q(\beta_{jk}|\eta_k, \gamma_{jk}) [\sum_{\eta_{-k}} \prod_{k' \neq k} q(\eta_{k'}) \sum_{\boldsymbol{\gamma}_{-k}} \prod_{k' \neq k} q(\boldsymbol{\gamma}_{k'}) \\
&\int \log \Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}) \prod_{k' \neq k} q(\boldsymbol{\beta}_k|\eta_k, \boldsymbol{\gamma}_k) d\boldsymbol{\beta}_{k'}] d\boldsymbol{\beta}_k \\
&- \sum_{\eta_k} q(\eta_k) \sum_{\boldsymbol{\gamma}_k} \prod_{j}^{l_k} q(\gamma_{jk}) \int \prod_{j}^{l_k} q(\beta_{jk}|\eta_k, \gamma_{jk}) \log q(\eta_k, \boldsymbol{\gamma}_k, \boldsymbol{\beta}_k) d\boldsymbol{\beta}_k + \text{const} \\
=&\mathbb{E}_{q(\eta_k, \boldsymbol{\gamma}_k, \boldsymbol{\beta}_k)} \left[\mathbb{E}_{k' \neq k} \log \Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}) - \log q(\eta_k, \boldsymbol{\gamma}_k, \boldsymbol{\beta}_k) \right] + \text{const} \\
=&\mathbb{E}_{q(\eta_k)} \left[\mathbb{E}_{q(\boldsymbol{\gamma}_k, \boldsymbol{\beta}_k|\eta_k)} \left[\mathbb{E}_{k' \neq k} \log \Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}) - \log q(\eta_k, \boldsymbol{\gamma}_k, \boldsymbol{\beta}_k) \right] \right] + \text{const} \\
=&q(\eta_k = 1) \left[\mathbb{E}_{q(\boldsymbol{\gamma}_k, \boldsymbol{\beta}_k|\eta_k = 1)} \left[\mathbb{E}_{k' \neq k} \log \Pr(\mathbf{y}, \boldsymbol{\eta}_{-k}, \eta_k = 1, \boldsymbol{\gamma}, \boldsymbol{\beta}) - \log q(\eta_k = 1, \boldsymbol{\gamma}_k, \boldsymbol{\beta}_k) \right] \right] \\
&+ q(\eta_k = 0) \left[\mathbb{E}_{q(\boldsymbol{\gamma}_k, \boldsymbol{\beta}_k|\eta_k = 0)} \left[\mathbb{E}_{k' \neq k} \log \Pr(\mathbf{y}, \boldsymbol{\eta}_{-k}, \eta_k = 0, \boldsymbol{\gamma}, \boldsymbol{\beta}) - \log q(\eta_k = 0, \boldsymbol{\gamma}_k, \boldsymbol{\beta}_k) \right] \right] \\
&+ \text{const},
\end{aligned} \tag{2.28}$$

where η_k is from Bernoulli distribution and $\boldsymbol{\eta}_{-k}$ is a vector obtained by removing the k-th term from $\boldsymbol{\eta}$. $\mathbb{E}_{k' \neq k}(\cdot)$ denotes taking expectation with respect to the terms outside the k-th group. Now given $q(\eta_k)$, when $\eta_k = 1$, we can focus on the expectations in

Equation (2.28):

$$
\begin{aligned}
&\mathbb{E}_{q(\boldsymbol{\gamma}_k,\boldsymbol{\beta}_k|\eta_k=1)}\left[\mathbb{E}_{k'\neq k}\log\Pr(\mathbf{y},\boldsymbol{\eta}_{-k},\eta_k=1,\boldsymbol{\gamma},\boldsymbol{\beta})-\log q(\eta_k=1,\boldsymbol{\gamma}_k,\boldsymbol{\beta}_k)\right]\\
=&\sum_{\boldsymbol{\gamma}_k}\prod_j^{l_k}q(\gamma_{jk})\int_{\boldsymbol{\beta}_k}\left(\mathbb{E}_{k'\neq k}\log\Pr(\mathbf{y},\boldsymbol{\eta}_{-k},\eta_k=1,\boldsymbol{\gamma},\boldsymbol{\beta})-\log q(\eta_k=1,\boldsymbol{\gamma}_k,\boldsymbol{\beta}_k)\right)\\
&\prod_j^{l_k}q(\beta_{jk}|\eta_k,\gamma_{jk})d\boldsymbol{\beta}_k\\
=&\sum_{\boldsymbol{\gamma}_k}\prod_j^{l_k}q(\gamma_{jk})\int_{\boldsymbol{\beta}_k}\left(\mathbb{E}_{k'\neq k}\log\Pr(\mathbf{y},\boldsymbol{\eta}_{-k},\eta_k=1,\boldsymbol{\gamma},\boldsymbol{\beta})-\log q(\boldsymbol{\gamma}_k,\boldsymbol{\beta}_k|\eta_k=1)\right)\\
&\prod_j^{l_k}q(\beta_{jk}|\eta_k,\gamma_{jk})d\boldsymbol{\beta}_k+\text{const}\\
=&\sum_{\gamma_{jk}}q(\gamma_{jk})\int q(\beta_{jk}|\gamma_{jk},\eta_k=1)[\sum_{\gamma_{-j|k}}\prod_{j'\neq j|k}q(\gamma_{j'k})\int\mathbb{E}_{k'\neq k}\left[\log\Pr(\mathbf{y},\boldsymbol{\eta}_{-k},\eta_k=1,\boldsymbol{\gamma},\boldsymbol{\beta})\right]\\
&\prod_{j'\neq j|k}q(\beta_{j'k},\gamma_{j'k}|\eta_k=1)d\beta_{j'k}]d\beta_{jk}\\
&-\sum_{\gamma_{jk}}q(\gamma_{jk})\int q(\beta_{jk}|\gamma_{jk},\eta_k=1)\log q(\beta_{jk},\gamma_{jk}|\eta_k=1)d\beta_{jk}+\text{const}\\
=&\mathbb{E}_{q(\beta_{jk},\gamma_{jk}|\eta_k=1)}\left[\mathbb{E}_{j'\neq j|k}\left[\mathbb{E}_{k'\neq k}\log\Pr(\mathbf{y},\boldsymbol{\eta}_{-k},\eta_k=1,\boldsymbol{\gamma},\boldsymbol{\beta})\right]-\log q(\beta_{jk},\gamma_{jk}|\eta_k=1)\right]+\text{const}\\
=&q(\gamma_{jk}=1)\mathbb{E}_{q(\beta_{jk}|\eta_k=1,\gamma_{jk}=1)}[\mathbb{E}_{j'\neq j|k}\left[\mathbb{E}_{k'\neq k}\log\Pr(\mathbf{y},\boldsymbol{\eta}_{-k},\eta_k=1,\boldsymbol{\gamma}_{-jk},\gamma_{jk}=1,\boldsymbol{\beta})\right]\\
&-\log q(\beta_{jk},\gamma_{jk}=1|\eta_k=1)]\\
&+q(\gamma_{jk}=0)\mathbb{E}_{q(\beta_{jk}|\eta_k=1,\gamma_{jk}=0)}[\mathbb{E}_{j'\neq j|k}\left[\mathbb{E}_{k'\neq k}\log\Pr(\mathbf{y},\boldsymbol{\eta}_{-k},\eta_k=1,\boldsymbol{\gamma}_{-jk},\gamma_{jk}=0,\boldsymbol{\beta})\right]\\
&-\log q(\beta_{jk},\gamma_{jk}=0|\eta_k=1)].
\end{aligned}
\tag{2.29}
$$

where the last equation is because of the assumption $q(\beta_{jk},\gamma_{jk}|\eta_k)=q(\beta_{jk}|\gamma_{jk},\eta_k)q(\gamma_{jk})$ and $\boldsymbol{\gamma}_{-jk}$ is a vector obtained by removing the jk-th term in $\boldsymbol{\gamma}$. $\mathbb{E}_{j'\neq j|k}(\cdot)$ denotes taking the expectation with respect to all variables inside the k-th group except the j-th one. Again, given $q(\gamma_{jk})$, when $\gamma_{jk}=1$, we can further derive with a similar procedure from the expectation in Equation (2.29) that:

$$
\begin{aligned}
&\mathbb{E}_{q(\beta_{jk}|\eta_k=1,\gamma_{jk}=1)}[\mathbb{E}_{j'\neq j|k}\left[\mathbb{E}_{k'\neq k}\log\Pr(\mathbf{y},\boldsymbol{\eta}_{-k},\eta_k=1,\boldsymbol{\gamma}_{-jk},\gamma_{jk}=1,\boldsymbol{\beta})\right]\\
&-\log q(\beta_{jk},\gamma_{jk}=1|\eta_k=1)]\\
=&\mathbb{E}_{q(\beta_{jk}|\eta_k=1,\gamma_{jk}=1)}[\mathbb{E}_{j'\neq j|k}\left[\mathbb{E}_{k'\neq k}\log\Pr(\mathbf{y},\boldsymbol{\eta}_{-k},\eta_k=1,\boldsymbol{\gamma}_{-jk},\gamma_{jk}=1,\boldsymbol{\beta})\right]\\
&-\log q(\beta_{jk}|\eta_k=1,\gamma_{jk}=1)]+\text{const},
\end{aligned}
\tag{2.30}
$$

which is a KL Divergence between $\mathbb{E}_{j'\neq j|k}\left[\mathbb{E}_{k'\neq k}\log\Pr(\mathbf{y},\boldsymbol{\eta}_{-k},\eta_k=1,\boldsymbol{\gamma}_{-jk},\gamma_{jk}=1,\boldsymbol{\beta})\right]$ and $q(\beta_{jk}|\eta_k=1,\gamma_{jk}=1)$ given $\eta_k=1$ and $\gamma_{jk}=1$. Hence the optimal form of $q^*(\beta_{jk}|\eta_k=1,\gamma_{jk}=1)$ is given by

$$\log q^*(\beta_{jk}|\eta_k=1,\gamma_{jk}=1)=\mathbb{E}_{j'\neq j|k}\left[\mathbb{E}_{k'\neq k}\log\Pr(\mathbf{y},\boldsymbol{\eta}_{-k},\eta_k=1,\boldsymbol{\gamma}_{-jk},\gamma_{jk}=1,\boldsymbol{\beta})\right]. \tag{2.31}$$

Here we only derive the case when $\eta_k=\gamma_{jk}=1$, other cases can be easily derived following the same procedure. Since both η_k and γ_{jk} are from Bernoulli distribution, with the expression in equation (2.31), we can first impose some variational parameters on $q(\gamma_{jk})$ and $q(\eta_k)$, then derive the conditional distribution of β_{jk} given η_k and γ_{jk}, and lastly optimize the lower bound to find the variational parameters.
First, we derive $q(\beta_{jk}|\eta_k,\gamma_{jk})$ which involves the joint probability function. The logarithm of the joint probability function is given as

$$\begin{aligned}
&\log\Pr(\mathbf{y},\boldsymbol{\eta},\boldsymbol{\gamma},\boldsymbol{\beta}|\mathbf{X},\mathbf{Z};\boldsymbol{\theta})\\
=&-\frac{n}{2}\log(2\pi\sigma_e^2)-\frac{\mathbf{y}^T\mathbf{y}}{2\sigma_e^2}-\frac{(\mathbf{Z}\boldsymbol{\omega})^T(\mathbf{Z}\boldsymbol{\omega})}{2\sigma_e^2}\\
&+\frac{\sum_k^K\sum_j^{l_k}\eta_k\gamma_{jk}\beta_{jk}\mathbf{x}_{jk}^T\mathbf{y}}{\sigma_e^2}+\frac{\mathbf{y}^T(\mathbf{Z}\boldsymbol{\omega})}{\sigma_e^2}-\frac{\sum_k^K\sum_j^{l_k}\eta_k\gamma_{jk}\beta_{jk}\mathbf{x}_{jk}^T(\mathbf{Z}\boldsymbol{\omega})}{\sigma_e^2}\\
&-\frac{1}{2\sigma_e^2}\sum_k^K\sum_j^{l_k}\left(\left(\eta_k\gamma_{jk}\beta_{jk}\right)^2\mathbf{x}_{jk}^T\mathbf{x}_{jk}\right)\\
&-\frac{1}{2\sigma_e^2}\left(\sum_k^K\sum_{k'\neq k}^K\sum_j^{j_k}\sum_{j'}^{l_{k'}}\left(\eta_{k'}\gamma_{j'k'}\beta_{j'k'}\right)\left(\eta_k\gamma_{jk}\beta_{jk}\right)\mathbf{x}_{j'k'}^T\mathbf{x}_{jk}\right)\\
&-\frac{1}{2\sigma_e^2}\left(\sum_k^K\sum_j^{l_k}\sum_{j'\neq j}^{l_k}\left(\eta_k\gamma_{j'k}\beta_{j'k}\right)\left(\eta_k\gamma_{jk}\beta_{jk}\right)\mathbf{x}_{j'k}^T\mathbf{x}_{jk}\right)\\
&-\frac{p}{2}\log(2\pi\sigma_\beta^2)-\frac{1}{2\sigma_\beta^2}\sum_{k=1}^K\sum_{j=1}^{l_k}\beta_{jk}^2\\
&+\log(\alpha)\sum_{k=1}^K\sum_{j=1}^{l_k}\gamma_{jk}+\log(1-\alpha)\sum_{k=1}^K\sum_{j=1}^{l_k}(1-\gamma_{jk})\\
&+\log(\pi)\sum_{k=1}^K\eta_k+\log(1-\pi)\sum_{k=1}^K(1-\eta_k).
\end{aligned} \tag{2.32}$$

To find the optimal form in Equation (2.31), We then rearrange Equation (2.32) and only retain the terms regarding jk

$$
\begin{aligned}
&\log\Pr(\mathbf{y},\boldsymbol{\eta},\boldsymbol{\gamma},\boldsymbol{\beta}|\mathbf{X},\mathbf{Z};\boldsymbol{\theta}) \\
=&-\frac{n}{2}\log(2\pi\sigma_e^2)-\frac{\mathbf{y}^T\mathbf{y}}{2\sigma_e^2} \\
&+\frac{\eta_k\gamma_{jk}\beta_{jk}\mathbf{x}_{jk}^T\mathbf{y}}{\sigma_e^2}-\frac{\eta_k\gamma_{jk}\beta_{jk}\mathbf{x}_{jk}^T(\mathbf{Z}\boldsymbol{\omega})}{\sigma_e^2} \\
&-\frac{1}{2\sigma_e^2}\left((\eta_k\gamma_{jk}\beta_{jk})^2\,\mathbf{x}_{jk}^T\mathbf{x}_{jk}\right) \\
&-\frac{1}{2\sigma_e^2}\left(\sum_{k'\neq k}^{K}\sum_{j'}^{l_{k'}}(\eta_k\gamma_{jk}\beta_{jk})\,(\eta_k\gamma_{j'k'}\beta_{j'k'})\,\mathbf{x}_{jk}^T\mathbf{x}_{j'k'}\right) \\
&-\frac{1}{2\sigma_e^2}\left(\sum_{j'\neq j}^{l_k}(\eta_k\gamma_{jk}\beta_{jk})\,(\eta_k\gamma_{j'k}\beta_{j'k})\,\mathbf{x}_{jk}^T\mathbf{x}_{j'k}\right)-\frac{1}{2\sigma_\beta^2}\beta_{jk}^2 \\
&+\log(\alpha)\gamma_{jk}+\log(1-\alpha)(1-\gamma_{jk}) \\
&+\log(\pi)\eta_k+\log(1-\pi)(1-\eta_{jk}) \\
&+\text{const.}
\end{aligned}
\tag{2.33}
$$

Now we can derive the log $q(\beta_{jk}|\eta_k=1,\gamma_{jk}=1)$ by taking the expectation in Equation (2.31). When $\eta_k=\gamma_{jk}=1\Leftrightarrow\eta_k\gamma_{jk}=1$, we have

$$
\begin{aligned}
&\log\ q(\beta_{jk}|\eta_k=1,\gamma_{jk}=1) \\
=&\left(-\frac{1}{2\sigma_e^2}\mathbf{x}_{jk}^T\mathbf{x}_{jk}-\frac{1}{2\sigma_\beta^2}\right)\beta_{jk}^2 \\
+&\left(\frac{\mathbf{x}_{jk}^T(\mathbf{y}-\mathbf{Z}\boldsymbol{\omega})-\sum_{k'\neq k}^{K}\sum_{j'}^{l_{k'}}\mathbb{E}_{k'\neq k}[\eta_{k'}\gamma_{j'k'}\beta_{j'k'}]\mathbf{x}_{jk}^T\mathbf{x}_{j'k'}-\sum_{j'\neq j}^{l_k}\mathbb{E}_{j'\neq j|k}[\gamma_{j'k}\beta_{j'k}]\mathbf{x}_{jk}^T\mathbf{x}_{j'k}}{\sigma_e^2}\right)\beta_{jk} \\
&+\text{const.}
\end{aligned}
\tag{2.34}
$$

Since Equation (2.34) is a quadratic form of β_{jk}, the posterior of $q(\beta_{jk}|\eta_k=1,\gamma_{jk}=1)$ follows a Gaussian of the form $\mathcal{N}(\mu_{jk},s_{jk}^2)$, where

$$s_{jk}^2 = \frac{\sigma_e^2}{\mathbf{x}_{jk}^T \mathbf{x}_{jk} + \frac{\sigma_e^2}{\sigma_\beta^2}} \tag{2.35}$$

$$\mu_{jk} = \frac{\mathbf{x}_{jk}^T(\mathbf{y} - \mathbf{Z}\boldsymbol{\omega}) - \sum_{k' \neq k}^{K} \sum_{j'}^{l_{k'}} \mathbb{E}_{k' \neq k}[\eta_{k'}\gamma_{j'k'}\beta_{j'k'}]\mathbf{x}_{jk}^T \mathbf{x}_{j'k'} - \sum_{j' \neq j}^{l_k} \mathbb{E}_{j' \neq j|k}[\gamma_{j'k}\beta_{j'k}]\mathbf{x}_{jk}^T \mathbf{x}_{j'k}}{\mathbf{x}_{jk}^T \mathbf{x}_{jk} + \frac{\sigma_e^2}{\sigma_\beta^2}}.$$

Similarly, for $\eta_k\gamma_{jk} = 0$, we have

$$\log\, q(\beta_{jk}|\eta_k\gamma_{jk} = 0) = -\frac{1}{2\sigma_\beta^2}\beta_{jk}^2 + \text{const}, \tag{2.36}$$

which implies that $q(\beta_{jk}|\eta_k\gamma_{jk} = 0) \sim \mathcal{N}(0, \sigma_\beta^2)$. Thus, the conditional posterior of β_{jk} is exactly the same as the prior if this variable is irrelevant in either one of the two levels ($\eta_k\gamma_{jk} = 0$). Now we turn to $q(\eta_k)$ and $q(\gamma_{jk})$. Denote $\pi_k = q(\eta_k)$ and $\alpha_{jk} = q(\gamma_{jk})$, we have

$$q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}) = \prod_k^K \left(\pi_k^{\eta_k}(1-\pi_k)^{1-\eta_k} \prod_j^{l_k} \left(\alpha_{jk}^{\gamma_{jk}}(1-\alpha_{jk})^{1-\gamma_{jk}} \mathcal{N}(\mu_{jk}, s_{jk}^2)^{\eta_k\gamma_{jk}} \mathcal{N}(0, \sigma_\beta)^{1-\eta_k\gamma_{jk}} \right) \right). \tag{2.37}$$

And the second term in $\mathcal{L}(q)$ can be derived as:

$$
\begin{aligned}
&- \mathbb{E}_q[\log q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta})] \\
&= - \mathbb{E}_q \left[\sum_k^K (\eta_k \log(\pi_k) + (1-\eta_k)\log(1-\pi_k)) \right] - \mathbb{E}_q \left[\sum_k^K \sum_j^{l_k} \eta_k \gamma_{jk} \log \mathcal{N}(\mu_{jk}, s_{jk}^2) \right] \\
&\quad - \mathbb{E}_q \left[\sum_k^K \sum_j^{l_k} (1-\eta_k\gamma_{jk})\mathcal{N}(0, \sigma_\beta^2) + \gamma_{jk}\log(\alpha_{jk}) + (1-\gamma_{jk})\log(1-\alpha_{jk}) \right] \\
&= - \sum_k^K \mathbb{E}_q \left[(\eta_k \log(\pi_k) + (1-\eta_k)\log(1-\pi_k)) \right] - \sum_k^K \sum_j^{l_k} \mathbb{E}_q \left[\eta_k \gamma_{jk} \log \mathcal{N}(\mu_{jk}, s_{jk}^2) \right] \\
&\quad - \sum_k^K \sum_j^{l_k} \mathbb{E}_q \left[(1-\eta_k\gamma_{jk})\mathcal{N}(0, \sigma_\beta^2) + \gamma_{jk}\log(\alpha_{jk}) + (1-\gamma_{jk})\log(1-\alpha_{jk}) \right] \\
&= - \sum_k^K \sum_j^{l_k} \mathbb{E}_{\eta_k, \gamma_{jk}} \{ \mathbb{E}_{\beta|\eta_k=1,\gamma_{jk}=1}[\log \mathcal{N}(\mu_{jk}, s_{jk}^2)] + \mathbb{E}_{\beta|\eta_k=1,\gamma_{jk}=0}[\log \mathcal{N}(0, \sigma_\beta^2)] \\
&\quad + \mathbb{E}_{\beta|\eta_k=0,\gamma_{jk}=1}[\log \mathcal{N}(0, \sigma_\beta^2)] + \mathbb{E}_{\beta|\eta_k=0,\gamma_{jk}=0}[\log \mathcal{N}(0, \sigma_\beta^2)] \} \\
&\quad - \sum_k^K \sum_j^{l_k} \mathbb{E}_q[\gamma_{jk}\log(\alpha_{jk}) + (1-\gamma_{jk})\log(1-\alpha_{jk})] \\
&\quad - \sum_k^K \mathbb{E}_q[\eta_k \log(\pi_k) + (1-\eta_k)\log(1-\pi_k)] \qquad (2.38)
\end{aligned}
$$

Note that $-\mathbb{E}_{\beta|\eta_k=1,\gamma_{jk}=1}[\log \mathcal{N}(\mu_{jk}, s_{jk}^2)]$ is the entropy of Gaussian, so we have $-\mathbb{E}_{\beta|\eta_k=1,\gamma_{jk}=1}[\log \mathcal{N}(\mu_{jk}, s_{jk}^2)] = \frac{1}{2}\log(s_{jk}^2) + \frac{1}{2}(1+\log(2\pi))$, similarly, $-\mathbb{E}_{\beta|\eta_k=1,\gamma_{jk}=0}[\log \mathcal{N}(\mu_{jk}, s_{jk}^2)] = \frac{1}{2}\log(\sigma_\beta^2) + \frac{1}{2}(1+\log(2\pi))$ and so on. Consequently, Equation (2.38) can be further

derived in:

$$
\begin{aligned}
&- \mathbb{E}[\log q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta})] \\
=&\sum_{k}^{K}\sum_{j}^{l_k}\{[\frac{1}{2}\log(s_{jk}^2) + \frac{1}{2}(1+\log(2\pi))]\pi_k\alpha_{jk} + [\frac{1}{2}\log(\sigma_\beta^2) + \frac{1}{2}(1+\log(2\pi))](1-\pi_k\alpha_{jk}) \\
&- \alpha_{jk}\log(\alpha_{jk}) - (1-\alpha_{jk})\log(1-\alpha_{jk})\} - \sum_{k}^{K}\{\pi_k\log(\pi_k) + (1-\pi_k)\log(1-\pi_k)\} \\
=&\sum_{k}^{K}\sum_{j}^{l_k}\frac{1}{2}\pi_k\alpha_{jk}(\log s_{jk}^2 - \log\sigma_\beta^2) + \frac{p}{2}\log(\sigma_\beta^2) + \frac{p}{2} + \frac{p}{2}\log(2\pi) \\
&- \sum_{k}^{K}\sum_{j}^{l_k}[\alpha_{jk}\log(\alpha_{jk}) + (1-\alpha_{jk})\log(1-\alpha_{jk})] - \sum_{k}^{K}[\pi_k\log(\pi_k) + (1-\pi_k)\log(1-\pi_k)].
\end{aligned}
\tag{2.39}
$$

Combine Equation (2.39) with Equation (2.32), the lower bound is obtained as

follow:

$$
\begin{aligned}
&\mathbb{E}_q[\log \Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta} | \mathbf{X}; \boldsymbol{\theta})] - \mathbb{E}_q[\log q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta})] \\
=& -\frac{n}{2}\log(2\pi\sigma_e^2) - \frac{\mathbf{y}^T\mathbf{y}}{2\sigma_e^2} - \frac{(\mathbf{Z}\boldsymbol{\omega})^T(\mathbf{Z}\boldsymbol{\omega})}{2\sigma_e^2} \\
&+ \frac{\sum_k^K \sum_j^{l_k} \mathbb{E}_q\left[\eta_k \gamma_{jk} \beta_{jk}\right] \mathbf{x}_{jk}^T \mathbf{y}}{\sigma_e^2} + \frac{\mathbf{y}^T(\mathbf{Z}\boldsymbol{\omega})}{\sigma_e^2} - \frac{\sum_k^K \sum_j^{l_k} \mathbb{E}_q\left[\eta_k \gamma_{jk} \beta_{jk}\right] \mathbf{x}_{jk}^T (\mathbf{Z}\boldsymbol{\omega})}{\sigma_e^2} \\
&- \frac{1}{2\sigma_e^2} \sum_k^K \sum_j^{l_k} \left(\mathbb{E}_q\left[\left(\eta_k \gamma_{jk} \beta_{jk}\right)^2 \right] \mathbf{x}_{jk}^T \mathbf{x}_{jk} \right) \\
&- \frac{1}{2\sigma_e^2} \left(\sum_k^K \sum_{k' \neq k}^K \sum_j^{j_k} \sum_{j'}^{l_{k'}} \mathbb{E}_q\left[\eta_{k'} \gamma_{j'k'} \beta_{j'k'}\right] \mathbb{E}_q\left[\eta_k \gamma_{jk} \beta_{jk}\right] \mathbf{x}_{j'k'}^T \mathbf{x}_{jk} \right) \\
&- \frac{1}{2\sigma_e^2} \left(\sum_k^K \sum_j^{l_k} \sum_{j' \neq j}^{l_k} \mathbb{E}_q\left[\eta_k^2 \gamma_{j'k} \beta_{j'k} \gamma_{jk} \beta_{jk}\right] \mathbf{x}_{j'k}^T \mathbf{x}_{jk} \right) \\
&- \frac{p}{2}\log(2\pi\sigma_\beta^2) - \frac{1}{2\sigma_\beta^2} \sum_{k=1}^K \sum_{j=1}^{l_k} \mathbb{E}_q\left[\beta_{jk}^2\right] \\
&+ \log(\alpha) \sum_{k=1}^K \sum_{j=1}^{l_k} \mathbb{E}_q\left[\gamma_{jk}\right] + \log(1-\alpha) \sum_{k=1}^K \sum_{j=1}^{l_k} \mathbb{E}_q\left[1 - \gamma_{jk}\right] \\
&+ \log(\pi) \sum_{k=1}^K \mathbb{E}_q\left[\eta_k\right] + \log(1-\pi) \sum_{k=1}^K \mathbb{E}_q\left[1 - \eta_k\right] \\
&+ \sum_k^K \sum_j^{l_k} \frac{1}{2}\pi_k \alpha_{jk} (\log s_{jk}^2 - \log\sigma_\beta^2) + \frac{p}{2}\log(\sigma_\beta^2) + \frac{p}{2} + \frac{p}{2}\log(2\pi) \\
&- \sum_k^K \sum_j^{l_k} [\alpha_{jk}\log(\alpha_{jk})] - \sum_k^K \sum_j^{l_k} [(1-\alpha_{jk})\log(1-\alpha_{jk})] \\
&- \sum_k^K [\pi_k \log(\pi_k)] - \sum_k^K [(1-\pi_k)\log(1-\pi_k)],
\end{aligned}
\tag{2.40}
$$

where expectations in are derived as follows:

$$
\mathbb{E}_q\left[\eta_k\right] = \pi_k, \quad \mathbb{E}_q\left[\gamma_{jk}\right] = \alpha_{jk}, \tag{2.41}
$$

$$\begin{aligned}\mathbb{E}\left[\eta_k\gamma_{jk}\beta_{jk}\right] &= \sum_{\gamma_{jk}}\sum_{\eta_k}\int_{\beta_{jk}} \eta_k\gamma_{jk}\beta_{jk}q(\beta_{jk}|\eta_k,\gamma_{jk})q(\eta_k)q(\gamma_{jk})d\beta_{jk} \\ &= \pi_k\alpha_{jk}\cdot\mu_{jk} + (1-\pi_k\alpha_{jk})\cdot 0 \\ &= \pi_k\alpha_{jk}\mu_{jk} \end{aligned} \tag{2.42}$$

$$\begin{aligned}\mathbb{E}_q\left[\beta_{jk}^2\right] &= \int_{\beta_{jk}} \beta_{jk}^2 q(\beta_{jk})d\beta_{jk} \\ &= \sum_{\eta_k}\sum_{\gamma_{jk}}\int_{\beta_{jk}} \beta_{jk}^2 q(\beta_{jk}|\eta_k,\gamma_{jk})q(\eta_k)q(\gamma_{jk})d\beta_{jk} \\ &= \int_{\beta_{jk}} \beta_{jk}^2\cdot\left[\pi_k\alpha_{jk}\mathcal{N}(\mu_{jk},s_{jk}^2) + (1-\pi_k\alpha_{jk})\mathcal{N}(0,\sigma_\beta^2)\right]d\beta_{jk} \\ &= \pi_k\alpha_{jk}(s_{jk}^2+\mu_{jk}^2) + (1-\pi_k\alpha_{jk})\sigma_\beta^2 \end{aligned} \tag{2.43}$$

$$\begin{aligned}\mathbb{E}_q\left[(\eta_k\gamma_{jk}\beta_{jk})^2\right] &= \sum_{\eta_k}\sum_{\gamma_{jk}}\int_{\beta_{jk}} \eta_k\gamma_{jk}\beta_{jk}^2 q(\beta_{jk}|\eta_k,\gamma_{jk})q(\eta_k)q(\gamma_{jk})d\beta_{jk} \\ &= \pi_k\alpha_{jk}\int_{\beta_{jk}} \beta_{jk}^2\mathcal{N}(\mu_{jk},s_{jk}^2))d\beta_{jk} \\ &= \pi_k\alpha_{jk}(s_{jk}^2+\mu_{jk}^2) \end{aligned} \tag{2.44}$$

$$\begin{aligned}&\mathbb{E}_q\left[\eta_k^2\gamma_{j'k}\beta_{j'k}\gamma_{jk}\beta_{jk}\right] \\ =&\mathbb{E}_q\left[\eta_k\gamma_{j'k}\beta_{j'k}\gamma_{jk}\beta_{jk}\right] \\ =&\sum_{\eta_k}\sum_{\gamma_{jk},\gamma_{j'k}}\int\int \eta_k\gamma_{j'k}\beta_{j'k}\gamma_{jk}\beta_{jk}q(\beta_{jk}|\eta_k,\gamma_{jk})q(\gamma_{jk})q(\beta_{j'k}|\eta_k,\gamma_{j'k})q(\gamma_{j'k})q(\eta_k)d\beta_{jk}d\beta_{j'k} \\ =&\pi_k\alpha_{j'k}\mu_{j'k}\alpha_{jk}\mu_{jk} \end{aligned} \tag{2.45}$$

We plug in the evaluations from Equation (2.41) to (2.45), $\mathcal{L}(q)$ then becomes

$$
\begin{aligned}
&\mathbb{E}_q[\log \Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta} | \mathbf{X}; \boldsymbol{\theta})] - \mathbb{E}_q[\log q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta})] \\
=& -\frac{n}{2}\log(2\pi\sigma_e^2) - \frac{||\mathbf{y} - \mathbf{Z}\boldsymbol{\omega} - \sum_k^K \sum_j^{l_k} \pi_k \alpha_{jk} \mu_{jk} \mathbf{x}_{jk}||^2}{2\sigma_e^2} \\
&- \frac{1}{2\sigma_e^2} \sum_k^K \sum_j^{l_k} \underbrace{[\pi_k \alpha_{jk}(s_{jk}^2 + \mu_{jk}^2) - (\pi_k \alpha_{jk} \mu_{jk})^2]}_{\mathrm{Var}[\eta_k \gamma_{jk} \beta_{jk}]} \mathbf{x}_{jk}^T \mathbf{x}_{jk} \\
&- \frac{1}{2\sigma_e^2} \sum_k^K \left(\pi_k - \pi_k^2\right) \left(\sum_j^{l_k} \sum_{j' \neq j}^{l_k} \alpha_{j'k} \mu_{j'k} \alpha_{jk} \mu_{jk} \mathbf{x}_{j'k}^T \mathbf{x}_{jk} \right) \\
&- \frac{p}{2}\log(2\pi\sigma_\beta^2) - \frac{1}{2\sigma_\beta^2} \sum_{k=1}^K \sum_{j=1}^{l_k} [\pi_k \alpha_{jk}(s_{jk}^2 + \mu_{jk}^2) + (1 - \pi_k \alpha_{jk})\sigma_\beta^2] \\
&+ \sum_{k=1}^K \sum_{j=1}^{l_k} \alpha_{jk} \log(\frac{\alpha}{\alpha_{jk}}) + \sum_{k=1}^K \sum_{j=1}^{l_k} (1 - \alpha_{jk}) \log(\frac{1-\alpha}{1-\alpha_{jk}}) \\
&+ \sum_{k=1}^K \pi_k \log(\frac{\pi}{\pi_k}) + \sum_{k=1}^K (1 - \pi_k) \log(\frac{1-\pi}{1-\pi_k}) \\
&+ \sum_k^K \sum_j^{l_k} \frac{1}{2} \pi_k \alpha_{jk} \log(\frac{s_{jk}^2}{\sigma_\beta^2}) + \frac{p}{2}\log(\sigma_\beta^2) + \frac{p}{2} + \frac{p}{2}\log(2\pi)
\end{aligned}
\tag{2.46}
$$

To get π_k and α_{jk}, we set

$$
\frac{\partial \mathbb{E}_q[\log \Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta} | \mathbf{X}; \boldsymbol{\theta})] - \mathbb{E}_q[\log q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta})]}{\partial \pi_k} = 0,
$$

$$
\frac{\partial \mathbb{E}_q[\log \Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta} | \mathbf{X}; \boldsymbol{\theta})] - \mathbb{E}_q[\log q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta})]}{\partial \alpha_{jk}} = 0,
$$

which gives

$$
\begin{aligned}
\pi_k &= \frac{1}{1 + \exp(-u_k)}, \\
\text{where } u_k &= \log\frac{\pi}{1-\pi} + \frac{1}{2} \sum_j^{l_k} \alpha_{jk} \left(\log\frac{s_{jk}^2}{\sigma_\beta^2} + \frac{\mu_{jk}^2}{s_{jk}^2} \right); \\
\text{and } \alpha_{jk} &= \frac{1}{1 + \exp(-v_{jk})}, \\
\text{where } v_{jk} &= \log\frac{\alpha}{1-\alpha} + \frac{1}{2}\pi_k \left(\log\frac{s_{jk}^2}{\sigma_\beta^2} + \frac{\mu_{jk}^2}{s_{jk}^2} \right).
\end{aligned}
\tag{2.47}
$$

The derivation is as follow:

$$
\begin{aligned}
u_k =& \log\frac{\pi}{1-\pi} + \frac{1}{2}\sum_j^{l_k}\alpha_{jk}\log\frac{s_{jk}^2}{\sigma_\beta^2} \\
&+ \frac{\sum_j^{l_k}\alpha_{jk}\mu_{jk}\mathbf{x}_{jk}^T\mathbf{y}}{\sigma_e^2} - \frac{\sum_j^{l_k}\alpha_{jk}\mu_{jk}\mathbf{x}_{jk}^T(\mathbf{Z}\boldsymbol{\omega})}{\sigma_e^2} - \frac{1}{2\sigma_e^2}\sum_j^{l_k}\alpha_{jk}(s_{jk}^2+\mu_{jk}^2)\mathbf{x}_{jk}^T\mathbf{x}_{jk} \\
&- \frac{1}{\sigma_e^2}\left(\sum_{k'\neq k}^{K}\sum_j^{l_k}\sum_{j'}^{l_{k'}}\pi_{k'}\alpha_{j'k'}\mu_{j'k'}\alpha_{jk}\mu_{jk}\mathbf{x}_{j'k'}^T\mathbf{x}_{jk}\right) \\
&- \frac{1}{\sigma_e^2}\left(\sum_j^{l_k}\sum_{j'\neq j}^{l_k}\alpha_{j'k}\mu_{j'k}\alpha_{jk}\mu_{jk}\mathbf{x}_{j'k}^T\mathbf{x}_{jk}\right) \\
&- \frac{1}{2\sigma_\beta}\sum_j^{l_k}\alpha_{jk}(s_{jk}^2+\mu_{jk}^2) + \frac{1}{2}\sum_j^{l_k}\alpha_{jk} \\
=& \log\frac{\pi}{1-\pi} + \frac{1}{2}\sum_j^{l_k}\alpha_{jk}\log\frac{s_{jk}^2}{\sigma_\beta^2} \\
&+ \alpha_{jk}\mu_{jk}\sum_j^{K}\underbrace{\left(\frac{\mathbf{x}_{jk}^T(\mathbf{y}-\mathbf{Z}\boldsymbol{\omega}) - \sum_{k'\neq k}^{K}\pi_{k'}\sum_{j'}^{l_{k'}}\alpha_{j'k'}\mu_{j'k'}\mathbf{x}_{j'k'}^T\mathbf{x}_{jk} - \sum_{j'\neq j}^{l_k}\alpha_{j'k}\mu_{j'k}\mathbf{x}_{j'k}^T\mathbf{x}_{jk}}{\sigma_e^2}\right)}_{\mu_{jk}/s_{jk}^2} \\
&- \frac{1}{2}\sum_j^{l_k}\alpha_{jk}(s_{jk}^2+\mu_{jk}^2)\underbrace{\left(\frac{\mathbf{x}_{jk}^T\mathbf{x}_{jk}}{\sigma_e^2}+\frac{1}{\sigma_\beta^2}\right)}_{1/s_{jk}^2} + \frac{1}{2}\sum_j^{l_k}\alpha_{jk} \\
=& \log\frac{\pi}{1-\pi} + \frac{1}{2}\sum_j^{l_k}\alpha_{jk}\log\frac{s_{jk}^2}{\sigma_\beta^2} + \sum_j^{l_k}\frac{\alpha_{jk}\mu_{jk}^2}{s_{jk}^2} - \sum_j^{l_k}\frac{\alpha_{jk}\mu_{jk}^2}{2s_{jk}^2} \\
=& \log\frac{\pi}{1-\pi} + \frac{1}{2}\sum_j^{l_k}\alpha_{jk}\left(\log\frac{s_{jk}^2}{\sigma_\beta^2} + \frac{\mu_{jk}^2}{s_{jk}^2}\right),
\end{aligned}
\tag{2.48}
$$

where we have used Equation (2.35) in the third equation. Derivation of v_{jk} follows the same procedure.

M-step

At M-step, we update the parameters $\boldsymbol{\theta} = \{\alpha, \pi, \sigma_\beta^2, \sigma_e^2, \boldsymbol{\omega}\}$. By setting $\frac{\partial \mathbb{E}_q[\log \Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta} | \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta})]}{\partial \sigma_e^2} = 0$, we have

$$\begin{aligned}\sigma_e^2 =& \frac{||\mathbf{y} - \mathbf{Z}\boldsymbol{\omega} - \sum_k^K \sum_j^{l_k} \pi_k \alpha_{jk} \mu_{jk} \mathbf{x}_{jk}||^2}{n} \\ &+ \frac{\sum_k^K \sum_j^{l_k} [\pi_k \alpha_{jk}(s_{jk}^2 + \mu_{jk}^2) - (\pi_k \alpha_{jk} \mu_{jk})^2] \mathbf{x}_{jk}^T \mathbf{x}_{jk}}{n} \\ &+ \frac{\sum_k^K (\pi_k - \pi_k^2)[\sum_j^{l_k} \sum_{j'}^{l_k} \alpha_{j'k} \mu_{j'k} \alpha_{jk} \mu_{jk}] \mathbf{x}_{j'k}^T \mathbf{x}_{jk}}{n}.\end{aligned} \tag{2.49}$$

To get σ_β^2, we set $\frac{\partial \mathcal{L}(q)}{\partial \sigma_\beta^2} = 0$, which gives

$$\sigma_\beta^2 = \frac{\sum_k^K \sum_j^{l_k} \pi_k \alpha_{jk}(s_{jk}^2 + \mu_{jk}^2)}{\sum_k^K \sum_j^{l_k} \pi_k \alpha_{jk}}. \tag{2.50}$$

Accordingly,

$$\alpha = \frac{1}{p} \sum_k^K \sum_j^{l_k} \alpha_{jk}, \tag{2.51}$$

$$\pi = \frac{1}{K} \sum_k^K \pi_k. \tag{2.52}$$

$$\boldsymbol{\omega} = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T (\mathbf{y} - \sum_k^K \sum_j^{l_k} \pi_k \alpha_{jk} \mu_{jk} \mathbf{x}_{jk}). \tag{2.53}$$

2.5.2 Variational EM Algorithm: Multi-task Learning with BIVAS

E-step

Let $\boldsymbol{\theta} = \{\alpha, \pi, \sigma_{\beta_j}^2, \sigma_{e_j}^2, \boldsymbol{\omega}_j\}_{j=1}^l$ be the collection of model parameters. The joint probabilistic model is

$$\begin{aligned}&\Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta} | \mathbf{X}, \mathbf{Z}; \theta) = \Pr(\mathbf{y} | \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \mathbf{X}, \mathbf{Z}, \theta) \Pr(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta} | \theta) \\ =& \prod_{j=1}^L \mathcal{N}(\mathbf{y}_j | \mathbf{Z}_j \boldsymbol{\omega}_j + \sum_k^K \eta_k \gamma_{jk} \beta_{jk} \mathbf{x}_{jk}) \prod_{k=1}^K \pi^{\eta_k} (1 - \pi)^{1 - \eta_k} \prod_{j=1}^L \mathcal{N}(0, \sigma_{\beta_j}^2) \alpha^{\gamma_{jk}} (1 - \alpha)^{1 - \gamma_{jk}}.\end{aligned} \tag{2.54}$$

The logarithm of the marginal likelihood is

$$\begin{aligned}
\log\, p(\mathbf{y}|\mathbf{X},\mathbf{Z};\theta) &= \log \sum_{\boldsymbol{\gamma}} \sum_{\boldsymbol{\eta}} \int_{\beta} \Pr(\mathbf{y},\boldsymbol{\eta},\boldsymbol{\gamma},\boldsymbol{\beta}|\mathbf{X},\mathbf{Z};\theta) d\boldsymbol{\beta} \\
&\geq \sum_{\boldsymbol{\gamma}} \sum_{\eta} \int_{\beta} q(\boldsymbol{\eta},\boldsymbol{\gamma},\boldsymbol{\beta}) \log \frac{\Pr(\mathbf{y},\boldsymbol{\eta},\boldsymbol{\gamma},\boldsymbol{\beta}|\mathbf{X},\mathbf{Z};\theta)}{q(\boldsymbol{\eta},\boldsymbol{\gamma},\boldsymbol{\beta})} \\
&= \mathbb{E}_q[\log \Pr(\mathbf{y},\boldsymbol{\eta},\boldsymbol{\gamma},\boldsymbol{\beta}|\mathbf{X},\mathbf{Z};\theta) - q(\boldsymbol{\eta},\boldsymbol{\gamma},\boldsymbol{\beta})] \\
&\equiv \mathcal{L}(q).
\end{aligned} \tag{2.55}$$

Again, we assume that the variational distribution takes the form

$$q(\boldsymbol{\eta},\boldsymbol{\gamma},\boldsymbol{\beta}) = \prod_{k}^{K} \left(q(\eta_k) \prod_{j}^{L} (q(\beta_{jk}|\eta_k,\gamma_{jk}) q(\gamma_{jk})) \right). \tag{2.56}$$

Actually, the variational approximation only assumes 'between group' factorizability ($\prod_{k=1}^{K} q(\eta_k,\gamma_{jk},\beta_{jk})$) because given the group, the tasks inside are independent due to model assumption. Follow the same procedure in Section 1.1, the optimal form of q is given by

$$\log q^*(\beta_{jk}|\eta_k=1,\gamma_{jk}=1) = \mathbb{E}_{j'\neq j|k}\left[\mathbb{E}_{k'\neq k} \log \Pr(\mathbf{y},\boldsymbol{\eta}_{-k},\eta_k=1,\boldsymbol{\gamma}_{-jk},\gamma_{jk}=1,\boldsymbol{\beta})\right]. \tag{2.57}$$

The Equation (2.57) contains the logarithm of joint probability funcion, which is

$$
\begin{aligned}
&\log\Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}|\mathbf{X}, \mathbf{Z}; \theta) \\
=&\sum_{j=1}^{L} \Bigg\{ -\frac{N_j}{2}\log(2\pi\sigma_{e_j}^2) - \frac{\mathbf{y}_j^T\mathbf{y}_j}{2\sigma_{e_j}^2} - \frac{(\mathbf{Z}_j\boldsymbol{\omega}_j)^T(\mathbf{Z}_j\boldsymbol{\omega}_j)}{2\sigma_{e_j}^2} \\
&+ \frac{\sum_k^K \eta_k\gamma_{jk}\beta_{jk}\mathbf{x}_{jk}^T\mathbf{y}_j}{\sigma_{e_j}^2} + \frac{\mathbf{y}_j^T(\mathbf{Z}_j\boldsymbol{\omega}_j)}{\sigma_{e_j}^2} - \frac{\sum_k^K \eta_k\gamma_{jk}\beta_{jk}\mathbf{x}_{jk}^T(\mathbf{Z}_j\boldsymbol{\omega}_j)}{\sigma_{e_j}^2} \\
&- \frac{1}{2\sigma_{e_j}^2}\sum_k^K \left((\eta_k\gamma_{jk}\beta_{jk})^2\,\mathbf{x}_{jk}^T\mathbf{x}_{jk}\right) \\
&- \frac{1}{2\sigma_{e_j}^2}\left(\sum_k^K\sum_{k'\neq k}^K (\eta_{k'}\gamma_{jk'}\beta_{jk'})\,(\eta_k\gamma_{jk}\beta_{jk})\,\mathbf{x}_{jk'}^T\mathbf{x}_{jk}\right)\Bigg\} \\
&- \frac{K}{2}\sum_{j=1}^{L}\log(2\pi\sigma_{\beta_j}^2) - \sum_{j=1}^{L}\frac{\sum_{k=1}^K \beta_{jk}^2}{2\sigma_{\beta_j}^2} \\
&+ \log(\alpha)\sum_{k=1}^{K}\sum_{j=1}^{L}\gamma_{jk} + \log(1-\alpha)\sum_{k=1}^{K}\sum_{j=1}^{L}(1-\gamma_{jk}) \\
&+ \log(\pi)\sum_{k=1}^{K}\eta_k + \log(1-\pi)\sum_{k=1}^{K}(1-\eta_k).
\end{aligned}
\tag{2.58}
$$

We then rearrange Equation (2.58) and retain the terms regarding jk

$$
\begin{aligned}
&\log\Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}|\mathbf{X}, \mathbf{Z}; \theta) \\
=&-\frac{N_j}{2}\log(2\pi\sigma_{e_j}^2) - \frac{\mathbf{y}_j^T\mathbf{y}_j}{2\sigma_{e_j}^2} \\
&+ \frac{\eta_k\gamma_{jk}\beta_{jk}\mathbf{x}_{jk}^T\mathbf{y}_j}{\sigma_{e_j}^2} - \frac{\eta_k\gamma_{jk}\beta_{jk}\mathbf{x}_{jk}^T(\mathbf{Z}_j\boldsymbol{\omega}_j)}{\sigma_{e_j}^2} \\
&- \frac{1}{2\sigma_{e_j}^2}\left((\eta_k\gamma_{jk}\beta_{jk})^2\,\mathbf{x}_{jk}^T\mathbf{x}_{jk}\right) \\
&- \frac{1}{2\sigma_{e_j}^2}\left(\sum_{k'\neq k}^K (\eta_k\gamma_{jk}\beta_{jk})\,(\eta_{k'}\gamma_{jk'}\beta_{jk'})\,\mathbf{x}_{jk}^T\mathbf{x}_{jk'}\right) - \frac{1}{2\sigma_{\beta_j}^2}\beta_{jk}^2 \\
&+ \log(\alpha)\gamma_{jk} + \log(1-\alpha)(1-\gamma_{jk}) \\
&+ \log(\pi)\eta_k + \log(1-\pi)(1-\eta_{jk}) \\
&+ \text{const.}
\end{aligned}
\tag{2.59}
$$

When $\eta_k = \gamma_{jk} = 1 \Leftrightarrow \eta_k \gamma_{jk} = 1$, we have

$$\begin{aligned}
&\log q(\beta_{jk} | \eta_k = 1, \gamma_{jk} = 1) \\
&= \left(-\frac{1}{2\sigma_{e_j}^2} \mathbf{x}_{jk}^T \mathbf{x}_{jk} - \frac{1}{2\sigma_{\beta_j}^2} \right) \beta_{jk}^2 \\
&\quad + \left(\frac{\mathbf{x}_{jk}^T (\mathbf{y}_j - \mathbf{Z}_j \boldsymbol{\omega}_j) - \sum_{k' \neq k}^K \mathbb{E}_{k' \neq k} [\eta_{k'} \gamma_{jk'} \beta_{jk'}] \mathbf{x}_{jk}^T \mathbf{x}_{jk'}}{\sigma_{e_j}^2} \right) \beta_{jk} \\
&\quad + \text{const},
\end{aligned} \tag{2.60}$$

from which we can see that the conditional posterior $q(\beta_{jk} | \eta_k = 1, \gamma_{jk} = 1) \sim \mathcal{N}(\mu_{jk}, s_{jk}^2)$, where

$$\begin{aligned}
s_{jk}^2 &= \frac{\sigma_{e_j}^2}{\mathbf{x}_{jk}^T \mathbf{x}_{jk} + \frac{\sigma_{e_j}^2}{\sigma_{\beta_j}^2}} \\
\mu_{jk} &= \frac{\mathbf{x}_{jk}^T (\mathbf{y}_j - \mathbf{Z}_j \boldsymbol{\omega}_j) - \sum_{k' \neq k}^K \mathbb{E}_{k' \neq k} [\eta_{k'} \gamma_{jk'} \beta_{jk'}] \mathbf{x}_{jk}^T \mathbf{x}_{jk'}}{\mathbf{x}_{jk}^T \mathbf{x}_{jk} + \frac{\sigma_{e_j}^2}{\sigma_{\beta_j}^2}}.
\end{aligned} \tag{2.61}$$

For $\eta_k \gamma_{jk} = 0$, we have

$$\log q(\beta_{jk} | \eta_k \gamma_{jk} = 0) = -\frac{1}{2\sigma_{\beta_j}^2} \beta_{jk}^2 + \text{const}, \tag{2.62}$$

which implies that $q(\beta_{jk} | \eta_k \gamma_{jk} = 0) \sim \mathcal{N}(0, \sigma_{\beta_j}^2)$. Thus, the posterior is exactly the same as the prior if this variable is irrelevant in either one of the two levels ($\eta_k \gamma_{jk} = 0$). Therefore we have

$$q(\eta, \gamma, \beta) = \prod_k^K \left(\pi_k^{\eta_k} (1 - \pi_k)^{1 - \eta_k} \prod_j^L \left(\alpha_{jk}^{\gamma_{jk}} (1 - \alpha_{jk})^{1 - \gamma_{jk}} \mathcal{N}(\mu_{jk}, s_{jk}^2)^{\eta_k \gamma_{jk}} \mathcal{N}(0, \sigma_{\beta_j}^2)^{1 - \eta_k \gamma_{jk}} \right) \right), \tag{2.63}$$

where we denote $\pi_k = q(\eta_k)$ and $\alpha_{jk} = q(\gamma_{jk})$.

Now we evaluate the second term of $\mathcal{L}(q)$ in the lower bound:

$$
\begin{aligned}
&-\mathbb{E}_q[\log q(\boldsymbol{\eta},\boldsymbol{\gamma},\boldsymbol{\beta})] \\
=&-\mathbb{E}_q\left[\sum_k^K(\eta_k\log(\pi_k)+(1-\eta_k)\log(1-\pi_k))\right]-\mathbb{E}_q\left[\sum_k^K\sum_j^L\eta_k\gamma_{jk}\log\mathcal{N}(\mu_{jk},s_{jk}^2)\right] \\
&-\mathbb{E}_q\left[\sum_k^K\sum_j^L(1-\eta_k\gamma_{jk})\mathcal{N}(0,\sigma_{\beta_j}^2)+\gamma_{jk}\log(\alpha_{jk})+(1-\gamma_{jk})\log(1-\alpha_{jk})\right] \\
=&-\sum_k^K\mathbb{E}_q\left[(\eta_k\log(\pi_k)+(1-\eta_k)\log(1-\pi_k))\right]-\sum_k^K\sum_j^L\mathbb{E}_q\left[\eta_k\gamma_{jk}\log\mathcal{N}(\mu_{jk},s_{jk}^2)\right] \\
&-\sum_k^K\sum_j^L\mathbb{E}_q\left[(1-\eta_k\gamma_{jk})\mathcal{N}(0,\sigma_{\beta_j}^2)+\gamma_{jk}\log(\alpha_{jk})+(1-\gamma_{jk})\log(1-\alpha_{jk})\right] \\
=&-\sum_k^K\sum_j^L\mathbb{E}_{\eta_k,\gamma_{jk}}\{\mathbb{E}_{\beta|\eta_k=1,\gamma_{jk}=1}[\log\mathcal{N}(\mu_{jk},s_{jk}^2)]+\mathbb{E}_{\beta|\eta_k=1,\gamma_{jk}=0}[\log\mathcal{N}(0,\sigma_{\beta_j}^2)] \qquad (2.64)\\
&+\mathbb{E}_{\beta|\eta_k=0,\gamma_{jk}=1}[\log\mathcal{N}(0,\sigma_{\beta_j}^2)]+\mathbb{E}_{\beta|\eta_k=0,\gamma_{jk}=0}[\log\mathcal{N}(0,\sigma_{\beta_j}^2)]\} \\
&-\sum_k^K\sum_j^L\mathbb{E}_q[\gamma_{jk}\log(\alpha_{jk})+(1-\gamma_{jk})\log(1-\alpha_{jk})]-\sum_k^K\mathbb{E}_q[\eta_k\log(\pi_k)+(1-\eta_k)\log(1-\pi_k)].
\end{aligned}
$$

Note that $-\mathbb{E}_{\beta|\eta_k=1,\gamma_{jk}=1}[\log\mathcal{N}(\mu_{jk},s_{jk}^2)]$ is the entropy of Gaussian, so we have
$-\mathbb{E}_{\beta|\eta_k=1,\gamma_{jk}=1}[\log\mathcal{N}(\mu_{jk},s_{jk}^2)]=\frac{1}{2}\log(s_{jk}^2)+\frac{1}{2}(1+\log(2\pi))$, similarly,
$-\mathbb{E}_{\beta|\eta_k=1,\gamma_{jk}=0}[\log\mathcal{N}(\mu_{jk},s_{jk}^2)]=\frac{1}{2}\log(\sigma_{\beta_j}^2)+\frac{1}{2}(1+\log(2\pi))$ and so on. Consequently,

$$
\begin{aligned}
&-\mathbb{E}_q[\log q(\boldsymbol{\eta},\boldsymbol{\gamma},\boldsymbol{\beta})] \\
=&\sum_k^K\sum_j^L\{[\frac{1}{2}\log(s_{jk}^2)+\frac{1}{2}(1+\log(2\pi))]\pi_k\alpha_{jk}+[\frac{1}{2}\log(\sigma_{\beta_j}^2)+\frac{1}{2}(1+\log(2\pi))](1-\pi_k\alpha_{jk}) \\
&-\alpha_{jk}\log(\alpha_{jk})-(1-\alpha_{jk})\log(1-\alpha_{jk})\}-\sum_k^K\{\pi_k\log(\pi_k)+(1-\pi_k)\log(1-\pi_k)\} \qquad (2.65)\\
=&\sum_k^K\sum_j^L\frac{1}{2}\pi_k\alpha_{jk}(\log s_{jk}^2-\log\sigma_{\beta_j}^2)+\frac{K}{2}\sum_j^L\log(\sigma_{\beta_j}^2)+\frac{p}{2}+\frac{p}{2}\log(2\pi) \\
&-\sum_k^K\sum_j^L[\alpha_{jk}\log(\alpha_{jk})+(1-\alpha_{jk})\log(1-\alpha_{jk})]-\sum_k^K[\pi_k\log(\pi_k)+(1-\pi_k)\log(1-\pi_k)].
\end{aligned}
$$

Combine Equation (2.65) and Equation (2.59), we can find the lower bound:

$$
\begin{aligned}
&\mathbb{E}_q[\log \Pr(\mathbf{y}_j, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta} | \mathbf{X}; \theta)] - \mathbb{E}_q[\log q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta})] \\
=&\sum_{j=1}^{L} \Bigg\{ -\frac{N_j}{2}\log(2\pi\sigma_{e_j}^2) - \frac{\mathbf{y}_j^T\mathbf{y}_j}{2\sigma_{e_j}^2} - \frac{(\mathbf{Z}_j\boldsymbol{\omega}_j)^T(\mathbf{Z}_j\boldsymbol{\omega}_j)}{2\sigma_{e_j}^2} \\
&+ \frac{\sum_k^K \mathbb{E}_q\left[\eta_k\gamma_{jk}\beta_{jk}\right]\mathbf{x}_{jk}^T\mathbf{y}_j}{\sigma_{e_j}^2} + \frac{\mathbf{y}_j^T(\mathbf{Z}_j\boldsymbol{\omega}_j)}{\sigma_{e_j}^2} - \frac{\sum_k^K \mathbb{E}_q\left[\eta_k\gamma_{jk}\beta_{jk}\right]\mathbf{x}_{jk}^T(\mathbf{Z}_j\boldsymbol{\omega}_j)}{\sigma_{e_j}^2} \\
&- \frac{1}{2\sigma_{e_j}^2}\sum_k^K \left(\mathbb{E}_q\left[(\eta_k\gamma_{jk}\beta_{jk})^2\right]\mathbf{x}_{jk}^T\mathbf{x}_{jk}\right) \\
&- \frac{1}{2\sigma_{e_j}^2}\left(\sum_k^K\sum_{k'\neq k}^K \mathbb{E}_q\left[\eta_{k'}\gamma_{jk'}\beta_{jk'}\right]\mathbb{E}_q\left[\eta_k\gamma_{jk}\beta_{jk}\right]\mathbf{x}_{jk'}^T\mathbf{x}_{jk}\right)\Bigg\} \\
&- \frac{K}{2}\sum_j^L \log(2\pi\sigma_{\beta_j}^2) - \sum_{k=1}^K \frac{\sum_{j=1}^L \mathbb{E}_q\left[\beta_{jk}^2\right]}{2\sigma_{\beta_j}^2} \\
&+ \log(\alpha)\sum_{k=1}^K\sum_{j=1}^L \mathbb{E}_q\left[\gamma_{jk}\right] + \log(1-\alpha)\sum_{k=1}^K\sum_{j=1}^L \mathbb{E}_q\left[1-\gamma_{jk}\right] \\
&+ \log(\pi)\sum_{k=1}^K \mathbb{E}_q\left[\eta_k\right] + \log(1-\pi)\sum_{k=1}^K \mathbb{E}_q\left[1-\eta_k\right] \\
&+ \sum_k^K\sum_j^L \frac{1}{2}\pi_k\alpha_{jk}(\log s_{jk}^2 - \log\sigma_{\beta_j}^2) + \frac{K}{2}\sum_j^L \log(\sigma_{\beta_j}^2) + \frac{p}{2} + \frac{p}{2}\log(2\pi) \\
&- \sum_k^K\sum_j^L [\alpha_{jk}\log(\alpha_{jk})] - \sum_k^K\sum_j^L [(1-\alpha_{jk})\log(1-\alpha_{jk})] \\
&- \sum_k^K [\pi_k\log(\pi_k)] - \sum_k^K [(1-\pi_k)\log(1-\pi_k)].
\end{aligned}
\tag{2.66}
$$

Again we can show with the same technique in Section 1.1 that that $\mathbb{E}_q\left[\eta_k\gamma_{jk}\beta_{jk}\right] = \pi_k\alpha_{jk}\mu_{jk}$, $\mathbb{E}_q\left[(\eta_k\gamma_{jk}\beta_{jk})^2\right] = \pi_k\alpha_{jk}(s_{jk}^2 + \mu_{jk}^2)$, $\mathbb{E}_q\left[\beta_{jk}^2\right] = \pi_k\alpha_{jk}(s_{jk}^2 + \mu_{jk}^2) + (1 - \pi_k\alpha_{jk})\sigma_{\beta_j}^2$, $\mathbb{E}_q\left[\eta_k\right] = \pi_k$, $\mathbb{E}_q\left[\gamma_{jk}\right] = \alpha_{jk}$. We plug in the expectations, then Equation (2.66) becomes

$$
\begin{aligned}
&\mathbb{E}_q[\log \Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta} | \mathbf{X}; \theta)] - \mathbb{E}_q[\log q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta})] \\
=&\sum_{j=1}^{L} \Bigg\{ -\frac{N_j}{2}\log(2\pi\sigma_{e_j}^2) - \frac{||\mathbf{y}_j - \mathbf{Z}_j\boldsymbol{\omega}_j - \sum_k^K \pi_k\alpha_{jk}\mu_{jk}\mathbf{x}_{jk}||^2}{2\sigma_{e_j}^2} \\
&- \frac{1}{2\sigma_{e_j}^2}\sum_k^K \underbrace{[\pi_k\alpha_{jk}(s_{jk}^2 + \mu_{jk}^2) - (\pi_k\alpha_{jk}\mu_{jk})^2]}_{\mathrm{Var}[\eta_k\gamma_{jk}\beta_{jk}]} \mathbf{x}_{jk}^T\mathbf{x}_{jk} \Bigg\} \\
&- \frac{K}{2}\sum_{j=1}^{L}\log(2\pi\sigma_{\beta_j}^2) - \frac{1}{2\sigma_{\beta_j}^2}\sum_{k=1}^{K}\sum_{j=1}^{L}[\pi_k\alpha_{jk}(s_{jk}^2 + \mu_{jk}^2) + (1 - \pi_k\alpha_{jk})\sigma_{\beta_j}^2] \\
&+ \sum_{k=1}^{K}\sum_{j=1}^{L}\alpha_{jk}\log(\frac{\alpha}{\alpha_{jk}}) + \sum_{k=1}^{K}\sum_{j=1}^{L}(1 - \alpha_{jk})\log(\frac{1-\alpha}{1-\alpha_{jk}}) \\
&+ \sum_{k=1}^{K}\pi_k\log(\frac{\pi}{\pi_k}) + \sum_{k=1}^{K}(1 - \pi_k)\log(\frac{1-\pi}{1-\pi_k}) \\
&+ \sum_{k}^{K}\sum_{j}^{L}\frac{1}{2}\pi_k\alpha_{jk}\log(\frac{s_{jk}^2}{\sigma_{\beta_j}^2}) + \frac{K}{2}\sum_{j}^{L}\log(\sigma_{\beta_j}^2) + \frac{p}{2} + \frac{p}{2}\log(2\pi).
\end{aligned}
\tag{2.67}
$$

To get π_k and α_{jk}, we let

$$
\frac{\partial \mathbb{E}_q[\log \Pr(\mathbf{y}_j, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta} | \mathbf{X}; \theta)] - \mathbb{E}_q[\log q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta})]}{\partial \pi_k} = 0,
$$

$$
\frac{\partial \mathbb{E}_q[\log \Pr(\mathbf{y}_j, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta} | \mathbf{X}; \theta)] - \mathbb{E}_q[\log q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta})]}{\partial \alpha_{jk}} = 0,
$$

which gives us

$$
\begin{aligned}
\pi_k &= \frac{1}{1 + \exp(-u_k)}, \\
\text{where } u_k =& \log\frac{\pi}{1-\pi} + \frac{1}{2}\sum_j^L \alpha_{jk}\left(\log\frac{s_{jk}^2}{\sigma_{\beta_j}^2} + \frac{\mu_{jk}^2}{s_{jk}^2}\right); \\
\text{and } \alpha_{jk} &= \frac{1}{1 + \exp(-v_{jk})}, \\
\text{where } v_{jk} =& \log\frac{\alpha}{1-\alpha} + \frac{1}{2}\pi_k\left(\log\frac{s_{jk}^2}{\sigma_{\beta_j}^2} + \frac{\mu_{jk}^2}{s_{jk}^2}\right).
\end{aligned}
\tag{2.68}
$$

The derivation is as follow:

$$
\begin{aligned}
u_k =& \log\frac{\pi}{1-\pi} + \frac{1}{2}\sum_j^L \alpha_{jk}\log\frac{s_{jk}^2}{\sigma_{\beta_j}^2} \\
&+ \sum_j^L \frac{1}{\sigma_{e_j}^2}\left\{\alpha_{jk}\mu_{jk}\mathbf{x}_{jk}^T\mathbf{y}_j - \alpha_{jk}\mu_{jk}\mathbf{x}_{jk}^T(\mathbf{Z}_j\boldsymbol{\omega}_j) - \sum_{k'\neq k}^K \pi_{k'}\alpha_{jk'}\mu_{jk'}\alpha_{jk}\mu_{jk}\mathbf{x}_{jk'}^T\mathbf{x}_{jk}\right\} \\
&- \frac{1}{2\sigma_{e_j}^2}\sum_j^L \alpha_{jk}(s_{jk}^2+\mu_{jk}^2)\mathbf{x}_{jk}^T\mathbf{x}_{jk} - \frac{1}{2\sigma_{\beta_j}}\sum_j^L \alpha_{jk}(s_{jk}^2+\mu_{jk}^2) + \frac{1}{2}\sum_j^L \alpha_{jk} \\
=& \log\frac{\pi}{1-\pi} + \frac{1}{2}\sum_j^L \alpha_{jk}\log\frac{s_{jk}^2}{\sigma_{\beta_j}^2} \\
&+ \sum_j^K \alpha_{jk}\mu_{jk}\underbrace{\left(\frac{\mathbf{x}_{jk}^T(\mathbf{y}_j - \mathbf{Z}_j\boldsymbol{\omega}_j) - \sum_{k'\neq k}^K \pi_{k'}\alpha_{jk'}\mu_{jk'}\mathbf{x}_{jk'}^T\mathbf{x}_{jk}}{\sigma_{e_j}^2}\right)}_{\mu_{jk}/s_{jk}^2} \\
&- \frac{1}{2}\sum_j^L \alpha_{jk}(s_{jk}^2+\mu_{jk}^2)\underbrace{\left(\frac{\mathbf{x}_{jk}^T\mathbf{x}_{jk}}{\sigma_{e_j}^2} + \frac{1}{\sigma_{\beta_j}^2}\right)}_{1/s_{jk}^2} + \frac{1}{2}\sum_j^L \alpha_{jk} \\
=& \log\frac{\pi}{1-\pi} + \frac{1}{2}\sum_j^L \alpha_{jk}\log\frac{s_{jk}^2}{\sigma_{\beta_j}^2} + \sum_j^L \frac{\alpha_{jk}\mu_{jk}^2}{s_{jk}^2} - \sum_j^L \frac{\alpha_{jk}\mu_{jk}^2}{2s_{jk}^2} \\
=& \log\frac{\pi}{1-\pi} + \frac{1}{2}\sum_j^L \alpha_{jk}\left(\log\frac{s_{jk}^2}{\sigma_{\beta_j}^2} + \frac{\mu_{jk}^2}{s_{jk}^2}\right)
\end{aligned}
\tag{2.69}
$$

where we have used Equation (2.61). Similarly, we can derive v_{jk}.

M-step

At M-step, we update the parameters $\{\sigma_{e_j}^2, \sigma_{\beta_j}^2, \pi, \alpha, \boldsymbol{\omega}_j\}$. First we consider $\sigma_{e_j}^2$, by setting $\frac{\partial \mathbb{E}_q[\log\Pr(\mathbf{y},\boldsymbol{\eta},\boldsymbol{\gamma},\boldsymbol{\beta}|\mathbf{X},\mathbf{Z};\theta)]}{\partial\sigma_{e_j}^2} = 0$, we have

$$
\begin{aligned}
\sigma_{e_j}^2 =& \frac{||\mathbf{y}_j - \mathbf{Z}_j\boldsymbol{\omega}_j - \sum_k^K \pi_k\alpha_{jk}\mu_{jk}\mathbf{x}_{jk}||^2}{N_j} \\
&+ \frac{\sum_k^K[\pi_k\alpha_{jk}(s_{jk}^2+\mu_{jk}^2) - (\pi_k\alpha_{jk}\mu_{jk})^2]\mathbf{x}_{jk}^T\mathbf{x}_{jk}}{N_j}.
\end{aligned}
\tag{2.70}
$$

For $\sigma^2_{\beta_j}$, set $\frac{\partial \mathcal{L}(q)}{\partial \sigma^2_{\beta_j}} = 0$, we have

$$\sigma^2_{\beta_j} = \frac{\sum_k^K \pi_k \alpha_{jk}(s^2_{jk} + \mu^2_{jk})}{\sum_k^K \pi_k \alpha_{jk}}. \tag{2.71}$$

Accordingly,

$$\alpha = \frac{1}{p} \sum_k^K \sum_j^L \alpha_{jk}, \tag{2.72}$$

$$\pi = \frac{1}{K} \sum_k^K \pi_k. \tag{2.73}$$

$$\boldsymbol{\omega}_j = (\mathbf{Z}_j^T \mathbf{Z}_j)^{-1} \mathbf{Z}_j^T (\mathbf{y}_j - \sum_k^K \pi_k \alpha_{jk} \mu_{jk} \mathbf{x}_{jk}). \tag{2.74}$$

Chapter 3

Numerical Examples

In this section, we gauged the performance of BIVAS in comparison with alternative methods using both simulation and real data analysis. In the spirit of reproducibility, all the simulation codes are made publicly availible at https://github.com/mxcai/simbivas.

3.1 Simulation study

For group BIVAS, we compared it with varbvs [Carbonetto et al., 2012], cMCP [Breheny and Huang, 2009], and GEL [Breheny, 2015]. The simulation data sets were generated as follows. The design matrix $\mathbf{X}$ was generated from normal distribution with autoregressive correlation $\rho^{|j-j'|}$ between column j and j'. 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. Next, we generated coefficients with different sparsity proportion at the group and individual levels: $(\pi, \alpha) \in \{(0.05, 0.8),\ (0.1, 0.4),\ (0.2, 0.2),\ (0.4, 0.1),\ (0.8, 005)\}$. Note that the total sparsity was fixed at $\alpha \cdot \pi = 0.04$ for different combinations of π and α. Finally, we controled the signal-to-noise ratio (SNR) at SNR $= var(\mathbf{X}\beta)/\sigma_e^2 \in \{0.5,\ 1,\ 2\}$. For all the above settings, we had $n = 1,000$, $p = 5,000$, $K = 250$ with 20 variables in each group.

As cMCP and GEL did not provide FDR estimates for variable selection, we first compared BIVAS with varbvs. Figure 3.1 shows the performance of FDR control

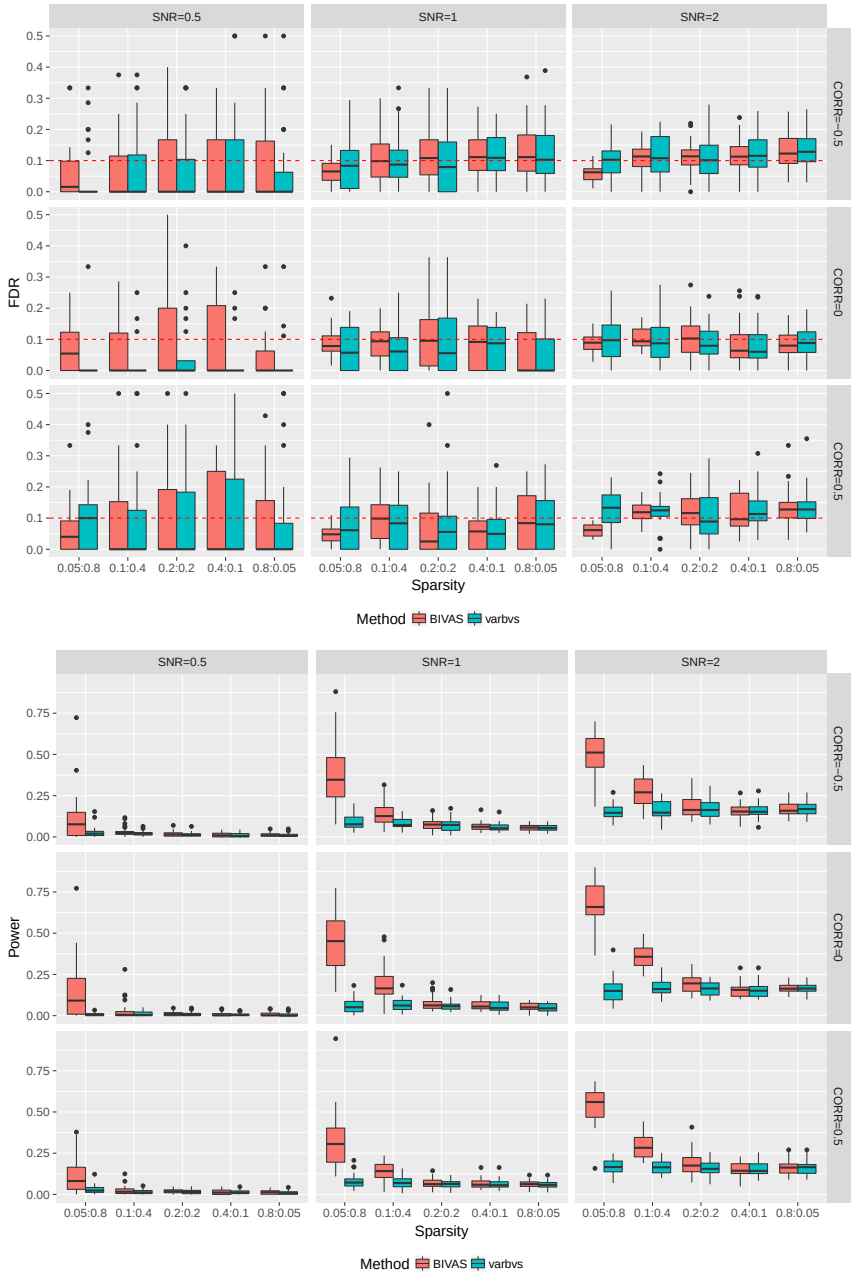


Figure 3.1: Comparison of BIVAS and varbvs for individual variable selection.

and statistical power for individual variable selection obtained by BIVAS and varbvs. When the signal is weak, both methods are underpowered. However, BIVAS gains more power as the group sparsity dominates and further enlarges the gap as signal increases. As ρ moves away from zero, empirical FDRs of both methods are slightly inflated.

Figures 3.2$\sim$ 3.5 show the comparison of BIVAS with varbvs, cMCP and GEL in terms of bi-level variable selection, estimation accuracy and computational efficiency.

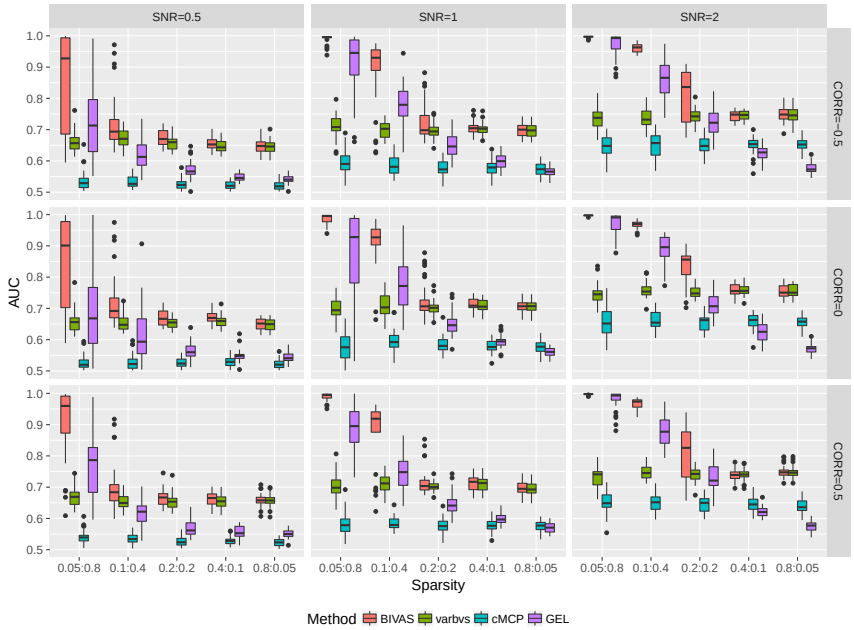


Figure 3.2: Comparisons of BIVAS, varbvs, cMCP and GEL (coupling parameter $\tau = 1/3$): AUC for individual variable selection.

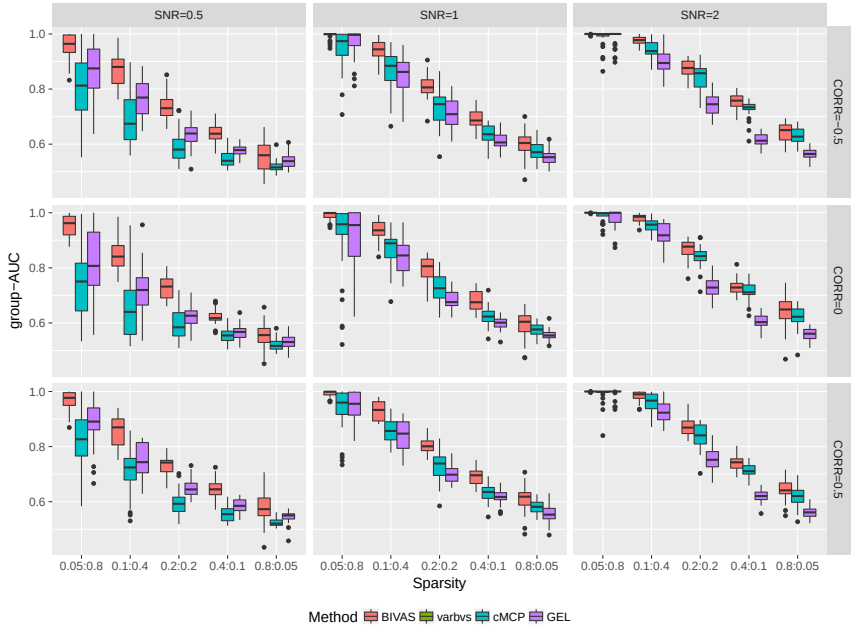


Figure 3.3: Comparisons of BIVAS, varbvs, cMCP and GEL (coupling parameter $\tau = 1/3$): AUC for group selection.

As varbvs only selects individual variables, we treat it as a base line for comparisons of BIVAS with other two alternatives. In the bottom left penal, estimation errors of all the three methods decrease steadily as signal increases when the sparsity-in-group dominates. BIVAS has similar performance with cMCP and GEL when signal is

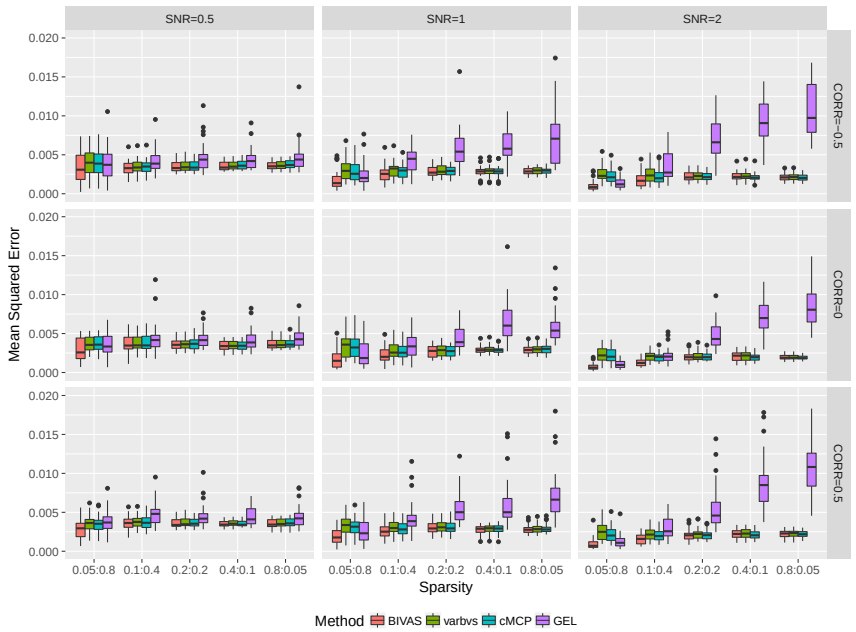


Figure 3.4: Comparisons of BIVAS, varbvs, cMCP and GEL (coupling parameter $\tau = 1/3$): Mean squared error (MSE) of coefficient estimates.

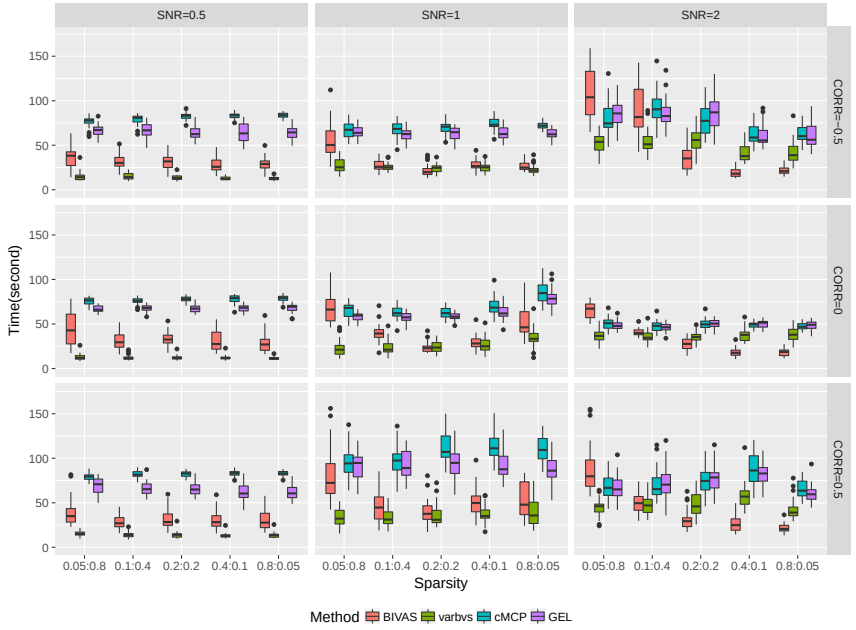


Figure 3.5: Comparisons of BIVAS, varbvs, cMCP and GEL (coupling parameter $\tau = 1/3$): Computational time.

moderate (SNR $= 0.5$) but outperforms them when signal is relatively large (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 mea-

sure of area under the receiver operating characteristic (ROC) curve (AUC) both at the group and individual levels. Figure 3.2 shows that the performance of variable selection for BIVAS is comparable with GEL when signal is strong. When the signal is weak (SNR = 0.5), the AUC of BIVAS is much larger than that of GEL. Moreover, as the 'bulk' of sparsity moves to individual variable level, BIVAS converges to varbvs while GEL becomes even worse than varbvs. This pattern is consistent with that we observe in the measurement of estimation error in Figure 3.4. In all settings we considered, the perfomance of cMCP is poor. Figure 3.3 shows the performance of variable selection at the group level (group-AUC). The pattern of group-AUC is similar to the individual level AUC. Figure 3.5 illustrates the computational efficiency of the four methods. With multi-thread 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 considered BSGS-SS [Xu et al., 2015]; we set $n = 200$, $p = 1,000$, $K = 100$ with 10 variables in each group and $\rho = 0.5$, SNR = 1. As illustrated in Figure 3.6, BIVAS achieves almost the same estimation accuracy as BSGS-SS but uses only around 1% of its computational time.

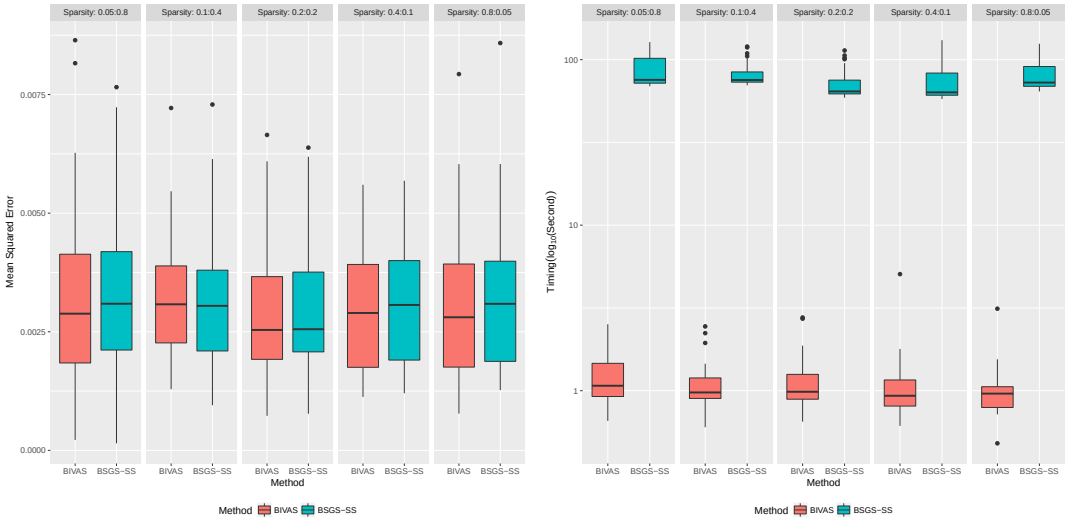


Figure 3.6: Comparison of BIVAS and BSGS-SS. Left: Mean Squared Error of coefficient estimates. Right: Time.

For multi-task BIVAS, we compared with varbvs and Lasso that are applied sep-

arately to each task. We simulated $L = 3$ tasks with sample sizes $n_1 = 600,\ n_2 = 500,\ n_3 = 400$. Number of variables $K = 2,000$ 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 3.7. As one can observe, BIVAS outperforms varbvs and Lasso when the group sparsity is

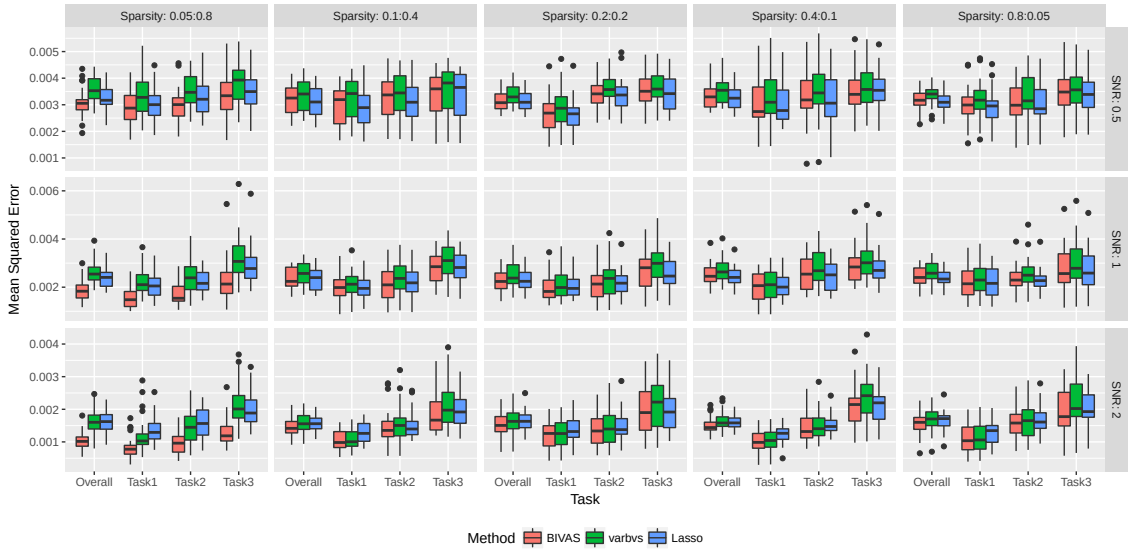


Figure 3.7: Comparison of BIVAS, varbvs, Ridge and Lasso in multi-task learning.

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

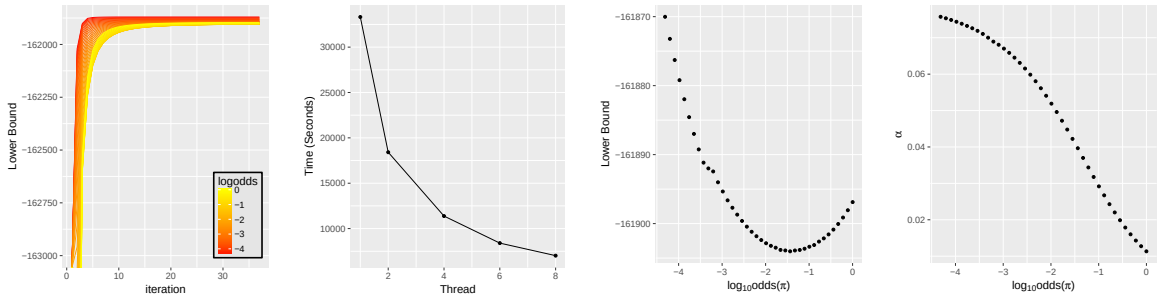
To examine the performance of BIVAS in large scale data, we provide three real examples: we first apply the regression model to the GWAS data from the Wellcome Trust Case Control Consortium (WTCCC) [Consortium et al., 2007] and the Northern Finland Birth Cohort (NFBC) [Sabatti et al., 2009]; then we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

3.2.1 GWAS data

In the GWAS data sets, 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 accessed by the database of Genotypes and Phenotypes (dbGaP) at `https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000276.v2.p1`. This data set was composed of $5,123$ individuals and $319,147$ SNPs. In our analysis, the SNPs were first annotated with their corresponding gene region using ANNOVAR [Wang et al., 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 3.8 (a) shows the convergence of each EM procedure for BIVAS. One can observe that the EM algorithm converges faster for smaller values of π, suggesting the evidence of group sparsity. Computational times for different numbers of threads are presented in Figure 3.8 (b). When $h = 40$, BIVAS took around 3.2 hours to converge using 4 threads and only took 1.6 hours using 8 threads, which indicates that the developed algorithm achieved almost perfect efficiency in parallelization. Estimates of lower bound and parameter α are shown in Figure 3.8 (c) and (d), suggesting the effectiveness of leveraging group structure using BIVAS. After the convergence, we identified the SNPs and genes based on $fdr < 0.05$. 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* containsed 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 3.9).

In the second example, we analyzed Rheumatoid Arthritis (RA) and Type 1 Diabetes (T1D) in the WTCCC data. These data sets were from European Genome-phenome Archive (EGA) websites `http://www.ebi.ac.uk/ega/studies/EGAS00000000011` and `http://www.ebi.ac.uk/ega/studies/EGAS00000000014`. After quality con-



(a) (b) (c) (d)

Figure 3.8: 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) $\hat{\alpha}$ for the 40 settings after convergence.

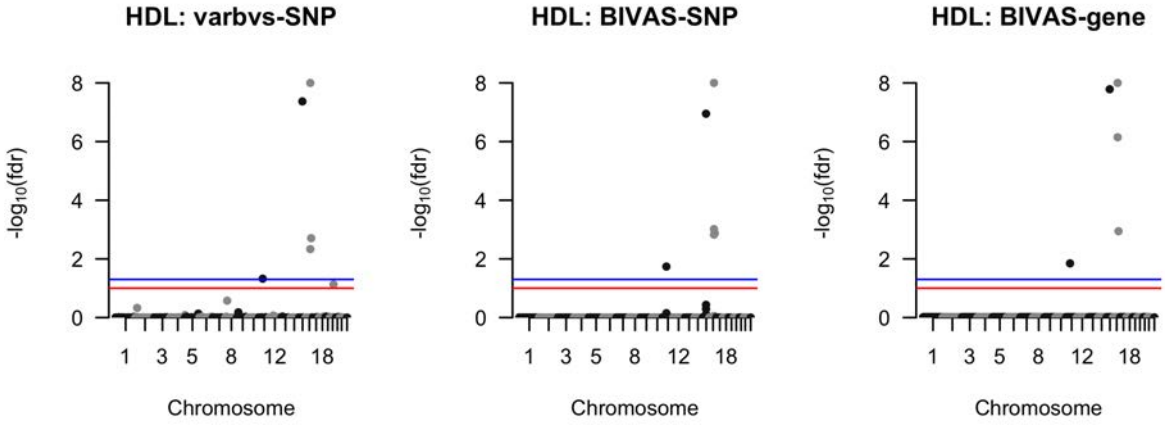


Figure 3.9: Manhattan plots of High-Density Lipoprotein (HDL). Red line represents $fdr = 0.1$ and blue line represents $fdr = 0.05$.

trol, we had 4,494 individuals and 307,089 SNPs for RA, and 4,986 individuals and 307,357 SNPs for T1D. The SNPs were then matched with corresponding genes using HapMap3 as reference, leading to 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 3.10. 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.

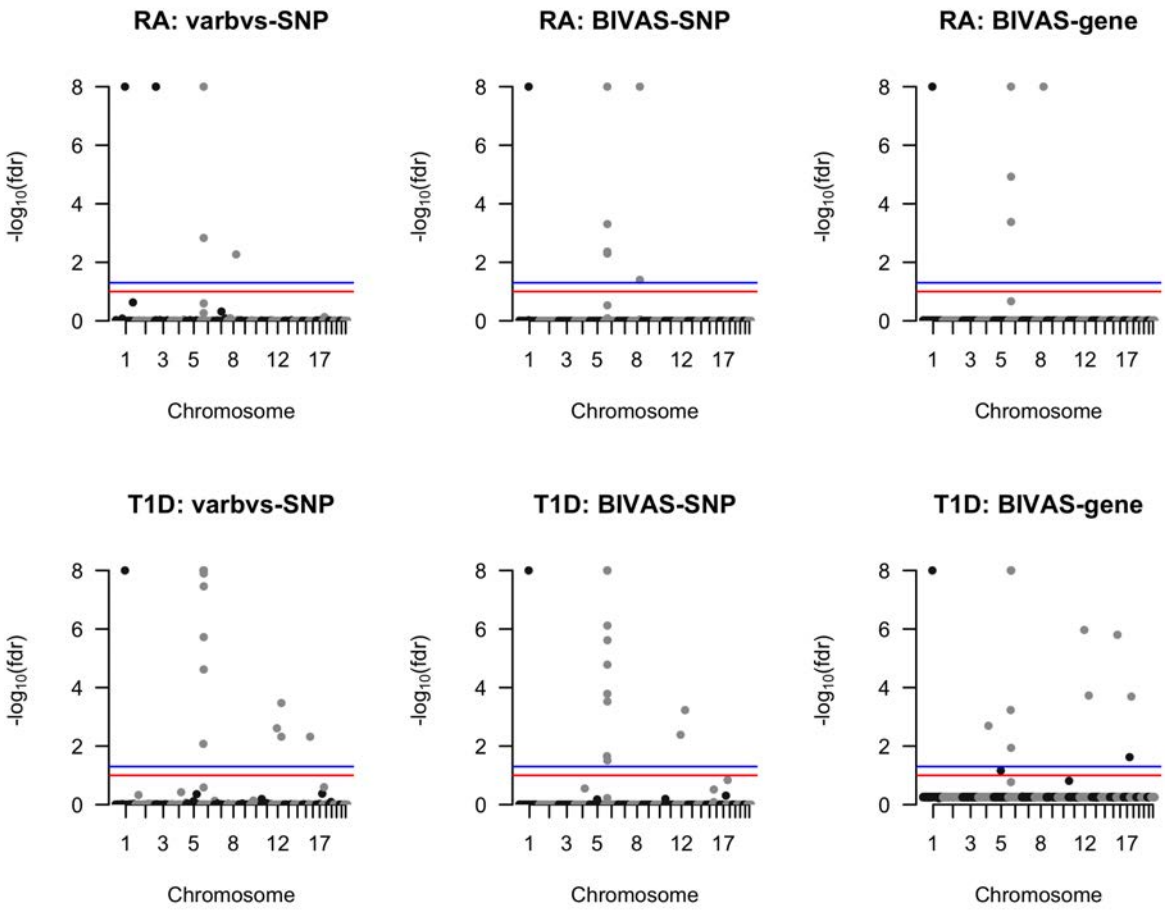


Figure 3.10: Manhattan plots of Rheumatoid Arthritis (RA) and Type 1 Diabetes (T1D). Red line represents $fdr = 0.1$ and blue line represents $fdr = 0.05$.

3.2.2 IMDB movie data

In the third example, we analyzed IMDb dataset [Maas et al., 2011] based on multi-task BIVAS. The IMDb data set was publicly availible at `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 6 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 $3,354$ for drama, $2,235$ for comedy, $1,175$ for horror, 346 for action, 258 for thriller and 139 for romance. We compared BIVAS against Ridge, Lasso and varbvs.

Table 3.1 shows the testing errors of the four methods. For the categories of Horror, Action, Thriller and Romance, BIVAS has better performance than the other 3 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.

	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	**6.13**	**6.67**	7.20	8.65	9.27	6.77
varbvs	7.14	6.20	6.90	8.94	8.37	11.48	6.91
BIVAS	**6.66**	6.32	7.01	**6.76**	**6.89**	**7.44**	**5.39**

Table 3.1: IMDb testing error.

The words selected by BIVAS and varbvs are presented in Figure 3.11 and Figure 3.12 using 'wordcloud' package in R. The words in blue and yellow represent the negative and positive effects, respectively. The size of words represents the effect size. As shown in Figure 3.11, 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 3.12, 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 multi-task 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.

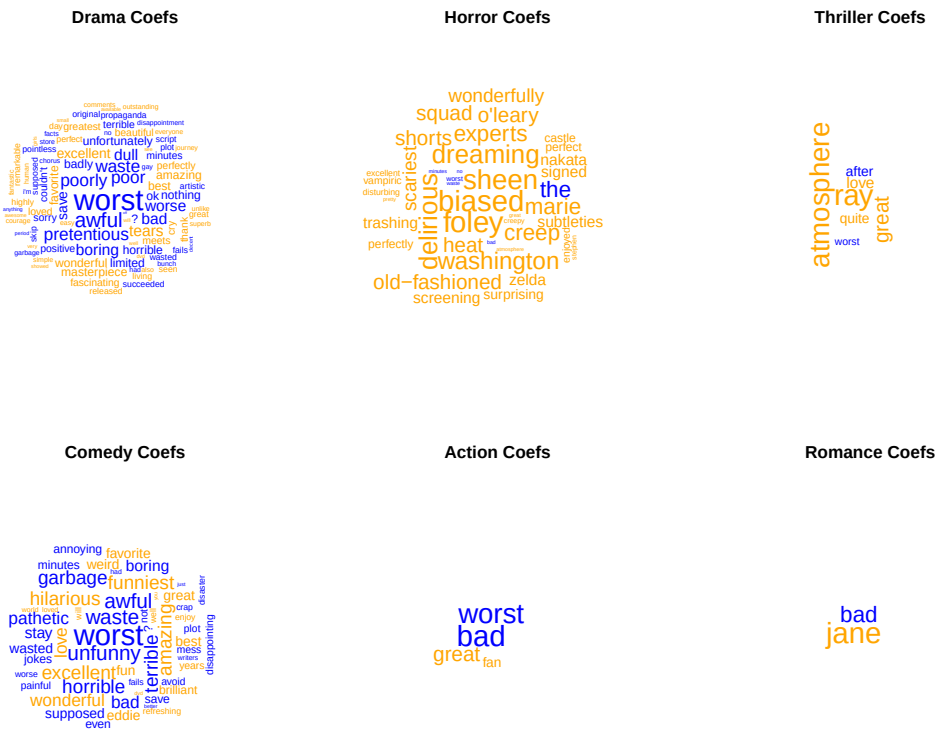


Figure 3.11: IMDb wordcloud generated by varbvs.

Figure 3.12: IMDb wordcloud generated by BIVAS.

Chapter 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 fdr. Despite the convenience, existing Bayesian bi-level variable selection methods are often computationally inefficient and unscalable to large data sets due to the intractable posterior.

In this thesis, we propose a hierarchically factorizable formulation to approximate the posterior distribution, by utilizing 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 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 selction.

Bibliography

K. Bae and B. K. Mallick. Gene selection using a two-level hierarchical bayesian model. *Bioinformatics*, 20(18):3423–3430, 2004.

C. M. Bishop. *Pattern recognition and machine learning*. springer, 2006.

P. Breheny. The group exponential lasso for bi-level variable selection. *Biometrics*, 71(3):731–740, 2015.

P. Breheny and J. Huang. Penalized methods for bi-level variable selection. *Statistics and its interface*, 2(3):369, 2009.

P. Carbonetto, M. Stephens, et al. Scalable variational inference for bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian analysis*, 7(1):73–108, 2012.

R. Caruana. Multitask learning. In *Learning to learn*, pages 95–133. Springer, 1998.

R.-B. Chen, C.-H. Chu, S. Yuan, and Y. N. Wu. Bayesian sparse group selection. *Journal of Computational and Graphical Statistics*, 25(3):665–683, 2016.

W. T. C. C. Consortium et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661, 2007.

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

M. A. Figueiredo. Adaptive sparseness for supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 25(9):1150–1159, 2003.

E. I. George and R. E. McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

E. I. George and R. E. McCulloch. Approaches for bayesian variable selection. *Statistica sinica*, pages 339–373, 1997.

T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

J. Huang, S. Ma, H. Xie, and C.-H. Zhang. A group bridge approach for variable selection. *Biometrika*, 96(2):339–355, 2009.

M. Kyung, J. Gill, M. Ghosh, G. Casella, et al. Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, 5(2):369–411, 2010.

A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 142–150. Association for Computational Linguistics, 2011.

D. Madigan and A. E. Raftery. Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of the American Statistical Association*, 89(428):1535–1546, 1994.

T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032, 1988.

T. Park and G. Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker, M. J. Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

S. Raman, T. J. Fuchs, P. J. Wild, E. Dahl, and V. Roth. The bayesian group-lasso for analyzing contingency tables. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 881–888. ACM, 2009.

C. Sabatti, A.-L. Hartikainen, A. Pouta, S. Ripatti, J. Brodsky, C. G. Jones, N. A. Zaitlen, T. Varilo, M. Kaakinen, U. Sovio, et al. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature genetics*, 41(1):35, 2009.

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

K. Wang, M. Li, and H. Hakonarson. Annovar: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38(16): e164, 2010. doi: 10.1093/nar/gkq603. URL `+http://dx.doi.org/10.1093/nar/gkq603`.

X. Xu, M. Ghosh, et al. Bayesian variable selection and estimation for group lasso. *Bayesian Analysis*, 10(4):909–936, 2015.

J. Yang, S. H. Lee, M. E. Goddard, and P. M. Visscher. Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):76–82, 2011.

M. Yuan and Y. Lin. Efficient empirical bayes variable selection and estimation in linear models. *Journal of the American Statistical Association*, 100(472):1215–1225, 2005.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

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

P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, pages 3468–3497, 2009.

N. Zhou and J. Zhu. Group variable selection via a hierarchical lasso and its oracle property. *Statistics and Its Interface*, 3(4):557–574, 2010.

CURRICULUM VITAE

Academic qualifications of the thesis author, Mr. CAI Mingxuan:

- Received Bachelor of Science in Statistics & Operational Research (Honor) from Hong Kong Baptist University University, June 2016.

May 2018