Supplementary: BIVAS: A scalable Bayesian method for bi-level variable selection with applications

Mingxuan Cai[1], Mingwei Dai[2,4], Jingsi Ming[1], Heng Peng[1*], Jin Liu[3*] and Can Yang[4*]

[1]Department of Mathematics, Hong Kong Baptist University, Hong Kong

[2]School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, China

[3]Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore

[4]Department of Mathematics, The Hong Kong University of Science and Technology, Hong Kong

Contents

*Correspondence should be addressed to Can Yang (macyang@ust.hk), Heng Peng (hpeng@hkbu.edu.hk) and Jin Liu (jin.liu@duke-nus.edu.sg)

1 Variational EM Algorithm: Regression with BIVAS

1.1 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

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

By integrating out the latent variables $\{\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}\}$, the logarithm of the marginal likelihood is given as

$$\begin{aligned} \log\ p(\mathbf{y} | \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}) &= \log \sum_{\boldsymbol{\gamma}} \sum_{\boldsymbol{\eta}} \int_{\boldsymbol{\beta}} \Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta} | \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}) d\boldsymbol{\beta} \\ &\geq \sum_{\boldsymbol{\gamma}} \sum_{\boldsymbol{\eta}} \int_{\boldsymbol{\beta}} q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}) \log \frac{\Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta} | \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta})}{q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta})} d\boldsymbol{\beta} \\ &= \mathbb{E}_q [\log \Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta} | \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}) - \log q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta})] \\ &\equiv \mathcal{L}(q), \end{aligned} \tag{2}$$

where we have adopted Jensen's inequality to obtain the lower bound $\mathcal{L}(q)$. Instead of working with the marginal likelihood directly, BIVAS iteratively maximize $\mathcal{L}(q)$ using the vairational EM algorithm. As illustrated in the main text, we use the following hierarchically factorized distribution to approximate the true posterior:

$$q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}) = \prod_k^K \left(q(\eta_k) \prod_j^{l_k} (q(\beta_{jk} | \eta_k, \gamma_{jk}) q(\gamma_{jk})) \right), \tag{3}$$

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:

$$\mathcal{L}(q) = \mathbb{E}_{q(\eta)} \left[\mathbb{E}_{q(\gamma, \beta | \eta)} \left[\log \Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}) - \log q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}) \right] \right]. \tag{4}$$

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})\left[\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'}\right]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] \qquad (5) \\
&+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}
$$

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 (5):

$$
\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)\left[\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}\right]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)}\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}_{-jk},\gamma_{jk}=1,\boldsymbol{\beta})\right] - \log q(\beta_{jk},\gamma_{jk}=1|\eta_k=1)\right] \qquad (6) \\
&+q(\gamma_{jk}=0)\mathbb{E}_{q(\beta_{jk}|\eta_k=1,\gamma_{jk}=0)}\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}_{-jk},\gamma_{jk}=0,\boldsymbol{\beta})\right] - \log q(\beta_{jk},\gamma_{jk}=0|\eta_k=1)\right].
\end{aligned}
$$

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 (6) that:

$$
\begin{aligned}
&\mathbb{E}_{q(\beta_{jk}|\eta_k=1,\gamma_{jk}=1)}\left[\mathbb{E}_{j'\neq j|k}\left[\mathbb{E}_{k'\neq k}\log\Pr(\mathbf{y},\boldsymbol{\eta}_{-k},\eta_k=1,\boldsymbol{\gamma}_{-jk},\gamma_{jk}=1,\boldsymbol{\beta})\right] - \log q(\beta_{jk},\gamma_{jk}=1|\eta_k=1)\right] \\
=&\mathbb{E}_{q(\beta_{jk}|\eta_k=1,\gamma_{jk}=1)}\left[\mathbb{E}_{j'\neq j|k}\left[\mathbb{E}_{k'\neq k}\log\Pr(\mathbf{y},\boldsymbol{\eta}_{-k},\eta_k=1,\boldsymbol{\gamma}_{-jk},\gamma_{jk}=1,\boldsymbol{\beta})\right] - \log q(\beta_{jk}|\eta_k=1,\gamma_{jk}=1)\right] + \text{const},
\end{aligned} \qquad (7)
$$

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

$$\log q^*(\beta_{jk}|\eta_k=1,\gamma_{jk}=1)=\mathbb{E}_{j'\neq j|k}\left[\mathbb{E}_{k'\neq k}\log\Pr(\mathbf{y},\boldsymbol{\eta}_{-k},\eta_k=1,\boldsymbol{\gamma}_{-jk},\gamma_{jk}=1,\boldsymbol{\beta})\right]. \tag{8}$$

Here we only derive the case when $\eta_k=\gamma_{jk}=1$, other cases can be obtained following the same procedure. Since both η_k and γ_{jk} are from Bernoulli distribution, with the expression in equation (8), 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

$$\begin{aligned}
\log\Pr(\mathbf{y},\boldsymbol{\eta},\boldsymbol{\gamma},\boldsymbol{\beta}|\mathbf{X},\mathbf{Z};\boldsymbol{\theta})=&-\frac{n}{2}\log(2\pi\sigma_e^2)-\frac{\mathbf{y}^T\mathbf{y}}{2\sigma_e^2}-\frac{(\mathbf{Z}\boldsymbol{\omega})^T(\mathbf{Z}\boldsymbol{\omega})}{2\sigma_e^2}\\
&+\frac{\sum_k^K\sum_j^{l_k}\eta_k\gamma_{jk}\beta_{jk}\mathbf{x}_{jk}^T\mathbf{y}}{\sigma_e^2}+\frac{\mathbf{y}^T(\mathbf{Z}\boldsymbol{\omega})}{\sigma_e^2}-\frac{\sum_k^K\sum_j^{l_k}\eta_k\gamma_{jk}\beta_{jk}\mathbf{x}_{jk}^T(\mathbf{Z}\boldsymbol{\omega})}{\sigma_e^2}\\
&-\frac{1}{2\sigma_e^2}\sum_k^K\sum_j^{l_k}\left(\left(\eta_k\gamma_{jk}\beta_{jk}\right)^2\mathbf{x}_{jk}^T\mathbf{x}_{jk}\right)\\
&-\frac{1}{2\sigma_e^2}\left(\sum_k^K\sum_{k'\neq k}^K\sum_j^{j_k}\sum_{j'}^{l_{k'}}\left(\eta_{k'}\gamma_{j'k'}\beta_{j'k'}\right)\left(\eta_k\gamma_{jk}\beta_{jk}\right)\mathbf{x}_{j'k'}^T\mathbf{x}_{jk}\right)\\
&-\frac{1}{2\sigma_e^2}\left(\sum_k^K\sum_j^{l_k}\sum_{j'\neq j}^{l_k}\left(\eta_k\gamma_{j'k}\beta_{j'k}\right)\left(\eta_k\gamma_{jk}\beta_{jk}\right)\mathbf{x}_{j'k}^T\mathbf{x}_{jk}\right)\\
&-\frac{p}{2}\log(2\pi\sigma_\beta^2)-\frac{1}{2\sigma_\beta^2}\sum_{k=1}^K\sum_{j=1}^{l_k}\beta_{jk}^2\\
&+\log(\alpha)\sum_{k=1}^K\sum_{j=1}^{l_k}\gamma_{jk}+\log(1-\alpha)\sum_{k=1}^K\sum_{j=1}^{l_k}(1-\gamma_{jk})\\
&+\log(\pi)\sum_{k=1}^K\eta_k+\log(1-\pi)\sum_{k=1}^K(1-\eta_k).
\end{aligned}\tag{9}$$

To find the optimal form in Equation (8), We then rearrange Equation (9) and only retain the terms regarding jk:

$$
\begin{aligned}
&\log\Pr(\mathbf{y},\boldsymbol{\eta},\boldsymbol{\gamma},\boldsymbol{\beta}|\mathbf{X},\mathbf{Z};\boldsymbol{\theta}) \\
=&-\frac{n}{2}\log(2\pi\sigma_e^2)-\frac{\mathbf{y}^T\mathbf{y}}{2\sigma_e^2} \\
&+\frac{\eta_k\gamma_{jk}\beta_{jk}\mathbf{x}_{jk}^T\mathbf{y}}{\sigma_e^2}-\frac{\eta_k\gamma_{jk}\beta_{jk}\mathbf{x}_{jk}^T(\mathbf{Z}\boldsymbol{\omega})}{\sigma_e^2} \\
&-\frac{1}{2\sigma_e^2}\left((\eta_k\gamma_{jk}\beta_{jk})^2\,\mathbf{x}_{jk}^T\mathbf{x}_{jk}\right) \\
&-\frac{1}{2\sigma_e^2}\left(\sum_{k'\neq k}^{K}\sum_{j'}^{l_{k'}}(\eta_k\gamma_{jk}\beta_{jk})\,(\eta_k\gamma_{j'k'}\beta_{j'k'})\,\mathbf{x}_{jk}^T\mathbf{x}_{j'k'}\right) \\
&-\frac{1}{2\sigma_e^2}\left(\sum_{j'\neq j}^{l_k}(\eta_k\gamma_{jk}\beta_{jk})\,(\eta_k\gamma_{j'k}\beta_{j'k})\,\mathbf{x}_{jk}^T\mathbf{x}_{j'k}\right)-\frac{1}{2\sigma_\beta^2}\beta_{jk}^2 \\
&+\log(\alpha)\gamma_{jk}+\log(1-\alpha)(1-\gamma_{jk}) \\
&+\log(\pi)\eta_k+\log(1-\pi)(1-\eta_{jk}) \\
&+\text{const.}
\end{aligned}
\tag{10}
$$

Now we can derive the $\log\ q(\beta_{jk}|\eta_k=1,\gamma_{jk}=1)$ by taking the expectation in Equation (8). When $\eta_k=\gamma_{jk}=1\Leftrightarrow\eta_k\gamma_{jk}=1$, we have

$$
\begin{aligned}
&\log\ q(\beta_{jk}|\eta_k=1,\gamma_{jk}=1) \\
=&\left(-\frac{1}{2\sigma_e^2}\mathbf{x}_{jk}^T\mathbf{x}_{jk}-\frac{1}{2\sigma_\beta^2}\right)\beta_{jk}^2 \\
&+\left(\frac{\mathbf{x}_{jk}^T(\mathbf{y}-\mathbf{Z}\boldsymbol{\omega})-\sum_{k'\neq k}^{K}\sum_{j'}^{l_{k'}}\mathbb{E}_{k'\neq k}[\eta_{k'}\gamma_{j'k'}\beta_{j'k'}]\mathbf{x}_{jk}^T\mathbf{x}_{j'k'}-\sum_{j'\neq j}^{l_k}\mathbb{E}_{j'\neq j|k}[\gamma_{j'k}\beta_{j'k}]\mathbf{x}_{jk}^T\mathbf{x}_{j'k}}{\sigma_e^2}\right)\beta_{jk} \\
&+\text{const.}
\end{aligned}
\tag{11}
$$

Since Equation (11) 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

$$
\begin{aligned}
s_{jk}^2&=\frac{\sigma_e^2}{\mathbf{x}_{jk}^T\mathbf{x}_{jk}+\frac{\sigma_e^2}{\sigma_\beta^2}} \\
\mu_{jk}&=\frac{\mathbf{x}_{jk}^T(\mathbf{y}-\mathbf{Z}\boldsymbol{\omega})-\sum_{k'\neq k}^{K}\sum_{j'}^{l_{k'}}\mathbb{E}_{k'\neq k}[\eta_{k'}\gamma_{j'k'}\beta_{j'k'}]\mathbf{x}_{jk}^T\mathbf{x}_{j'k'}-\sum_{j'\neq j}^{l_k}\mathbb{E}_{j'\neq j|k}[\gamma_{j'k}\beta_{j'k}]\mathbf{x}_{jk}^T\mathbf{x}_{j'k}}{\mathbf{x}_{jk}^T\mathbf{x}_{jk}+\frac{\sigma_e^2}{\sigma_\beta^2}}.
\end{aligned}
\tag{12}
$$

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

$$
\log\ q(\beta_{jk}|\eta_k\gamma_{jk}=0)=-\frac{1}{2\sigma_\beta^2}\beta_{jk}^2+\text{const},
\tag{13}
$$

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

$$q(\boldsymbol{\eta},\boldsymbol{\gamma},\boldsymbol{\beta}) = \prod_k^K \left(\pi_k^{\eta_k}(1-\pi_k)^{1-\eta_k} \prod_j^{l_k} \left(\alpha_{jk}^{\gamma_{jk}}(1-\alpha_{jk})^{1-\gamma_{jk}} \mathcal{N}(\mu_{jk},s_{jk}^2)^{\eta_k\gamma_{jk}} \mathcal{N}(0,\sigma_\beta)^{1-\eta_k\gamma_{jk}} \right) \right). \quad (14)$$

With this variational posterior probability function, the second term of the lower bound $\mathcal{L}(q)$ in (4) can be derived as:

$$\begin{aligned}
&-\mathbb{E}_q[\log q(\boldsymbol{\eta},\boldsymbol{\gamma},\boldsymbol{\beta})] \\
=&-\mathbb{E}_q\left[\sum_k^K (\eta_k\log(\pi_k) + (1-\eta_k)\log(1-\pi_k))\right] \\
&-\mathbb{E}_q\left[\sum_k^K\sum_j^{l_k} \eta_k\gamma_{jk}\log\mathcal{N}(\mu_{jk},s_{jk}^2) + (1-\eta_k\gamma_{jk})\mathcal{N}(0,\sigma_\beta^2) + \gamma_{jk}\log(\alpha_{jk}) + (1-\gamma_{jk})\log(1-\alpha_{jk})\right] \\
=&-\sum_k^K \mathbb{E}_q\left[(\eta_k\log(\pi_k) + (1-\eta_k)\log(1-\pi_k))\right] \\
&-\sum_k^K\sum_j^{l_k} \mathbb{E}_q\left[\eta_k\gamma_{jk}\log\mathcal{N}(\mu_{jk},s_{jk}^2) + (1-\eta_k\gamma_{jk})\mathcal{N}(0,\sigma_\beta^2) + \gamma_{jk}\log(\alpha_{jk}) + (1-\gamma_{jk})\log(1-\alpha_{jk})\right] \\
=&-\sum_k^K\sum_j^{l_k} \mathbb{E}_{\eta_k,\gamma_{jk}}\{\mathbb{E}_{\beta|\eta_k=1,\gamma_{jk}=1}[\log\mathcal{N}(\mu_{jk},s_{jk}^2)] + \mathbb{E}_{\beta|\eta_k=1,\gamma_{jk}=0}[\log\mathcal{N}(0,\sigma_\beta^2)] \\
&+\mathbb{E}_{\beta|\eta_k=0,\gamma_{jk}=1}[\log\mathcal{N}(0,\sigma_\beta^2)] + \mathbb{E}_{\beta|\eta_k=0,\gamma_{jk}=0}[\log\mathcal{N}(0,\sigma_\beta^2)]\} \\
&-\sum_k^K\sum_j^{l_k} \mathbb{E}_q[\gamma_{jk}\log(\alpha_{jk}) + (1-\gamma_{jk})\log(1-\alpha_{jk})] - \sum_k^K \mathbb{E}_q[\eta_k\log(\pi_k) + (1-\eta_k)\log(1-\pi_k)].
\end{aligned} \quad (15)$$

Note that $-\mathbb{E}_{\beta|\eta_k=1,\gamma_{jk}=1}[\log\mathcal{N}(\mu_{jk},s_{jk}^2)]$ is the entropy of Gaussian distribution, 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,

$$\begin{aligned}
&-\mathbb{E}_{\beta|\eta_k=1,\gamma_{jk}=0}[\log\mathcal{N}(\mu_{jk},s_{jk}^2)] \\
=&-\mathbb{E}_{\beta|\eta_k=0,\gamma_{jk}=1}[\log\mathcal{N}(\mu_{jk},s_{jk}^2)] \\
=&-\mathbb{E}_{\beta|\eta_k=0,\gamma_{jk}=0}[\log\mathcal{N}(\mu_{jk},s_{jk}^2)] \\
=&\frac{1}{2}\log(\sigma_\beta^2) + \frac{1}{2}(1+\log(2\pi)).
\end{aligned}$$

Using these results, Equation (15) can be written as:

$$
\begin{aligned}
&- \mathbb{E}[\log q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta})] \\
=&\sum_{k}^{K}\sum_{j}^{l_k}\{[\frac{1}{2}\log(s_{jk}^2) + \frac{1}{2}(1+\log(2\pi))]\pi_k\alpha_{jk} + [\frac{1}{2}\log(\sigma_\beta^2) + \frac{1}{2}(1+\log(2\pi))](1-\pi_k\alpha_{jk}) \\
&- \alpha_{jk}\log(\alpha_{jk}) - (1-\alpha_{jk})\log(1-\alpha_{jk})\} - \sum_{k}^{K}\{\pi_k\log(\pi_k) + (1-\pi_k)\log(1-\pi_k)\} \\
=&\sum_{k}^{K}\sum_{j}^{l_k}\frac{1}{2}\pi_k\alpha_{jk}(\log s_{jk}^2 - \log\sigma_\beta^2) + \frac{p}{2}\log(\sigma_\beta^2) + \frac{p}{2} + \frac{p}{2}\log(2\pi) \\
&- \sum_{k}^{K}\sum_{j}^{l_k}[\alpha_{jk}\log(\alpha_{jk}) + (1-\alpha_{jk})\log(1-\alpha_{jk})] - \sum_{k}^{K}[\pi_k\log(\pi_k) + (1-\pi_k)\log(1-\pi_k)].
\end{aligned} \tag{16}
$$

Combine (16) with (9), 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[(\eta_k\gamma_{jk}\beta_{jk})^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{17}
$$

where expectations in (17) are derived as follows:

$$\mathbb{E}_q\left[\eta_k\right] = \pi_k, \quad \mathbb{E}_q\left[\gamma_{jk}\right] = \alpha_{jk}, \tag{18}$$

$$\begin{aligned}
\mathbb{E}\left[\eta_k\gamma_{jk}\beta_{jk}\right] &= \sum_{\gamma_{jk}}\sum_{\eta_k}\int_{\beta_{jk}} \eta_k\gamma_{jk}\beta_{jk}q(\beta_{jk}|\eta_k,\gamma_{jk})q(\eta_k)q(\gamma_{jk})d\beta_{jk} \\
&= \pi_k\alpha_{jk}\cdot\mu_{jk} + (1-\pi_k\alpha_{jk})\cdot 0 \\
&= \pi_k\alpha_{jk}\mu_{jk}
\end{aligned} \tag{19}$$

$$\begin{aligned}
\mathbb{E}_q\left[\beta_{jk}^2\right] &= \int_{\beta_{jk}} \beta_{jk}^2 q(\beta_{jk})d\beta_{jk} \\
&= \sum_{\eta_k}\sum_{\gamma_{jk}}\int_{\beta_{jk}} \beta_{jk}^2 q(\beta_{jk}|\eta_k,\gamma_{jk})q(\eta_k)q(\gamma_{jk})d\beta_{jk} \\
&= \int_{\beta_{jk}} \beta_{jk}^2\cdot\left[\pi_k\alpha_{jk}\mathcal{N}(\mu_{jk},s_{jk}^2) + (1-\pi_k\alpha_{jk})\mathcal{N}(0,\sigma_\beta^2)\right]d\beta_{jk} \\
&= \pi_k\alpha_{jk}(s_{jk}^2+\mu_{jk}^2) + (1-\pi_k\alpha_{jk})\sigma_\beta^2
\end{aligned} \tag{20}$$

$$\begin{aligned}
\mathbb{E}_q\left[(\eta_k\gamma_{jk}\beta_{jk})^2\right] &= \sum_{\eta_k}\sum_{\gamma_{jk}}\int_{\beta_{jk}} \eta_k\gamma_{jk}\beta_{jk}^2 q(\beta_{jk}|\eta_k,\gamma_{jk})q(\eta_k)q(\gamma_{jk})d\beta_{jk} \\
&= \pi_k\alpha_{jk}\int_{\beta_{jk}} \beta_{jk}^2\mathcal{N}(\mu_{jk},s_{jk}^2))d\beta_{jk} \\
&= \pi_k\alpha_{jk}(s_{jk}^2+\mu_{jk}^2)
\end{aligned} \tag{21}$$

$$\begin{aligned}
&\mathbb{E}_q\left[\eta_k^2\gamma_{j'k}\beta_{j'k}\gamma_{jk}\beta_{jk}\right] \\
=&\mathbb{E}_q\left[\eta_k\gamma_{j'k}\beta_{j'k}\gamma_{jk}\beta_{jk}\right] \\
=&\sum_{\eta_k}\sum_{\gamma_{jk},\gamma_{j'k}}\int\int \eta_k\gamma_{j'k}\beta_{j'k}\gamma_{jk}\beta_{jk}q(\beta_{jk}|\eta_k,\gamma_{jk})q(\gamma_{jk})q(\beta_{j'k}|\eta_k,\gamma_{j'k})q(\gamma_{j'k})q(\eta_k)d\beta_{jk}d\beta_{j'k} \\
=&\pi_k\alpha_{j'k}\mu_{j'k}\alpha_{jk}\mu_{jk}
\end{aligned} \tag{22}$$

By plugging in the evaluations from Equation (18) to (22), the lower bound $\mathcal{L}(q)$ in Equation (17)

then becomes

$$
\begin{aligned}
&\mathbb{E}_q[\log \Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta} | \mathbf{X}; \boldsymbol{\theta})] - \mathbb{E}_q[\log q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta})] \\
=& -\frac{n}{2}\log(2\pi\sigma_e^2) - \frac{||\mathbf{y} - \mathbf{Z}\boldsymbol{\omega} - \sum_k^K \sum_j^{l_k} \pi_k \alpha_{jk} \mu_{jk} \mathbf{x}_{jk}||^2}{2\sigma_e^2} \\
&- \frac{1}{2\sigma_e^2} \sum_k^K \sum_j^{l_k} \underbrace{[\pi_k \alpha_{jk}(s_{jk}^2 + \mu_{jk}^2) - (\pi_k \alpha_{jk} \mu_{jk})^2]}_{\mathrm{Var}[\eta_k \gamma_{jk} \beta_{jk}]} \mathbf{x}_{jk}^T \mathbf{x}_{jk} \\
&- \frac{1}{2\sigma_e^2} \sum_k^K (\pi_k - \pi_k^2) \left(\sum_j^{l_k} \sum_{j' \neq j}^{l_k} \alpha_{j'k} \mu_{j'k} \alpha_{jk} \mu_{jk} \mathbf{x}_{j'k}^T \mathbf{x}_{jk} \right) \\
&- \frac{p}{2}\log(2\pi\sigma_\beta^2) - \frac{1}{2\sigma_\beta^2} \sum_{k=1}^K \sum_{j=1}^{l_k} [\pi_k \alpha_{jk}(s_{jk}^2 + \mu_{jk}^2) + (1 - \pi_k \alpha_{jk})\sigma_\beta^2] \\
&+ \sum_{k=1}^K \sum_{j=1}^{l_k} \alpha_{jk} \log(\frac{\alpha}{\alpha_{jk}}) + \sum_{k=1}^K \sum_{j=1}^{l_k} (1 - \alpha_{jk}) \log(\frac{1-\alpha}{1-\alpha_{jk}}) \\
&+ \sum_{k=1}^K \pi_k \log(\frac{\pi}{\pi_k}) + \sum_{k=1}^K (1 - \pi_k) \log(\frac{1-\pi}{1-\pi_k}) \\
&+ \sum_k^K \sum_j^{l_k} \frac{1}{2} \pi_k \alpha_{jk} \log(\frac{s_{jk}^2}{\sigma_\beta^2}) + \frac{p}{2}\log(\sigma_\beta^2) + \frac{p}{2} + \frac{p}{2}\log(2\pi).
\end{aligned}
\tag{23}
$$

To get π_k and α_{jk}, we set the derivative of $\mathcal{L}(q)$ in (23) to be zero, i.e.,

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

which gives

$$
\begin{aligned}
\pi_k &= \frac{1}{1 + \exp(-u_k)}, \\
\text{where } u_k &= \log\frac{\pi}{1-\pi} + \frac{1}{2} \sum_j^{l_k} \alpha_{jk} \left(\log\frac{s_{jk}^2}{\sigma_\beta^2} + \frac{\mu_{jk}^2}{s_{jk}^2} \right); \\
\text{and } \alpha_{jk} &= \frac{1}{1 + \exp(-v_{jk})}, \\
\text{where } v_{jk} &= \log\frac{\alpha}{1-\alpha} + \frac{1}{2} \pi_k \left(\log\frac{s_{jk}^2}{\sigma_\beta^2} + \frac{\mu_{jk}^2}{s_{jk}^2} \right).
\end{aligned}
\tag{24}
$$

The detailed derivations in (24) is as follows:

$$
\begin{aligned}
u_k =& \log\frac{\pi}{1-\pi} + \frac{1}{2}\sum_j^{l_k}\alpha_{jk}\log\frac{s_{jk}^2}{\sigma_\beta^2} \\
&+ \frac{\sum_j^{l_k}\alpha_{jk}\mu_{jk}\mathbf{x}_{jk}^T\mathbf{y}}{\sigma_e^2} - \frac{\sum_j^{l_k}\alpha_{jk}\mu_{jk}\mathbf{x}_{jk}^T(\mathbf{Z}\boldsymbol{\omega})}{\sigma_e^2} - \frac{1}{2\sigma_e^2}\sum_j^{l_k}\alpha_{jk}(s_{jk}^2+\mu_{jk}^2)\mathbf{x}_{jk}^T\mathbf{x}_{jk} \\
&- \frac{1}{\sigma_e^2}\left(\sum_{k'\neq k}^{K}\sum_j^{l_k}\sum_{j'}^{l_{k'}}\pi_{k'}\alpha_{j'k'}\mu_{j'k'}\alpha_{jk}\mu_{jk}\mathbf{x}_{j'k'}^T\mathbf{x}_{jk}\right) \\
&- \frac{1}{\sigma_e^2}\left(\sum_j^{l_k}\sum_{j'\neq j}^{l_k}\alpha_{j'k}\mu_{j'k}\alpha_{jk}\mu_{jk}\mathbf{x}_{j'k}^T\mathbf{x}_{jk}\right) \\
&- \frac{1}{2\sigma_\beta}\sum_j^{l_k}\alpha_{jk}(s_{jk}^2+\mu_{jk}^2) + \frac{1}{2}\sum_j^{l_k}\alpha_{jk} \\
=& \log\frac{\pi}{1-\pi} + \frac{1}{2}\sum_j^{l_k}\alpha_{jk}\log\frac{s_{jk}^2}{\sigma_\beta^2} \\
&+ \alpha_{jk}\mu_{jk}\sum_j^{K}\underbrace{\left(\frac{\mathbf{x}_{jk}^T(\mathbf{y}-\mathbf{Z}\boldsymbol{\omega}) - \sum_{k'\neq k}^{K}\pi_{k'}\sum_{j'}^{l_{k'}}\alpha_{j'k'}\mu_{j'k'}\mathbf{x}_{j'k'}^T\mathbf{x}_{jk} - \sum_{j'\neq j}^{l_k}\alpha_{j'k}\mu_{j'k}\mathbf{x}_{j'k}^T\mathbf{x}_{jk}}{\sigma_e^2}\right)}_{\mu_{jk}/s_{jk}^2} \\
&- \frac{1}{2}\sum_j^{l_k}\alpha_{jk}(s_{jk}^2+\mu_{jk}^2)\underbrace{\left(\frac{\mathbf{x}_{jk}^T\mathbf{x}_{jk}}{\sigma_e^2}+\frac{1}{\sigma_\beta^2}\right)}_{1/s_{jk}^2} + \frac{1}{2}\sum_j^{l_k}\alpha_{jk} \\
=& \log\frac{\pi}{1-\pi} + \frac{1}{2}\sum_j^{l_k}\alpha_{jk}\log\frac{s_{jk}^2}{\sigma_\beta^2} + \sum_j^{l_k}\frac{\alpha_{jk}\mu_{jk}^2}{s_{jk}^2} - \sum_j^{l_k}\frac{\alpha_{jk}\mu_{jk}^2}{2s_{jk}^2} \\
=& \log\frac{\pi}{1-\pi} + \frac{1}{2}\sum_j^{l_k}\alpha_{jk}\left(\log\frac{s_{jk}^2}{\sigma_\beta^2} + \frac{\mu_{jk}^2}{s_{jk}^2}\right),
\end{aligned}
\tag{25}
$$

where we have used Equation (11) in the third equation above. Derivation of v_{jk} follows the same procedure and is omitted here.

1.2 M-step

At M-step, we update the parameters $\boldsymbol{\theta} = \{\alpha, \pi, \sigma_\beta^2, \sigma_e^2, \boldsymbol{\omega}\}$. Setting the partial derivative of $\mathcal{L}(q)$ with respect to model parameters to be zero yielding the closed-form updates:

$$
\begin{aligned}
\sigma_e^2 =& \frac{||\mathbf{y} - \mathbf{Z}\boldsymbol{\omega} - \sum_k^K\sum_j^{l_k}\pi_k\alpha_{jk}\mu_{jk}\mathbf{x}_{jk}||^2}{n} \\
&+ \frac{\sum_k^K\sum_j^{l_k}[\pi_k\alpha_{jk}(s_{jk}^2+\mu_{jk}^2) - (\pi_k\alpha_{jk}\mu_{jk})^2]\mathbf{x}_{jk}^T\mathbf{x}_{jk}}{n} \\
&+ \frac{\sum_k^K(\pi_k-\pi_k^2)[\sum_j^{l_k}\sum_{j'}^{l_k}\alpha_{j'k}\mu_{j'k}\alpha_{jk}\mu_{jk}]\mathbf{x}_{j'k}^T\mathbf{x}_{jk}}{n}.
\end{aligned}
\tag{26}
$$

$$\sigma_\beta^2 = \frac{\sum_k^K \sum_j^{l_k} \pi_k \alpha_{jk}(s_{jk}^2 + \mu_{jk}^2)}{\sum_k^K \sum_j^{l_k} \pi_k \alpha_{jk}}. \tag{27}$$

$$\alpha = \frac{1}{p} \sum_k^K \sum_j^{l_k} \alpha_{jk}, \tag{28}$$

$$\pi = \frac{1}{K} \sum_k^K \pi_k. \tag{29}$$

$$\boldsymbol{\omega} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T(\mathbf{y} - \sum_k^K \sum_j^{l_k} \pi_k \alpha_{jk} \mu_{jk} \mathbf{x}_{jk}). \tag{30}$$

2 Variational EM Algorithm: Multi-task Learning with BI-VAS

2.1 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 parameters under the multi-task model formulated as (14-16) of the main text. The joint probabilistic model is

$$\begin{aligned} &\Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta} | \mathbf{X}, \mathbf{Z}; \theta) = \Pr(\mathbf{y} | \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \mathbf{X}, \mathbf{Z}, \theta) \Pr(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta} | \theta) \\ =& \prod_{j=1}^L \mathcal{N}(\mathbf{y}_j | \mathbf{Z}_j \boldsymbol{\omega}_j + \sum_k^K \eta_k \gamma_{jk} \beta_{jk} \mathbf{x}_{jk}) \prod_{k=1}^K \pi^{\eta_k} (1-\pi)^{1-\eta_k} \prod_{j=1}^L \mathcal{N}(0, \sigma_{\beta_j}^2) \alpha^{\gamma_{jk}} (1-\alpha)^{1-\gamma_{jk}}. \end{aligned} \tag{31}$$

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 estimate the parameters $\boldsymbol{\theta}$ by maximizing the marginal likelihood, and evaluate the posterior distribution of β_{jk} based on the parameter estimates. The logarithm of the marginal likelihood is

$$\begin{aligned} \log\, p(\mathbf{y}|\mathbf{X}, \mathbf{Z}; \theta) &= \log \sum_{\boldsymbol{\gamma}} \sum_{\boldsymbol{\eta}} \int_\beta \Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta} | \mathbf{X}, \mathbf{Z}; \theta) d\boldsymbol{\beta} \\ &\geq \sum_{\boldsymbol{\gamma}} \sum_{\eta} \int_\beta q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}) \log \frac{\Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta} | \mathbf{X}, \mathbf{Z}; \theta)}{q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta})} \\ &= \mathbb{E}_q[\log \Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta} | \mathbf{X}, \mathbf{Z}; \theta) - q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta})] \\ &\equiv \mathcal{L}(q). \end{aligned} \tag{32}$$

Again, we assume that the variational distribution takes the form

$$q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}) = \prod_k^K \left(q(\eta_k) \prod_j^L (q(\beta_{jk} | \eta_k, \gamma_{jk}) q(\gamma_{jk})) \right). \tag{33}$$

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

$$\log q^*(\beta_{jk}|\eta_k=1,\gamma_{jk}=1)=\mathbb{E}_{j'\neq j|k}\left[\mathbb{E}_{k'\neq k}\log\Pr(\mathbf{y},\boldsymbol{\eta}_{-k},\eta_k=1,\boldsymbol{\gamma}_{-jk},\gamma_{jk}=1,\boldsymbol{\beta})\right]. \tag{34}$$

The Equation (34) contains the logarithm of joint probability funcion, which is

$$\begin{aligned}
\log\Pr(\mathbf{y},\boldsymbol{\eta},\boldsymbol{\gamma},\boldsymbol{\beta}|\mathbf{X},\mathbf{Z};\theta)=&\sum_{j=1}^{L}\Bigg\{-\frac{N_j}{2}\log(2\pi\sigma_{e_j}^2)-\frac{\mathbf{y}_j^T\mathbf{y}_j}{2\sigma_{e_j}^2}-\frac{(\mathbf{Z}_j\boldsymbol{\omega}_j)^T(\mathbf{Z}_j\boldsymbol{\omega}_j)}{2\sigma_{e_j}^2}\\
&+\frac{\sum_k^K\eta_k\gamma_{jk}\beta_{jk}\mathbf{x}_{jk}^T\mathbf{y}_j}{\sigma_{e_j}^2}+\frac{\mathbf{y}_j^T(\mathbf{Z}_j\boldsymbol{\omega}_j)}{\sigma_{e_j}^2}-\frac{\sum_k^K\eta_k\gamma_{jk}\beta_{jk}\mathbf{x}_{jk}^T(\mathbf{Z}_j\boldsymbol{\omega}_j)}{\sigma_{e_j}^2}\\
&-\frac{1}{2\sigma_{e_j}^2}\sum_k^K\left(\left(\eta_k\gamma_{jk}\beta_{jk}\right)^2\mathbf{x}_{jk}^T\mathbf{x}_{jk}\right)\\
&-\frac{1}{2\sigma_{e_j}^2}\left(\sum_k^K\sum_{k'\neq k}^K\left(\eta_{k'}\gamma_{jk'}\beta_{jk'}\right)\left(\eta_k\gamma_{jk}\beta_{jk}\right)\mathbf{x}_{jk'}^T\mathbf{x}_{jk}\right)\Bigg\}\\
&-\frac{K}{2}\sum_{j=1}^{L}\log(2\pi\sigma_{\beta_j}^2)-\sum_{j=1}^{L}\frac{\sum_{k=1}^K\beta_{jk}^2}{2\sigma_{\beta_j}^2}\\
&+\log(\alpha)\sum_{k=1}^{K}\sum_{j=1}^{L}\gamma_{jk}+\log(1-\alpha)\sum_{k=1}^{K}\sum_{j=1}^{L}(1-\gamma_{jk})\\
&+\log(\pi)\sum_{k=1}^{K}\eta_k+\log(1-\pi)\sum_{k=1}^{K}(1-\eta_k).
\end{aligned} \tag{35}$$

We then rearrange Equation (35) and retain the terms regarding jk

$$\begin{aligned}
&\log\Pr(\mathbf{y},\boldsymbol{\eta},\boldsymbol{\gamma},\boldsymbol{\beta}|\mathbf{X},\mathbf{Z};\theta)\\
=&-\frac{N_j}{2}\log(2\pi\sigma_{e_j}^2)-\frac{\mathbf{y}_j^T\mathbf{y}_j}{2\sigma_{e_j}^2}\\
&+\frac{\eta_k\gamma_{jk}\beta_{jk}\mathbf{x}_{jk}^T\mathbf{y}_j}{\sigma_{e_j}^2}-\frac{\eta_k\gamma_{jk}\beta_{jk}\mathbf{x}_{jk}^T(\mathbf{Z}_j\boldsymbol{\omega}_j)}{\sigma_{e_j}^2}\\
&-\frac{1}{2\sigma_{e_j}^2}\left(\left(\eta_k\gamma_{jk}\beta_{jk}\right)^2\mathbf{x}_{jk}^T\mathbf{x}_{jk}\right)\\
&-\frac{1}{2\sigma_{e_j}^2}\left(\sum_{k'\neq k}^{K}\left(\eta_k\gamma_{jk}\beta_{jk}\right)\left(\eta_{k'}\gamma_{jk'}\beta_{jk'}\right)\mathbf{x}_{jk}^T\mathbf{x}_{jk'}\right)-\frac{1}{2\sigma_{\beta_j}^2}\beta_{jk}^2\\
&+\log(\alpha)\gamma_{jk}+\log(1-\alpha)(1-\gamma_{jk})\\
&+\log(\pi)\eta_k+\log(1-\pi)(1-\eta_{jk})\\
&+\text{const.}
\end{aligned} \tag{36}$$

Next, we evaluate $q(\beta_{jk}|\eta_k, \gamma_{jk})$. When $\eta_k = \gamma_{jk} = 1 \Leftrightarrow \eta_k\gamma_{jk} = 1$, using Equation (34), we have

$$\begin{aligned}
&\log\, q(\beta_{jk}|\eta_k = 1, \gamma_{jk} = 1) \\
&= \left(-\frac{1}{2\sigma_{e_j}^2}\mathbf{x}_{jk}^T\mathbf{x}_{jk} - \frac{1}{2\sigma_{\beta_j}^2} \right) \beta_{jk}^2 \\
&\quad + \left(\frac{\mathbf{x}_{jk}^T(\mathbf{y}_j - \mathbf{Z}_j\boldsymbol{\omega}_j) - \sum_{k'\neq k}^K \mathbb{E}_{k'\neq k}[\eta_{k'}\gamma_{jk'}\beta_{jk'}]\mathbf{x}_{jk}^T\mathbf{x}_{jk'}}{\sigma_{e_j}^2} \right) \beta_{jk} \\
&\quad + \text{const},
\end{aligned} \tag{37}$$

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

$$\begin{aligned}
s_{jk}^2 &= \frac{\sigma_{e_j}^2}{\mathbf{x}_{jk}^T\mathbf{x}_{jk} + \frac{\sigma_{e_j}^2}{\sigma_{\beta_j}^2}} \\
\mu_{jk} &= \frac{\mathbf{x}_{jk}^T(\mathbf{y}_j - \mathbf{Z}_j\boldsymbol{\omega}_j) - \sum_{k'\neq k}^K \mathbb{E}_{k'\neq k}[\eta_{k'}\gamma_{jk'}\beta_{jk'}]\mathbf{x}_{jk}^T\mathbf{x}_{jk'}}{\mathbf{x}_{jk}^T\mathbf{x}_{jk} + \frac{\sigma_{e_j}^2}{\sigma_{\beta_j}^2}}.
\end{aligned} \tag{38}$$

Again, applying the property of Gaussian, we have

$$\log\, q(\beta_{jk}|\eta_k\gamma_{jk} = 0) = -\frac{1}{2\sigma_{\beta_j}^2}\beta_{jk}^2 + \text{const}, \tag{39}$$

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

$$q(\eta, \gamma, \beta) = \prod_k^K \left(\pi_k^{\eta_k}(1-\pi_k)^{1-\eta_k} \prod_j^L \left(\alpha_{jk}^{\gamma_{jk}}(1-\alpha_{jk})^{1-\gamma_{jk}} \mathcal{N}(\mu_{jk}, s_{jk}^2)^{\eta_k\gamma_{jk}} \mathcal{N}(0, \sigma_{\beta_j}^2)^{1-\eta_k\gamma_{jk}} \right) \right), \tag{40}$$

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

With this variational posterior probability function, we can now evaluate the second term of the lower bound $\mathcal{L}(q)$ in Equation (32):

$$
\begin{aligned}
&- \mathbb{E}_q[\log q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta})] \\
=&- \mathbb{E}_q\left[\sum_k^K (\eta_k \log(\pi_k) + (1-\eta_k)\log(1-\pi_k))\right] \\
&- \mathbb{E}_q\left[\sum_k^K \sum_j^L \eta_k\gamma_{jk}\log\mathcal{N}(\mu_{jk}, s_{jk}^2) + (1-\eta_k\gamma_{jk})\mathcal{N}(0,\sigma_{\beta_j}^2) + \gamma_{jk}\log(\alpha_{jk}) + (1-\gamma_{jk})\log(1-\alpha_{jk})\right] \\
=&- \sum_k^K \mathbb{E}_q\left[(\eta_k\log(\pi_k) + (1-\eta_k)\log(1-\pi_k))\right] \\
&- \sum_k^K \sum_j^L \mathbb{E}_q\left[\eta_k\gamma_{jk}\log\mathcal{N}(\mu_{jk}, s_{jk}^2) + (1-\eta_k\gamma_{jk})\mathcal{N}(0,\sigma_{\beta_j}^2) + \gamma_{jk}\log(\alpha_{jk}) + (1-\gamma_{jk})\log(1-\alpha_{jk})\right] \\
=&- \sum_k^K \sum_j^L \mathbb{E}_{\eta_k,\gamma_{jk}}\{\mathbb{E}_{\beta|\eta_k=1,\gamma_{jk}=1}[\log\mathcal{N}(\mu_{jk}, s_{jk}^2)] + \mathbb{E}_{\beta|\eta_k=1,\gamma_{jk}=0}[\log\mathcal{N}(0,\sigma_{\beta_j}^2)] \\
&+ \mathbb{E}_{\beta|\eta_k=0,\gamma_{jk}=1}[\log\mathcal{N}(0,\sigma_{\beta_j}^2)] + \mathbb{E}_{\beta|\eta_k=0,\gamma_{jk}=0}[\log\mathcal{N}(0,\sigma_{\beta_j}^2)]\} \\
&- \sum_k^K \sum_j^L \mathbb{E}_q[\gamma_{jk}\log(\alpha_{jk}) + (1-\gamma_{jk})\log(1-\alpha_{jk})] - \sum_k^K \mathbb{E}_q[\eta_k\log(\pi_k) + (1-\eta_k)\log(1-\pi_k)].
\end{aligned} \tag{41}
$$

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,

$$
\begin{aligned}
&- \mathbb{E}_{\beta|\eta_k=1,\gamma_{jk}=0}[\log\mathcal{N}(\mu_{jk}, s_{jk}^2)] \\
=&- \mathbb{E}_{\beta|\eta_k=0,\gamma_{jk}=1}[\log\mathcal{N}(\mu_{jk}, s_{jk}^2)] \\
=&- \mathbb{E}_{\beta|\eta_k=0,\gamma_{jk}=0}[\log\mathcal{N}(\mu_{jk}, s_{jk}^2)] \\
=&\frac{1}{2}\log(\sigma_{\beta_j}^2) + \frac{1}{2}(1+\log(2\pi)).
\end{aligned}
$$

Plugging these entropy terms into Equation (41), we have

$$
\begin{aligned}
&- \mathbb{E}_q[\log q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta})] \\
=&\sum_k^K \sum_j^L \{[\frac{1}{2}\log(s_{jk}^2) + \frac{1}{2}(1+\log(2\pi))]\pi_k\alpha_{jk} + [\frac{1}{2}\log(\sigma_{\beta_j}^2) + \frac{1}{2}(1+\log(2\pi))](1-\pi_k\alpha_{jk}) \\
&- \alpha_{jk}\log(\alpha_{jk}) - (1-\alpha_{jk})\log(1-\alpha_{jk})\} - \sum_k^K \{\pi_k\log(\pi_k) + (1-\pi_k)\log(1-\pi_k)\} \\
=&\sum_k^K \sum_j^L \frac{1}{2}\pi_k\alpha_{jk}(\log s_{jk}^2 - \log\sigma_{\beta_j}^2) + \frac{K}{2}\sum_j^L \log(\sigma_{\beta_j}^2) + \frac{p}{2} + \frac{p}{2}\log(2\pi) \\
&- \sum_k^K \sum_j^L [\alpha_{jk}\log(\alpha_{jk}) + (1-\alpha_{jk})\log(1-\alpha_{jk})] - \sum_k^K [\pi_k\log(\pi_k) + (1-\pi_k)\log(1-\pi_k)].
\end{aligned} \tag{42}
$$

Combine Equation (42) and Equation (35), the lower bound can be written in the following form:

$$
\begin{aligned}
&\mathbb{E}_q[\log \Pr(\mathbf{y}_j, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta} | \mathbf{X}; \theta)] - \mathbb{E}_q[\log q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta})] \\
=&\sum_{j=1}^{L} \Bigg\{ -\frac{N_j}{2}\log(2\pi\sigma_{e_j}^2) - \frac{\mathbf{y}_j^T\mathbf{y}_j}{2\sigma_{e_j}^2} - \frac{(\mathbf{Z}_j\boldsymbol{\omega}_j)^T(\mathbf{Z}_j\boldsymbol{\omega}_j)}{2\sigma_{e_j}^2} \\
&+ \frac{\sum_k^K \mathbb{E}_q\left[\eta_k\gamma_{jk}\beta_{jk}\right]\mathbf{x}_{jk}^T\mathbf{y}_j}{\sigma_{e_j}^2} + \frac{\mathbf{y}_j^T(\mathbf{Z}_j\boldsymbol{\omega}_j)}{\sigma_{e_j}^2} - \frac{\sum_k^K \mathbb{E}_q\left[\eta_k\gamma_{jk}\beta_{jk}\right]\mathbf{x}_{jk}^T(\mathbf{Z}_j\boldsymbol{\omega}_j)}{\sigma_{e_j}^2} \\
&- \frac{1}{2\sigma_{e_j}^2}\sum_k^K \left(\mathbb{E}_q\left[\left(\eta_k\gamma_{jk}\beta_{jk}\right)^2\right]\mathbf{x}_{jk}^T\mathbf{x}_{jk}\right) \\
&- \frac{1}{2\sigma_{e_j}^2}\left(\sum_k^K\sum_{k'\neq k}^K \mathbb{E}_q\left[\eta_{k'}\gamma_{jk'}\beta_{jk'}\right]\mathbb{E}_q\left[\eta_k\gamma_{jk}\beta_{jk}\right]\mathbf{x}_{jk'}^T\mathbf{x}_{jk}\right)\Bigg\} \\
&- \frac{K}{2}\sum_j^L \log(2\pi\sigma_{\beta_j}^2) - \sum_{k=1}^K \frac{\sum_{j=1}^L \mathbb{E}_q\left[\beta_{jk}^2\right]}{2\sigma_{\beta_j}^2} \\
&+ \log(\alpha)\sum_{k=1}^K\sum_{j=1}^L \mathbb{E}_q\left[\gamma_{jk}\right] + \log(1-\alpha)\sum_{k=1}^K\sum_{j=1}^L \mathbb{E}_q\left[1-\gamma_{jk}\right] \\
&+ \log(\pi)\sum_{k=1}^K \mathbb{E}_q\left[\eta_k\right] + \log(1-\pi)\sum_{k=1}^K \mathbb{E}_q\left[1-\eta_k\right] \\
&+ \sum_k^K\sum_j^L \frac{1}{2}\pi_k\alpha_{jk}(\log s_{jk}^2 - \log\sigma_{\beta_j}^2) + \frac{K}{2}\sum_j^L\log(\sigma_{\beta_j}^2) + \frac{p}{2} + \frac{p}{2}\log(2\pi) \\
&- \sum_k^K\sum_j^L[\alpha_{jk}\log(\alpha_{jk})] - \sum_k^K\sum_j^L[(1-\alpha_{jk})\log(1-\alpha_{jk})] \\
&- \sum_k^K[\pi_k\log(\pi_k)] - \sum_k^K[(1-\pi_k)\log(1-\pi_k)].
\end{aligned}
\tag{43}
$$

Again we can show with the same technique used in Section 1.1 that $\mathbb{E}_q\left[\eta_k\gamma_{jk}\beta_{jk}\right] = \pi_k\alpha_{jk}\mu_{jk}$, $\mathbb{E}_q\left[(\eta_k\gamma_{jk}\beta_{jk})^2\right] = \pi_k\alpha_{jk}(s_{jk}^2+\mu_{jk}^2)$, $\mathbb{E}_q\left[\beta_{jk}^2\right] = \pi_k\alpha_{jk}(s_{jk}^2+\mu_{jk}^2) + (1-\pi_k\alpha_{jk})\sigma_{\beta_j}^2$, $\mathbb{E}_q\left[\eta_k\right] = \pi_k$, $\mathbb{E}_q\left[\gamma_{jk}\right] = \alpha_{jk}$. By plugging in the expectations, the lower bound (43) becomes

$$
\begin{aligned}
&\mathbb{E}_q[\log \Pr(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta} | \mathbf{X}; \theta)] - \mathbb{E}_q[\log q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta})] \\
=&\sum_{j=1}^{L} \Bigg\{ -\frac{N_j}{2}\log(2\pi\sigma_{e_j}^2) - \frac{||\mathbf{y}_j - \mathbf{Z}_j\boldsymbol{\omega}_j - \sum_k^K \pi_k\alpha_{jk}\mu_{jk}\mathbf{x}_{jk}||^2}{2\sigma_{e_j}^2} \\
&- \frac{1}{2\sigma_{e_j}^2}\sum_k^K \underbrace{[\pi_k\alpha_{jk}(s_{jk}^2 + \mu_{jk}^2) - (\pi_k\alpha_{jk}\mu_{jk})^2]}_{\mathrm{Var}[\eta_k\gamma_{jk}\beta_{jk}]} \mathbf{x}_{jk}^T\mathbf{x}_{jk} \Bigg\} \\
&- \frac{K}{2}\sum_{j=1}^{L}\log(2\pi\sigma_{\beta_j}^2) - \frac{1}{2\sigma_{\beta_j}^2}\sum_{k=1}^{K}\sum_{j=1}^{L}[\pi_k\alpha_{jk}(s_{jk}^2 + \mu_{jk}^2) + (1 - \pi_k\alpha_{jk})\sigma_{\beta_j}^2] \\
&+ \sum_{k=1}^{K}\sum_{j=1}^{L}\alpha_{jk}\log(\frac{\alpha}{\alpha_{jk}}) + \sum_{k=1}^{K}\sum_{j=1}^{L}(1 - \alpha_{jk})\log(\frac{1-\alpha}{1-\alpha_{jk}}) \\
&+ \sum_{k=1}^{K}\pi_k\log(\frac{\pi}{\pi_k}) + \sum_{k=1}^{K}(1 - \pi_k)\log(\frac{1-\pi}{1-\pi_k}) \\
&+ \sum_{k}^{K}\sum_{j}^{L}\frac{1}{2}\pi_k\alpha_{jk}\log(\frac{s_{jk}^2}{\sigma_{\beta_j}^2}) + \frac{K}{2}\sum_{j}^{L}\log(\sigma_{\beta_j}^2) + \frac{p}{2} + \frac{p}{2}\log(2\pi).
\end{aligned}
\tag{44}
$$

To get π_k and α_{jk}, we let

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

which gives us

$$
\begin{aligned}
\pi_k &= \frac{1}{1 + \exp(-u_k)}, \\
\text{where } u_k =& \log\frac{\pi}{1-\pi} + \frac{1}{2}\sum_{j}^{L}\alpha_{jk}\left(\log\frac{s_{jk}^2}{\sigma_{\beta_j}^2} + \frac{\mu_{jk}^2}{s_{jk}^2}\right); \\
\text{and } \alpha_{jk} &= \frac{1}{1 + \exp(-v_{jk})}, \\
\text{where } v_{jk} =& \log\frac{\alpha}{1-\alpha} + \frac{1}{2}\pi_k\left(\log\frac{s_{jk}^2}{\sigma_{\beta_j}^2} + \frac{\mu_{jk}^2}{s_{jk}^2}\right).
\end{aligned}
\tag{45}
$$

The derivation of (45) is as follows:

$$
\begin{aligned}
u_k =& \log\frac{\pi}{1-\pi} + \frac{1}{2}\sum_j^L \alpha_{jk}\log\frac{s_{jk}^2}{\sigma_{\beta_j}^2} \\
&+ \sum_j^L \frac{1}{\sigma_{e_j}^2}\left\{\alpha_{jk}\mu_{jk}\mathbf{x}_{jk}^T\mathbf{y}_j - \alpha_{jk}\mu_{jk}\mathbf{x}_{jk}^T(\mathbf{Z}_j\boldsymbol{\omega}_j) - \sum_{k'\neq k}^K \pi_{k'}\alpha_{jk'}\mu_{jk'}\alpha_{jk}\mu_{jk}\mathbf{x}_{jk'}^T\mathbf{x}_{jk}\right\} \\
&- \frac{1}{2\sigma_{e_j}^2}\sum_j^L \alpha_{jk}(s_{jk}^2+\mu_{jk}^2)\mathbf{x}_{jk}^T\mathbf{x}_{jk} - \frac{1}{2\sigma_{\beta_j}}\sum_j^L \alpha_{jk}(s_{jk}^2+\mu_{jk}^2) + \frac{1}{2}\sum_j^L \alpha_{jk} \\
=& \log\frac{\pi}{1-\pi} + \frac{1}{2}\sum_j^L \alpha_{jk}\log\frac{s_{jk}^2}{\sigma_{\beta_j}^2} \\
&+ \sum_j^K \alpha_{jk}\mu_{jk}\underbrace{\left(\frac{\mathbf{x}_{jk}^T(\mathbf{y}_j - \mathbf{Z}_j\boldsymbol{\omega}_j) - \sum_{k'\neq k}^K \pi_{k'}\alpha_{jk'}\mu_{jk'}\mathbf{x}_{jk'}^T\mathbf{x}_{jk}}{\sigma_{e_j}^2}\right)}_{\mu_{jk}/s_{jk}^2} \\
&- \frac{1}{2}\sum_j^L \alpha_{jk}(s_{jk}^2+\mu_{jk}^2)\underbrace{\left(\frac{\mathbf{x}_{jk}^T\mathbf{x}_{jk}}{\sigma_{e_j}^2} + \frac{1}{\sigma_{\beta_j}^2}\right)}_{1/s_{jk}^2} + \frac{1}{2}\sum_j^L \alpha_{jk} \\
=& \log\frac{\pi}{1-\pi} + \frac{1}{2}\sum_j^L \alpha_{jk}\log\frac{s_{jk}^2}{\sigma_{\beta_j}^2} + \sum_j^L \frac{\alpha_{jk}\mu_{jk}^2}{s_{jk}^2} - \sum_j^L \frac{\alpha_{jk}\mu_{jk}^2}{2s_{jk}^2} \\
=& \log\frac{\pi}{1-\pi} + \frac{1}{2}\sum_j^L \alpha_{jk}\left(\log\frac{s_{jk}^2}{\sigma_{\beta_j}^2} + \frac{\mu_{jk}^2}{s_{jk}^2}\right)
\end{aligned}
\tag{46}
$$

where we have used Equation (38). Similarly, we can derive v_{jk}.

2.2 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

$$
\begin{aligned}
\sigma_{e_j}^2 =& \frac{||\mathbf{y}_j - \mathbf{Z}_j\boldsymbol{\omega}_j - \sum_k^K \pi_k\alpha_{jk}\mu_{jk}\mathbf{x}_{jk}||^2}{N_j} \\
&+ \frac{\sum_k^K[\pi_k\alpha_{jk}(s_{jk}^2+\mu_{jk}^2) - (\pi_k\alpha_{jk}\mu_{jk})^2]\mathbf{x}_{jk}^T\mathbf{x}_{jk}}{N_j}.
\end{aligned}
\tag{47}
$$

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

$$
\sigma_{\beta_j}^2 = \frac{\sum_k^K \pi_k\alpha_{jk}(s_{jk}^2+\mu_{jk}^2)}{\sum_k^K \pi_k\alpha_{jk}}. \tag{48}
$$

Accordingly,

$$
\alpha = \frac{1}{p}\sum_k^K\sum_j^L \alpha_{jk}, \tag{49}
$$

$$\pi = \frac{1}{K}\sum_{k}^{K}\pi_k. \tag{50}$$

$$\boldsymbol{\omega}_j = (\mathbf{Z}_j^T\mathbf{Z}_j)^{-1}\mathbf{Z}_j^T(\mathbf{y}_j - \sum_{k}^{K}\pi_k\alpha_{jk}\mu_{jk}\mathbf{x}_{jk}). \tag{51}$$

2.3 Summary of multi-task EM algorithm

In summary, we have

$$\begin{aligned}
s_{jk}^2 &= \frac{\sigma_{e_j}^2}{\mathbf{x}_{jk}^T\mathbf{x}_{jk} + \frac{\sigma_{e_j}^2}{\sigma_{\beta_j}^2}} \\
\mu_{jk} &= \frac{\mathbf{x}_{jk}^T(\mathbf{y}_j - \mathbf{Z}_j\boldsymbol{\omega}_j - \tilde{\mathbf{y}}_{jk})}{\mathbf{x}_{jk}^T\mathbf{x}_{jk} + \frac{\sigma_{e_j}^2}{\sigma_{\beta_j}^2}} \\
\pi_k &= \frac{1}{1+\exp(-u_k)},\ \text{where } u_k = \log\frac{\pi}{1-\pi} + \frac{1}{2}\sum_{j}^{L}\alpha_{jk}\left(\log\frac{s_{jk}^2}{\sigma_{\beta_j}^2} + \frac{\mu_{jk}^2}{s_{jk}^2}\right) \\
\alpha_{jk} &= \frac{1}{1+\exp(-v_k)},\ \text{where } v_k = \log\frac{\alpha}{1-\alpha} + \frac{1}{2}\pi_k\left(\log\frac{s_{jk}^2}{\sigma_{\beta_j}^2} + \frac{\mu_{jk}^2}{s_{jk}^2}\right)
\end{aligned} \tag{52}$$

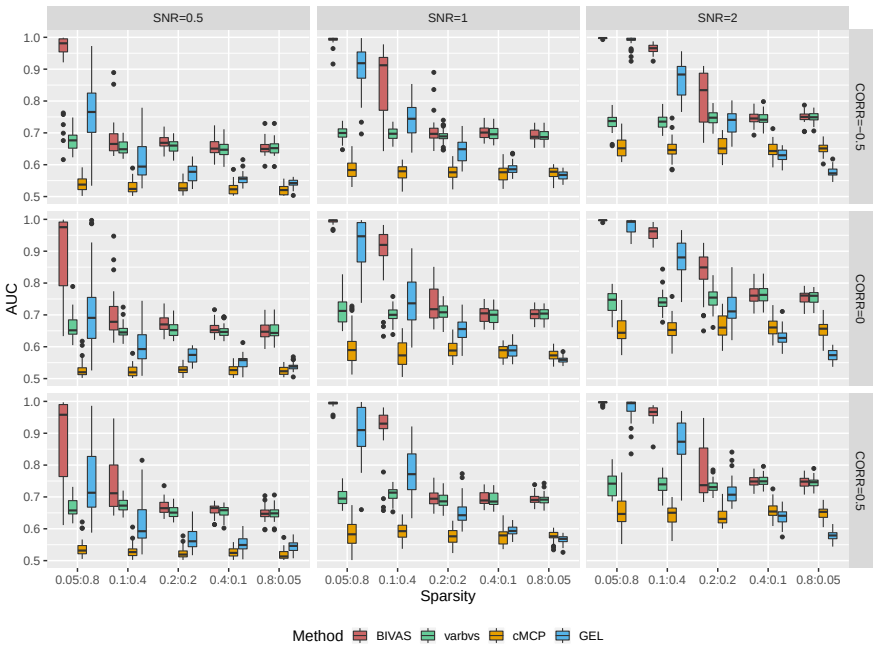
for E-step; and

$$\begin{aligned}
\sigma_{e_j}^2 =& \frac{||\mathbf{y}_j - \mathbf{Z}_j\boldsymbol{\omega}_j - \sum_k^K \pi_k\alpha_{jk}\mu_{jk}\mathbf{x}_{jk}||^2}{N_j} \\
&+ \frac{\sum_k^K[\pi_k\alpha_{jk}(s_{jk}^2+\mu_{jk}^2) - (\pi_k\alpha_{jk}\mu_{jk})^2]\mathbf{x}_{jk}^T\mathbf{x}_{jk}}{N_j}, \\
\sigma_{\beta_j}^2 =& \frac{\sum_k^K \pi_k\alpha_{jk}(s_{jk}^2+\mu_{jk}^2)}{\sum_k^K \pi_k\alpha_{jk}}, \\
\alpha =& \frac{1}{p}\sum_{k}^{K}\sum_{j}^{L}\alpha_{jk}, \\
\pi =& \frac{1}{K}\sum_{k}^{K}\pi_k, \\
\boldsymbol{\omega}_j =& (\mathbf{Z}_j^T\mathbf{Z}_j)^{-1}\mathbf{Z}_j^T(\mathbf{y}_j - \sum_{k}^{K}\pi_k\alpha_{jk}\mu_{jk}\mathbf{x}_{jk}),
\end{aligned} \tag{53}$$

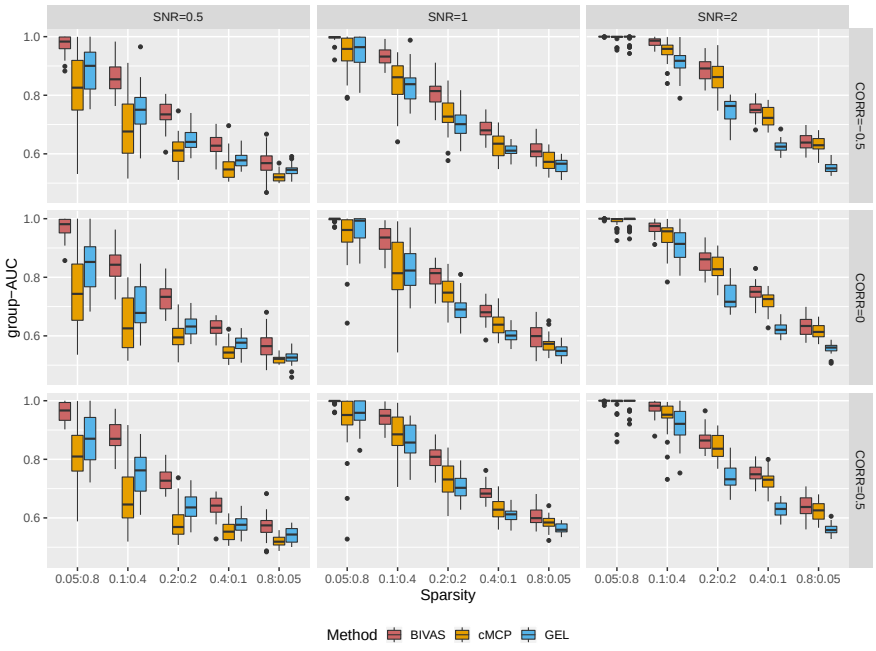
for M-step.

3 Complete outcomes of simulation study with $\rho \in \{-0.5, 0, 0.5\}$

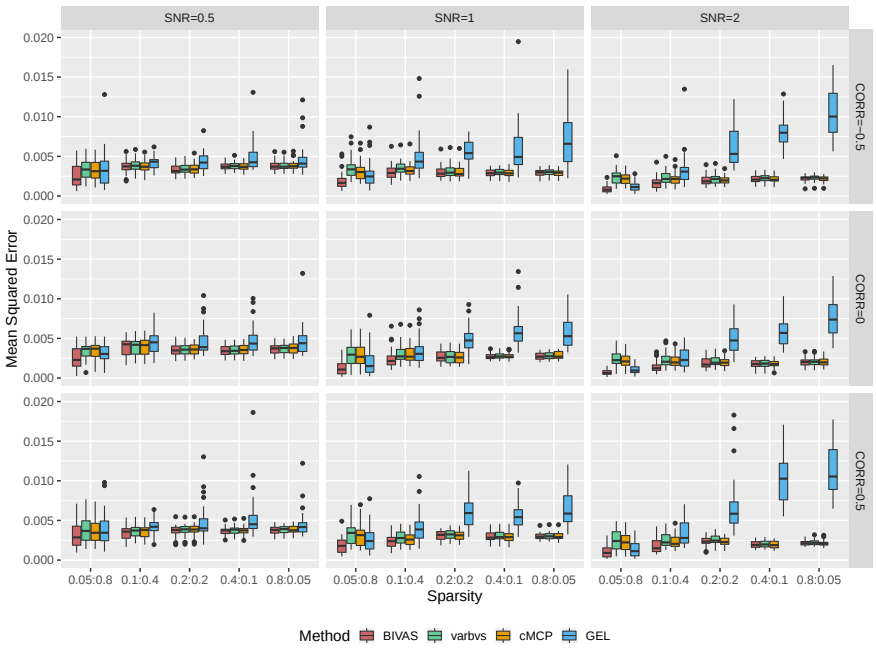
Here we present the complete outcomes of simulation study described in the main text with $\rho \in \{-0.5, 0, 0.5\}$.



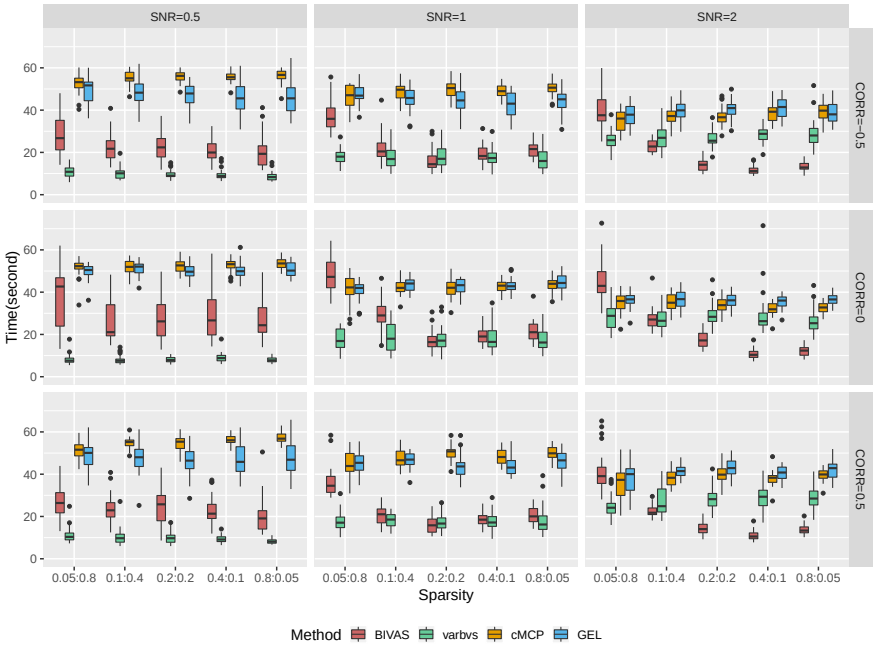
Supplementary Figure 1: AUC of BIVAS, varbvs, cMCP and GEL (coupling parameter $\tau = 1/3$) in identifying active predictors.



Supplementary Figure 2: AUC of BIVAS, varbvs, cMCP and GEL (coupling parameter $\tau = 1/3$) in identifying active groups.



Supplementary Figure 3: Mean aquared error of BIVAS, varbvs, cMCP and GEL (coupling parameter $\tau = 1/3$) in estimating $\boldsymbol{\beta}$.



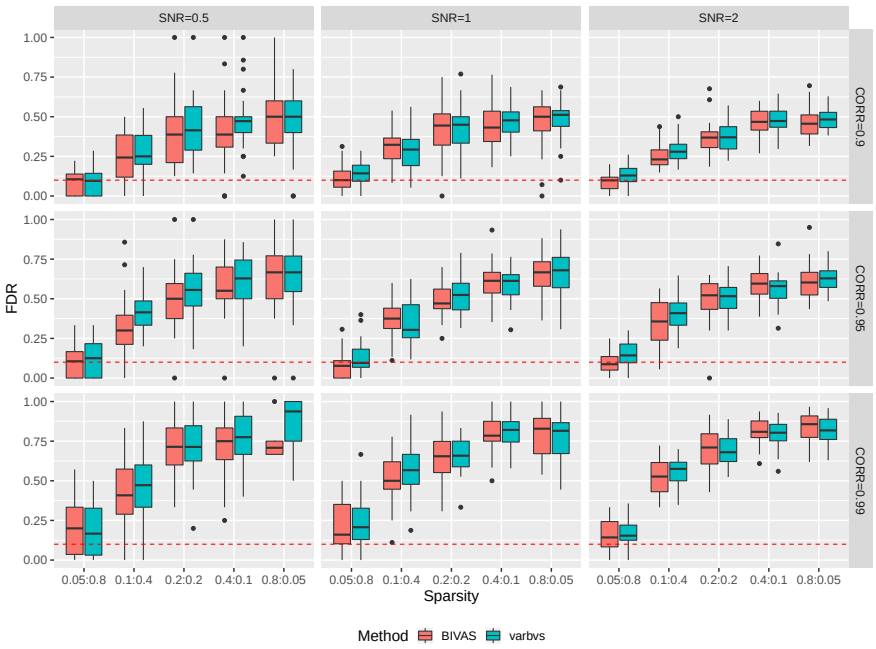
Supplementary Figure 4: Fitting times of BIVAS, varbvs, cMCP and GEL (coupling parameter $\tau = 1/3$).

4 Cases of strongly correlated variables: $\rho \in \{0.9, 0.95, 0.99\}$

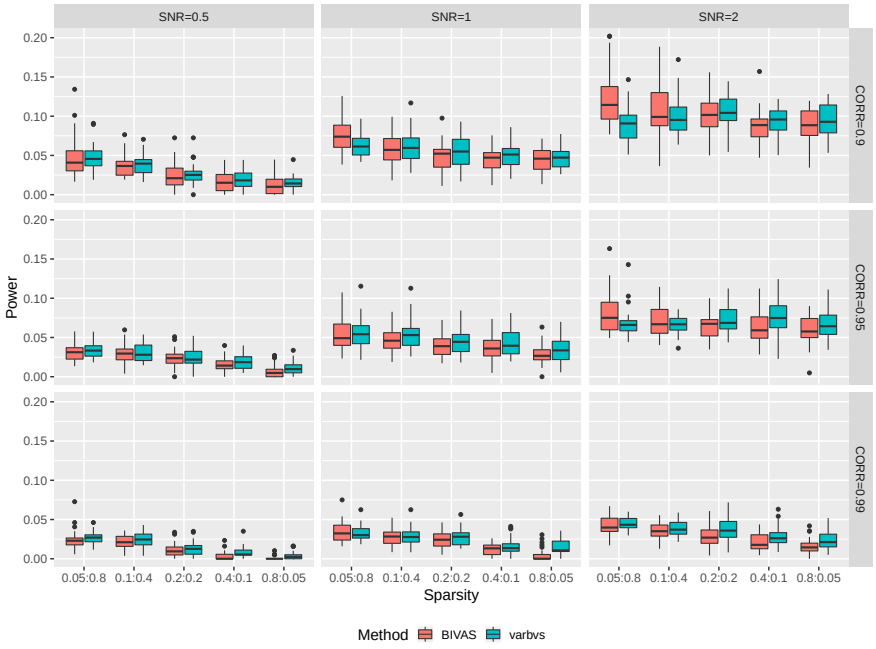
We further compared the performance of BIVAS with alternative approaches for $\rho \in \{0.9, 0.95, 0.99\}$, where the variational assumption is badly violated.

The results are shown in Supplementary Figures 5-10. Supplementary Figures 5-6 show that BIVAS and varbvs have similar performances in FDR and power when ρ is large. Because only the columns of $\mathbf{X}$ within the same group are correlated, the covariance matrix of the true posterior of $\boldsymbol{\beta}$ will be sparse if most of the groups are inactive (i.e., small π). In this case, the variational assumption is not badly violated. Hence both methods have well-controlled FDR and high power when the sparsity-in-group dominates (See left-most columns in the panels of Supplementary Figures 5-6). As more groups become active, the true covariance matrix of $\boldsymbol{\beta}$ is no longer sparse. Therefore, both methods have inflated FDR and decreased power when the 'bulk' of sparsity moves to individual variable level. According to Supplementary Figures 7-8, BIVAS still outperforms other methods in AUC and group-AUC although the gain is less than that we obseved when ρ is small. Supplementary Figure 9 shows that all approaches have comparable performance in the estimation accuray of $\boldsymbol{\beta}$. Regarding the computational efficiency, BIVAS is still faster than other methods in most cases (Supplementary Figure 10).

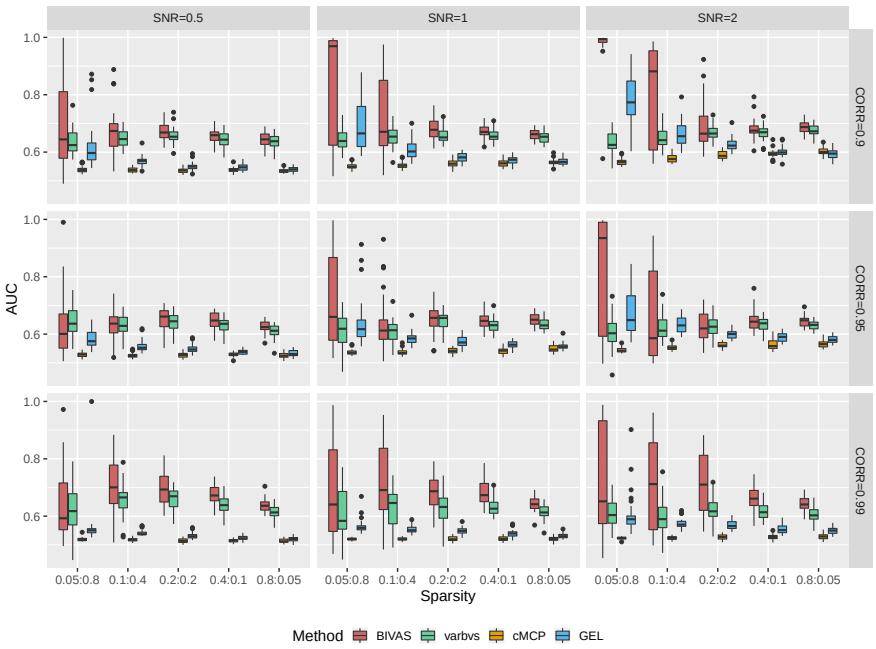
To sum up, our method has comparable performance with alternative approaches in most cases when there is large correlation between variables. While we have reduced benifit of using BIVAS in this scenario, we note that this situation is rare and unrealistic in most real data applications.



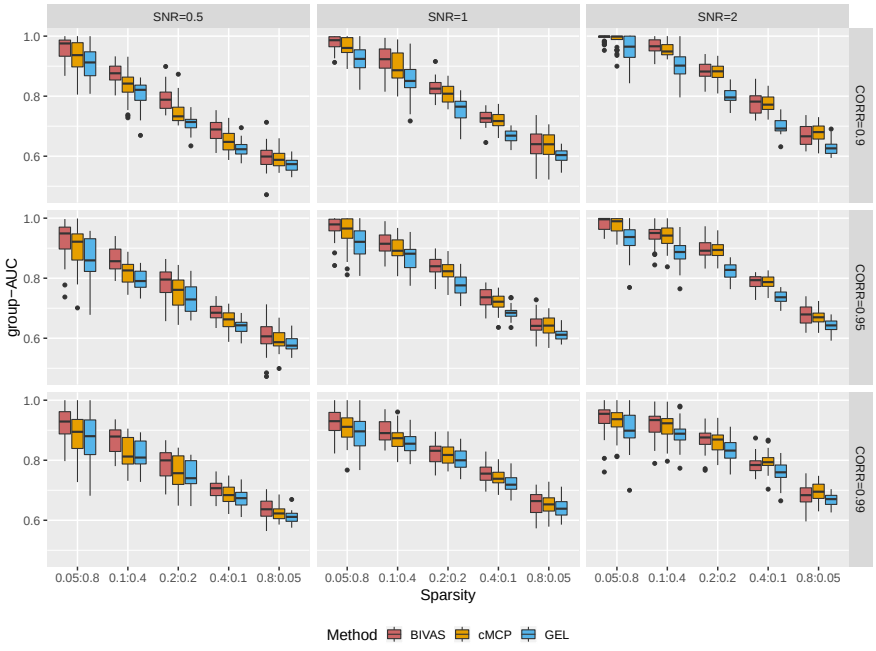
Supplementary Figure 5: FDR of BIVAS and varbvs for individual variable selection when $\rho \in \{0.9, 0.95, 0.99\}$. We controlled the global FDR at the nominal level 0.1 (the red dashed lines).



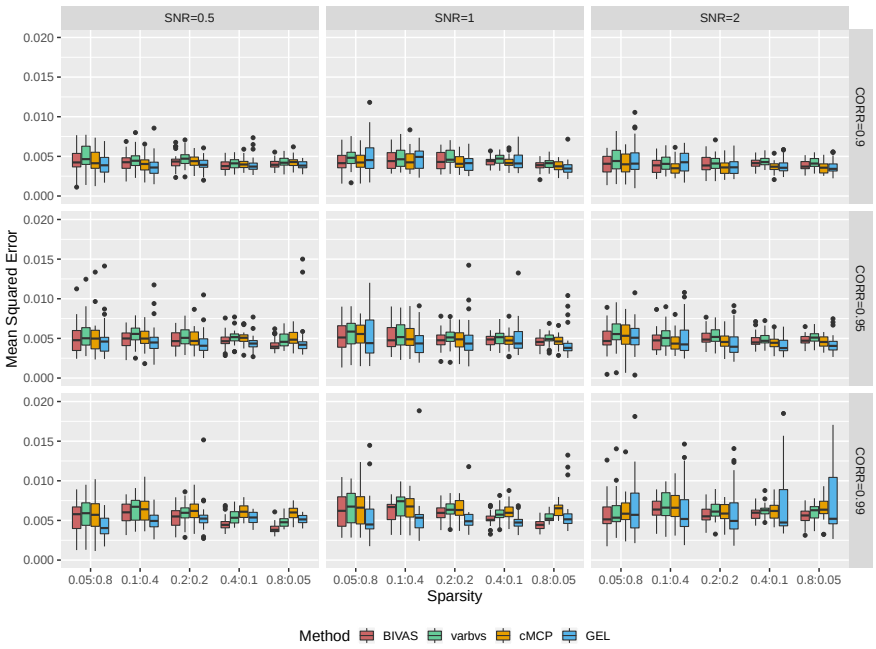
Supplementary Figure 6: Power of BIVAS and varbvs for individual variable selection when $\rho \in \{0.9, 0.95, 0.99\}$. We controlled the global FDR at the nominal level 0.1.



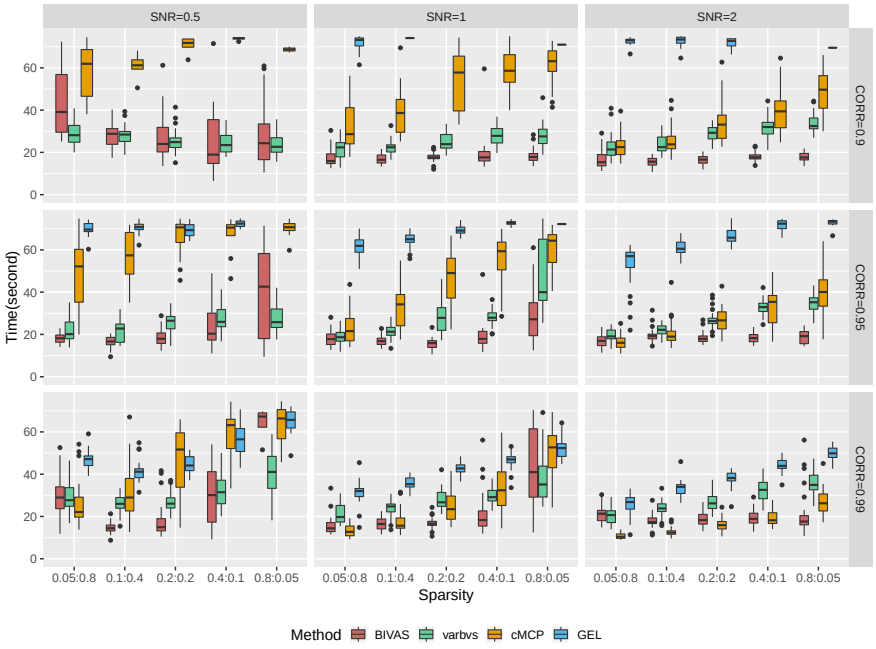
Supplementary Figure 7: AUC of BIVAS, varbvs, cMCP and GEL (coupling parameter $\tau = 1/3$) in identifying active predictors when $\rho \in \{0.9, 0.95, 0.99\}$.



Supplementary Figure 8: AUC of BIVAS, varbvs, cMCP and GEL (coupling parameter $\tau = 1/3$) in identifying active groups when $\rho \in \{0.9, 0.95, 0.99\}$.



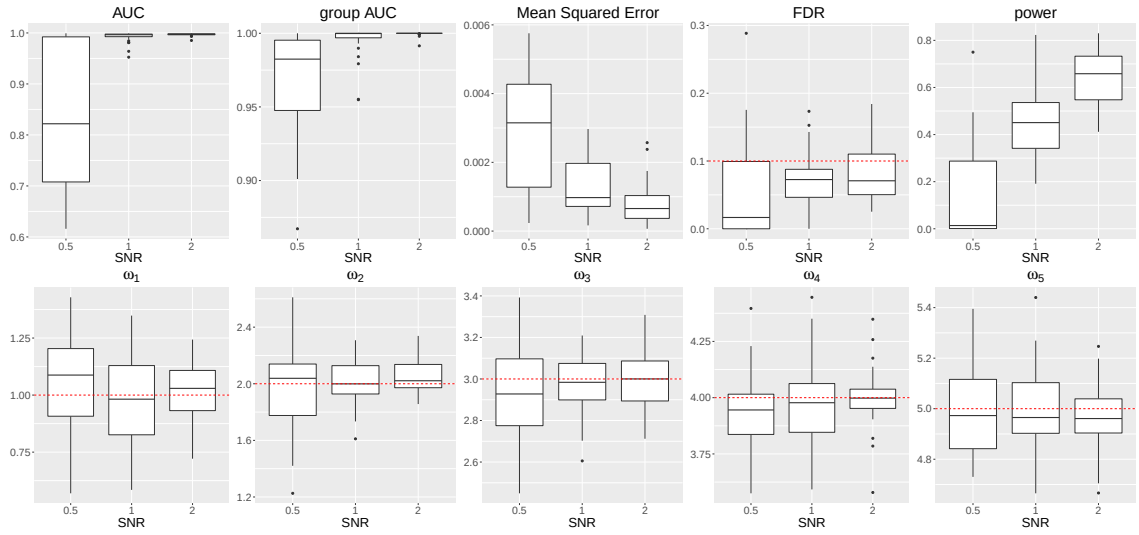
Supplementary Figure 9: Mean aquared error of BIVAS, varbvs, cMCP and GEL (coupling parameter $\tau = 1/3$) in estimating $\boldsymbol{\beta}$ when $\rho \in \{0.9, 0.95, 0.99\}$.



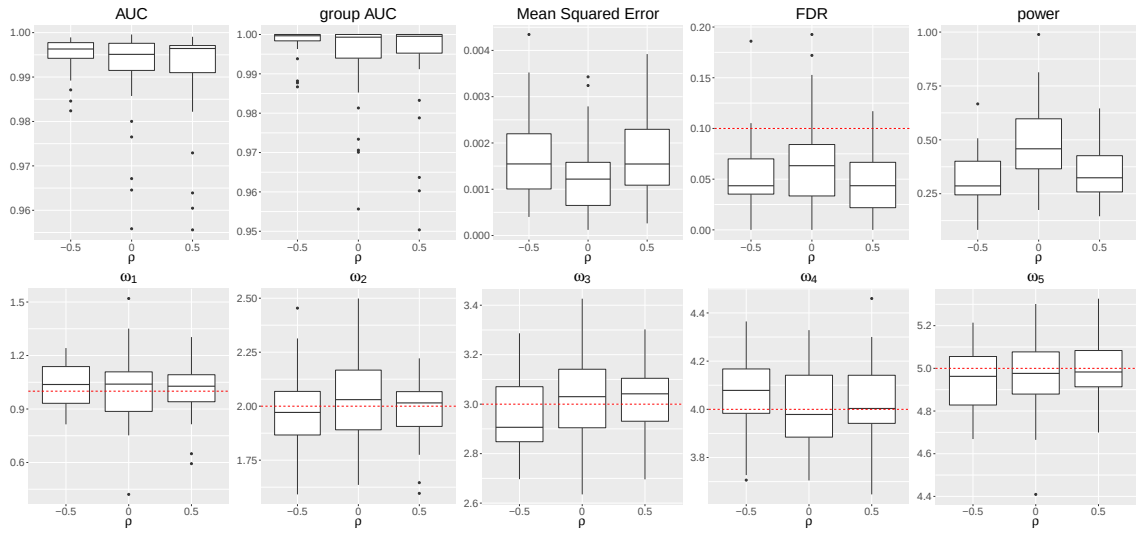
Supplementary Figure 10: Fitting times of BIVAS, varbvs, cMCP and GEL (coupling parameter $\tau = 1/3$) when $\rho \in \{0.9, 0.95, 0.99\}$.

5 Influence of fixed effects

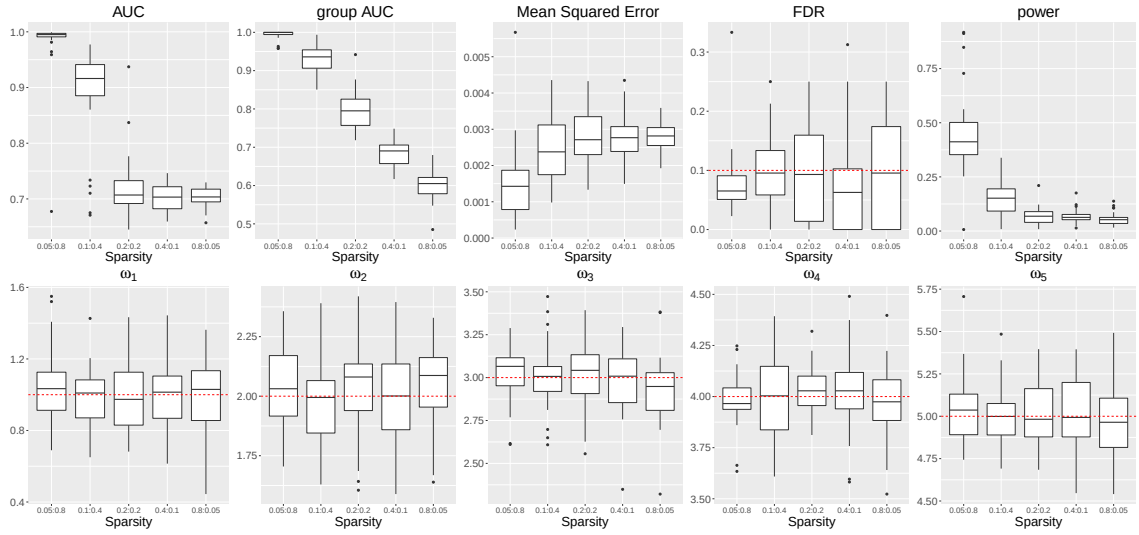
In this section, we conducted simulation in the presence of covariates $\mathbf{Z}$ with fiexed effects $\boldsymbol{\omega}$. In practice, the number of fixed covariates r is usually much smaller than the sample size n. Here we considered $n = 1,000$, $r = 5$, $p = 5,000$, and $K = 250$ with 20 variables in each group. The true values of fixed effects were set as $\boldsymbol{\omega} = [1, 2, 3, 4, 5]^T$. To analyze the performance of BIVAS under various situations, we varied SNR, ρ and (π, α) in turn and evaluated AUC, group AUC, MSE, FDR, power and the estimates of $\boldsymbol{\omega}$. The results are shown in Supplementary Figures 11-13. As shown in the bottom rows of the three Supplementary Figures, BIVAS provides accurate estimations of fixed effect $\boldsymbol{\omega}$ under various situations. Besides, by comparing the variable selection performance (i.e. AUC, group-AUC, MSE, FDR and power) with those evaluated without fixed effects term given in the main text, we can observe the patterns are not influenced by the covariates $\mathbf{Z}$.



Supplementary Figure 11: AUC, group-AUC, MSE of $\boldsymbol{\beta}$, FDR power and eistimates of $\boldsymbol{\omega}$ with SNR varied at $\{0.5, 1, 2\}$. We controlled FDR at 0.1 (the red dashed line in the FDR panel). The red dashed lines in the bottom rows represent the true values of $\boldsymbol{\omega}$.



Supplementary Figure 12: AUC, group-AUC, MSE of $\boldsymbol{\beta}$, FDR power and eistimates of $\boldsymbol{\omega}$ with $\rho \in \{-0.5, 0, 0.5\}$. We controlled FDR at 0.1 (the red dashed line in the FDR panel). The red dashed lines in the bottom rows represent the true values of $\boldsymbol{\omega}$.



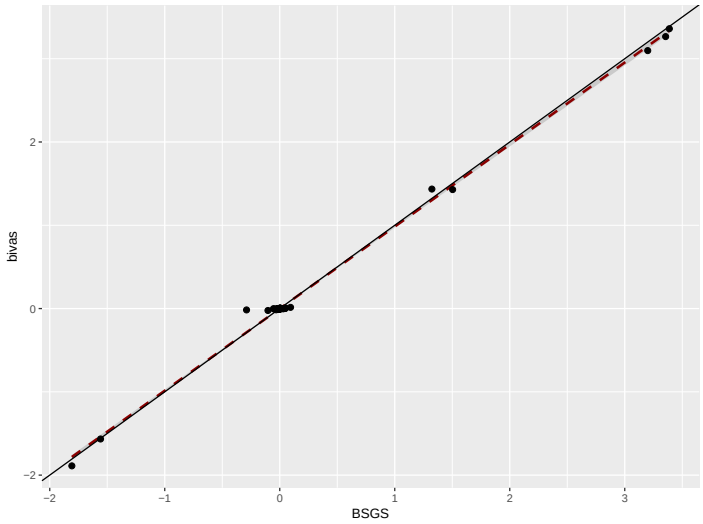
Supplementary Figure 13: AUC, group-AUC, MSE of $\boldsymbol{\beta}$, FDR power and eistimates of $\boldsymbol{\omega}$ with $(\pi, \alpha) \in \{(0.05, 0.8),\ (0.1, 0.4),\ (0.2, 0.2),\ (0.4, 0.1),\ (0.8, 005)\}$. We controlled FDR at 0.1 (the red dashed line in the FDR panel). The red dashed lines in the bottom rows represent the true values of $\boldsymbol{\omega}$.

	cMCP		GEL	
	SNP	gene	SNP	gene
HDL	98	95	123	55
RA	712	629	56	53
T1D	531	474	111	98

Supplementary Table 1: SNPs and genes identified by cMCP and GEL.

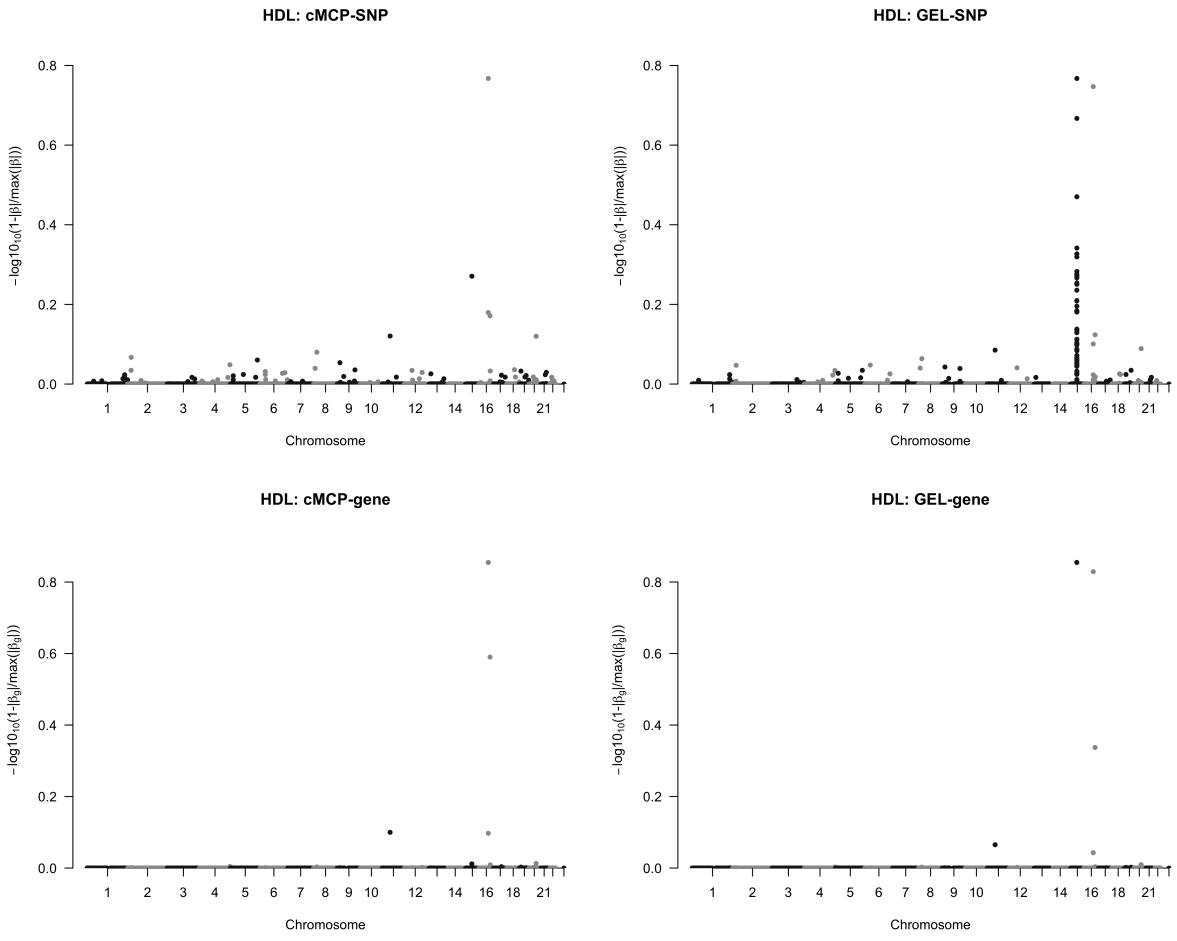
6 Posterior mean comparisons of BIVAS and BSGS

In this section, we compared the posterior means of $\boldsymbol{\beta}$ obtained by BIVAS and BSGS. We chose $n = 50$, $p = 100$ and $K = 10$ with 10 variables in each group. Then, we made group $1, 2, 5, 8$ active and set the corresponding nonzero effects $\beta_{7,1} = \beta_{8,1} = \beta_{9,1} = 3.2$, $\beta_{1,2} = \beta_{2,2} = 1.5$, $\beta_{3,5} = -1.5$ and $\beta_{7,8} = -2$, where $\beta_{j,k}$ represents the effect size of the j-th variable in the k-th group. The total number of iterations of the BSGS Gibbs sampler was set at 2000 and the fixed parameters were set as $\tau_{jk}^2 = 5$, $\rho_k = \theta_{jk} = 0.5$. We used the same values to initialize BIVAS. It took 192.648 seconds for BSGS and only 0.198 seconds for BIVAS to fit the model. The resulting posterior mean estimates of the two approaches are given in Supplementary Figure 14.

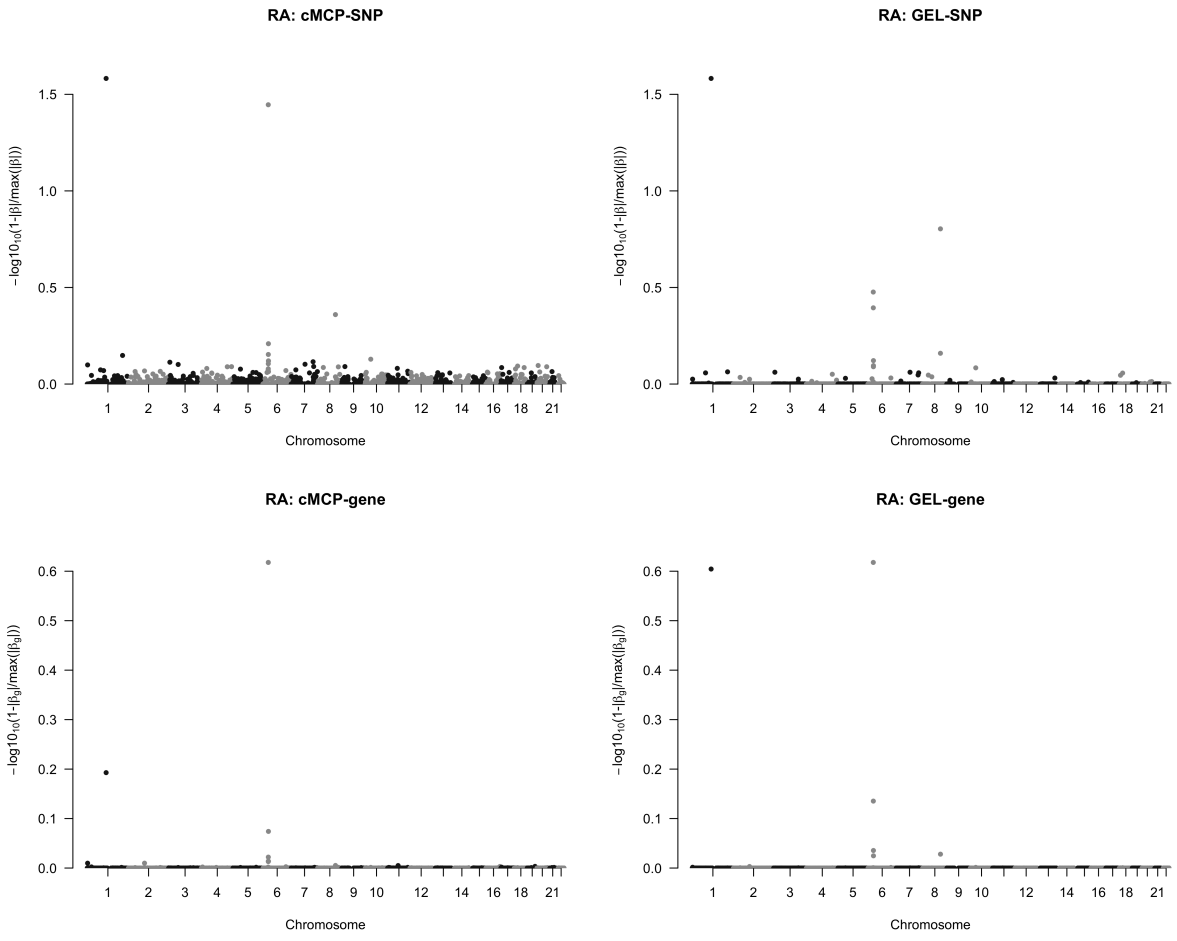


Supplementary Figure 14: Posterior means of $\boldsymbol{\beta}$ obtained by BIVAS and BSGS. The dashed regression line is fitted by the points. The solid line represents a perfect match.

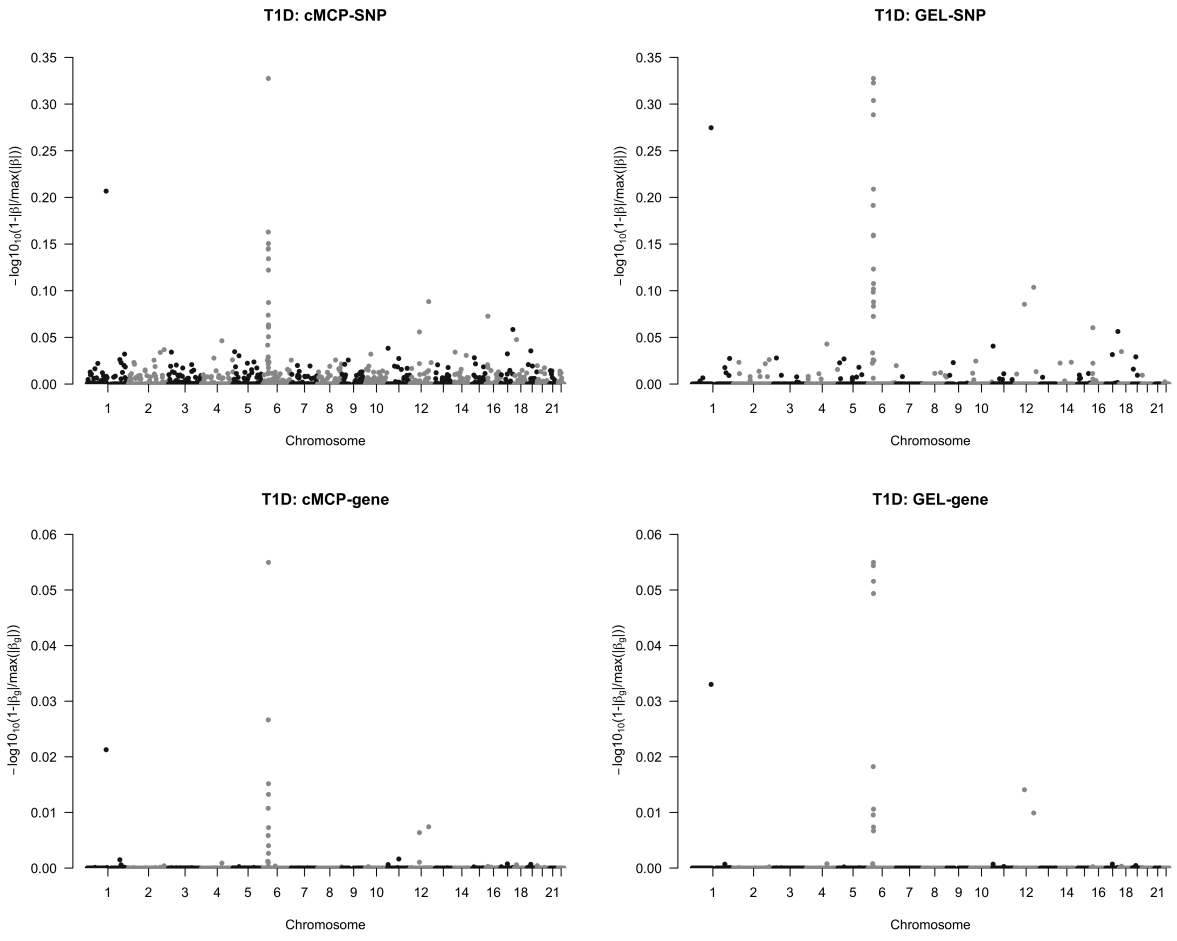
7 Real data results produced by penalized methods



Supplementary Figure 15: Manhattan plots of High-Density Lipoprotein (HDL) produced by cMCP and GEL. Plotted on the y-axis is the relative effect size.



Supplementary Figure 16: Manhattan plots of Rheumatoid Arthritis (RA) produced by cMCP and GEL. Plotted on the y-axis is the relative effect size.



Supplementary Figure 17: Manhattan plots of Type 1 Diabetes (T1D) produced by cMCP and GEL. Plotted on the y-axis is the relative effect size.