

Web-based Supplementary Materials

for

Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure

by

Yanming Li, Bin Nan and Ji Zhu

A Proofs of technical results

A.1 Some matrix algebra

Lemma A.1 *Let*

$$L^0(\mathbf{B}) = \frac{1}{2} \|\mathbf{Y} - \mathbf{XB}\|_2^2 = \frac{1}{2} \text{tr}(\mathbf{Y} - \mathbf{XB})^\top (\mathbf{Y} - \mathbf{XB}) = \frac{1}{2} \sum_{i=1}^n \sum_{k=1}^q (y_{ik} - \sum_{j=1}^p x_{ij} \beta_{jk})^2.$$

Then

$$\partial L^0(\mathbf{B}) / \partial \beta_{jk} = -\mathbf{x}_j^\top (\mathbf{Y} - \mathbf{XB})_{\cdot k} = -S_{jk} + \|\mathbf{x}_j\|_2^2 \beta_{jk},$$

where $S_{jk} = \mathbf{x}_j^\top (\mathbf{Y} - \mathbf{XB}^{(-j)})_{\cdot k}$.

A.2 Proof of Theorem 3.1

Following Lemma A.1,

$$L(\mathbf{B}) = \frac{1}{n} L^0(\mathbf{B}) + \sum_{1 \leq j \leq p; 1 \leq k \leq q} \lambda_{jk} |\beta_{jk}| + \sum_{g \in \mathcal{G}} \lambda_g \|\mathbf{B}_g\|_2.$$

For a coordinate β_{jk} in \mathbf{B} , denote $\mathcal{G}_{jk} = \{g : \beta_{jk} \in \mathbf{B}_g \in \mathcal{G}\}$, then when $L(\mathbf{B})$ is differentiable at β_{jk} ,

$$\frac{\partial L(\mathbf{B})}{\partial \beta_{jk}} = -S_{jk}/n + \|\mathbf{x}_j\|_2^2 \beta_{jk}/n + \lambda_{jk} \text{sign}(\beta_{jk}) + \sum_{\mathcal{G}_{jk}} \lambda_g \beta_{jk} / \|\mathbf{B}_g\|_2.$$

If $\beta_{jk} > 0$, then for any $\mathbf{B}_g \in \mathcal{G}_{jk}$, $\|\mathbf{B}_g\|_2 > 0$, and

$$\frac{\partial L(\mathbf{B})}{\partial \beta_{jk}} = -S_{jk}/n + \|\mathbf{x}_j\|_2^2 \beta_{jk}/n + \lambda_{jk} + \sum_{\mathcal{G}_{jk}} \lambda_g \beta_{jk} / \|\mathbf{B}_g\|_2.$$

Notice that $\partial L(\mathbf{B})/\partial \beta_{jk} \geq 0$ if and only if

$$\beta_{jk} \geq \frac{S_{jk} - n\lambda_{jk}}{\|\mathbf{x}_j\|_2^2 + n \sum_{\mathcal{G}_{jk}} \lambda_g / \|\mathbf{B}_g\|_2} \triangleq \tilde{\beta}_{jk}^+,$$

and $\partial L(\mathbf{B})/\partial \beta_{jk} < 0$ if and only if $\beta_{jk} < \tilde{\beta}_{jk}^+$. So fixing all other coordinates of \mathbf{B} , if $\beta_{jk} > 0$, then $L(\mathbf{B})$ is monotone increasing with respect to β_{jk} when $\beta_{jk} > \tilde{\beta}_{jk}^+$ and decreasing when $\beta_{jk} < \tilde{\beta}_{jk}^+$. Therefore, if $\hat{\beta}_{jk, \min}^+$ minimizes $L(\mathbf{B})$ with respect to β_{jk} when $\beta_{jk} > 0$, then

$$\hat{\beta}_{jk, \min}^+ = \begin{cases} \frac{S_{jk} - n\lambda_{jk}}{\|\mathbf{x}_j\|_2^2 + n \sum_{\mathcal{G}_{jk}} \lambda_g / \|\mathbf{B}_g\|_2}, & \text{if } S_{jk} > n\lambda_{jk} \\ 0, & \text{otherwise.} \end{cases} \quad (\text{A.1})$$

Similarly, if $\hat{\beta}_{jk, \min}^-$ minimizes $L(\mathbf{B})$ with respect to β_{jk} when $\beta_{jk} < 0$, then

$$\hat{\beta}_{jk, \min}^- = \begin{cases} \frac{S_{jk} + n\lambda_{jk}}{\|\mathbf{x}_j\|_2^2 + n \sum_{\mathcal{G}_{jk}} \lambda_g / \|\hat{\mathbf{B}}_g\|_2}, & \text{if } S_{jk} < -n\lambda_{jk} \\ 0, & \text{otherwise.} \end{cases} \quad (\text{A.2})$$

Based on both (A.1) and (A.2), when $\beta_{jk} \neq 0$, the minimizer $\hat{\beta}_{jk}$ can be summarized into a unified form:

$$\hat{\beta}_{jk} = \frac{\text{sgn}(S_{jk}) (|S_{jk}| - n\lambda_{jk})_+}{\|\mathbf{x}_j\|_2^2 + n \sum_{\{g \in \mathcal{G}_{jk}\}} \lambda_g / \|\hat{\mathbf{B}}_g\|_2}.$$

When $\beta_{jk} = 0$, we discuss two separate cases according to whether all the groups containing β_{jk} are zero groups or not. First, if none of the groups in \mathcal{G}_{jk} is a zero group, then $\hat{\beta}_{jk}$ needs to satisfy the subgradient equation

$$S_{jk}/n - \|\mathbf{x}_j\|_2^2 \hat{\beta}_{jk}/n = \lambda_{jk} u + \sum_{\mathcal{G}_{jk}} \hat{\beta}_{jk} / \|\hat{\mathbf{B}}_g\|_2 \quad (\text{A.3})$$

with $|u| \leq 1$. Then a similar discussion to the case when $\beta_{jk} \neq 0$ based on whether $u > 0$ or $u \leq 0$ in (A.3) yields the same expression:

$$\hat{\beta}_{jk} = \frac{\text{sgn}(S_{jk}) (|S_{jk}| - n\lambda_{jk})_+}{\|\mathbf{x}_j\|_2^2 + n \sum_{\{g \in \mathcal{G}_{jk}\}} \lambda_g / \|\hat{\mathbf{B}}_g\|_2}.$$

Second, if some of the groups in \mathcal{G}_{jk} are zero groups, then neither $|\cdot|$ nor $\|\cdot\|_2$ in $L(\mathbf{B})$ is differentiable at zero. Let $\mathcal{G}_{jk}^\otimes = \{g : \beta_{jk} \in \mathbf{B}_g \in \mathcal{G}, \|\mathbf{B}\|_g > 0\}$. Then for any zero group \mathbf{B}_{g_0} in \mathcal{G}_{jk} and for all $\beta_{jk} \in \mathbf{B}_{g_0}$, $\hat{\beta}_{jk}$ needs to satisfy the subgradient equation:

$$S_{jk}/n - \|\mathbf{x}_j\|_2^2 \hat{\beta}_{jk}/n = \lambda_{jk}u + \lambda_{g_0}v_{jk} + \sum_{\mathcal{G}_{jk}^\otimes} \lambda_g \hat{\beta}_{jk} / \|\hat{\mathbf{B}}_g\|_2, \quad (\text{A.4})$$

where u is the subgradient scalar for the L_1 norm $|\cdot|$ and \mathbf{v} is the subgradient vector for the L_2 norm $\|\cdot\|_2$ with constrains $|u| \leq 1$ and $\|\mathbf{v}\|_2 \leq 1$, and λ_{g_0} is the tuning parameter associated with \mathbf{B}_{g_0} . It can be seen directly (similar to the case of $\beta_{jk} = 0$) that (A.4) yields $\hat{\mathbf{B}}_{g_0} = 0$ if

$$\sqrt{\sum_{\{jk: \beta_{jk} \in \mathbf{B}_{g_0}\}} (|S_{jk}|/n - \lambda_{jk})_+^2} \leq \lambda_{g_0}.$$

□

A.3 Proof of Theorem 3.3

Lemma A.2 *Under the assumptions in Theorem 3.3, for any $\mathbf{B} \in \mathbb{R}^{p \times q}$, with probability at least $1 - (pq)^{1-A^2/2}$,*

$$\frac{1}{n} \|\mathbf{X}(\mathbf{B}^* - \hat{\mathbf{B}})\|_2^2 + \lambda \|\hat{\mathbf{B}} - \mathbf{B}\|_1 + 2 \sum_{g \in \mathcal{G}} \lambda_g \|\hat{\mathbf{B}}_g - \mathbf{B}_g\|_2 \quad (\text{A.5})$$

$$\leq \frac{1}{n} \|\mathbf{X}(\mathbf{B}^* - \mathbf{B})\|_2^2 + 4\lambda \sum_{jk \in J_1(\mathbf{B})} |\hat{\beta}_{jk} - \beta_{jk}| + 4 \sum_{g \in J_2(\mathbf{B})} \lambda_g \|\hat{\mathbf{B}}_g - \mathbf{B}_g\|_2,$$

$$M_1(\hat{\mathbf{B}}) \leq \frac{4}{\lambda^2 n^2} \sum_{jk \in J_1(\hat{\mathbf{B}})} |[\mathbf{X}^\top \mathbf{X}(\hat{\mathbf{B}} - \mathbf{B}^*)]_{jk}|^2 \leq \frac{4}{\lambda^2 n^2} \|\mathbf{X}^\top \mathbf{X}(\hat{\mathbf{B}} - \mathbf{B}^*)\|_2^2. \quad (\text{A.6})$$

Proof. For any $\mathbf{B} \in \mathbb{R}^{p \times q}$, we have

$$\frac{1}{n} \|\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}\|_2^2 + 2\lambda|\hat{\mathbf{B}}|_1 + \sum_{g \in \mathcal{G}} 2\lambda_g \|\hat{\mathbf{B}}_g\|_2 \leq \frac{1}{n} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_2^2 + 2\lambda|\mathbf{B}|_1 + \sum_{g \in \mathcal{G}} 2\lambda_g \|\mathbf{B}_g\|_2.$$

Plugging $\mathbf{Y} = \mathbf{X}\mathbf{B}^* + \mathbf{W}$ into the above inequality, we obtain

$$\begin{aligned} \frac{1}{n} \|\mathbf{X}(\mathbf{B}^* - \hat{\mathbf{B}})\|_2^2 &\leq \frac{1}{n} \|\mathbf{X}(\mathbf{B}^* - \mathbf{B})\|_2^2 + \frac{2}{n} \sum_{i=1}^n \sum_{k=1}^q [\mathbf{X}(\hat{\mathbf{B}} - \mathbf{B})]_{ik} \omega_{ik} \\ &\quad + 2\lambda(|\mathbf{B}|_1 - |\hat{\mathbf{B}}|_1) + \sum_{g \in \mathcal{G}} 2\lambda_g (\|\mathbf{B}_g\|_2 - \|\hat{\mathbf{B}}_g\|_2), \end{aligned}$$

where $[\mathbf{X}(\hat{\mathbf{B}} - \mathbf{B})]_{ik}$ denotes the ik^{th} element of the product matrix $\mathbf{X}(\hat{\mathbf{B}} - \mathbf{B})$ and ω_{ik} is the ik^{th} element of \mathbf{W} . Notice that

$$\begin{aligned} \sum_{i=1}^n \sum_{k=1}^q [\mathbf{X}(\hat{\mathbf{B}} - \mathbf{B})]_{ik} \omega_{ik} &= \sum_{i=1}^n \left\{ \sum_{k=1}^q \left[\sum_{j=1}^p x_{ij} (\hat{\beta}_{jk} - \beta_{jk}) \right] \omega_{ik} \right\} \\ &\leq \max_{1 \leq k \leq q, 1 \leq j \leq p} \left| \sum_{i=1}^n x_{ij} \omega_{ik} \right| \sum_{k=1}^q \sum_{j=1}^p |\hat{\beta}_{jk} - \beta_{jk}| = |\mathbf{X}^T \mathbf{W}|_\infty |\hat{\mathbf{B}} - \mathbf{B}|_1 \end{aligned}$$

where $|\mathbf{X}^T \mathbf{W}|_\infty = \max_{1 \leq k \leq q, 1 \leq j \leq p} |\sum_{i=1}^n x_{ij} \omega_{ik}|$ is the maximum absolute value of entries of $\mathbf{X}^T \mathbf{W}$.

Let $V_{jk} = \mathbf{x}_j^T \cdot \mathbf{w}_k$, $1 \leq j \leq p$, $1 \leq k \leq q$. Since $\mathbf{w}_k \sim N(0, \sigma_k^2 I_q)$ for $1 \leq k \leq q$, then $\text{var}(V_{jk}) = \mathbf{x}_p^T \text{cov}(\mathbf{w}_q) \mathbf{x}_p = n\sigma_q^2$. Therefore $(n\sigma_q^2)^{-1/2} V_{jk}$ are standard normal random variables. Consider the random event

$$\mathcal{A} = \left\{ \frac{2}{n} |\mathbf{X}^T \mathbf{W}|_\infty \leq \lambda \right\}.$$

It is easy to see that the complement of \mathcal{A} can be expressed as

$$\mathcal{A}^c = \left\{ \text{At least one } |V_{jk}| > \frac{\lambda n}{2}, 1 \leq j \leq p, 1 \leq k \leq q \right\}.$$

Denote $B(0, \lambda n/2)$ to be a 1-dimensional ball centered at 0 and with radius $\lambda n/2$, then

$$\begin{aligned} Pr\{\mathcal{A}^c\} &\leq \sum_{j=1}^p \sum_{k=1}^q Pr \left\{ V_{jk} \notin B \left(0, \frac{\lambda n}{2} \right) \right\} = p \sum_{k=1}^q Pr \left\{ (n\sigma_k^2)^{-1/2} V_{jk} \notin B \left(0, \frac{\lambda n^{1/2}}{2\sigma_k} \right) \right\} \\ &\leq pq \times Pr \left\{ |Z| \geq \frac{\lambda n^{1/2}}{2\sigma} \right\} \leq pq \exp \left(\frac{-\lambda^2 n}{8\sigma^2} \right) = (pq)^{1-A^2/2}, \end{aligned}$$

where Z is a standard normal random variable, and the last inequality is obtained by $Pr\{|Z| > a\} \leq \exp(-a^2/2)$. But on event \mathcal{A} , we have

$$\begin{aligned}
& \frac{1}{n} \|\mathbf{X}(\mathbf{B}^* - \hat{\mathbf{B}})\|_2^2 + \lambda |\hat{\mathbf{B}} - \mathbf{B}|_1 + 2 \sum_{g \in \mathcal{G}} \lambda_g \|\hat{\mathbf{B}}_g - \mathbf{B}_g\|_2 \\
& \leq \frac{1}{n} \|\mathbf{X}(\mathbf{B}^* - \mathbf{B})\|_2^2 + 2\lambda (|\hat{\mathbf{B}} - \mathbf{B}|_1 + |\mathbf{B}|_1 - |\hat{\mathbf{B}}|_1) \\
& \quad + 2 \sum_{g \in \mathcal{G}} \lambda_g (\|\hat{\mathbf{B}}_g - \mathbf{B}_g\|_2 + \|\mathbf{B}_g\|_2 - \|\hat{\mathbf{B}}_g\|_2) \\
& \leq \frac{1}{n} \|\mathbf{X}(\mathbf{B}^* - \mathbf{B})\|_2^2 + 4\lambda \sum_{jk \in J_1(\mathbf{B})} |\hat{\beta}_{jk} - \beta_{jk}| + 4 \sum_{g \in J_2(\mathbf{B})} \lambda_g (\|\hat{\mathbf{B}}_g - \mathbf{B}_g\|_2).
\end{aligned}$$

This completes the proof of the first inequality in Lemma A.2.

To prove the second inequality, we use the KKT conditions and obtain

$$\begin{cases} (1/n)[\mathbf{X}^\top(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})]_{jk} = 2\lambda \text{sgn}(\hat{\beta}_{jk}) + 2 \sum_{g \in \mathcal{G}} \lambda_g \hat{\beta}_{jk} / \|\hat{\mathbf{B}}_g\|_2, & \hat{\beta}_{jk} \neq 0; \\ (1/n)|[\mathbf{X}^\top(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})]_{jk}| \leq 2\lambda + 2 \sum_{g \in \mathcal{G}} \lambda_g, & \hat{\beta}_{jk} = 0. \end{cases}$$

From the first condition we can see that $\forall \hat{\beta}_{jk} \neq 0$,

$$\lambda \leq \frac{1}{n} |[\mathbf{X}^\top(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})]_{jk}|.$$

On the other hand, we have on \mathcal{A}

$$\begin{aligned}
\frac{1}{n} |[\mathbf{X}^\top(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})]_{jk}| & \leq \frac{1}{n} |[\mathbf{X}^\top \mathbf{X}(\mathbf{B}^* - \hat{\mathbf{B}})]_{jk}| + |[\mathbf{X}^\top \mathbf{W}]_{jk}| \\
& \leq \frac{1}{n} |[\mathbf{X}^\top \mathbf{X}(\mathbf{B}^* - \hat{\mathbf{B}})]_{jk}| + \frac{1}{n} |\mathbf{X}^\top \mathbf{W}|_\infty \\
& \leq \frac{1}{n} |[\mathbf{X}^\top \mathbf{X}(\mathbf{B}^* - \hat{\mathbf{B}})]_{jk}| + \frac{\lambda}{2}.
\end{aligned}$$

Then combine the above two inequalities, we have

$$\frac{\lambda}{2} \leq \frac{1}{n} |[\mathbf{X}^\top \mathbf{X}(\hat{\mathbf{B}} - \mathbf{B}^*)]_{jk}|.$$

Therefore

$$M_1(\hat{\mathbf{B}}) = |J_1(\hat{\mathbf{B}})| \leq \frac{4}{\lambda^2 n^2} \sum_{jk \in J_1(\hat{\mathbf{B}})} |[\mathbf{X}^\top \mathbf{X}(\hat{\mathbf{B}} - \mathbf{B}^*)]_{jk}|^2 \leq \frac{4}{\lambda^2 n^2} \|\mathbf{X}^\top \mathbf{X}(\hat{\mathbf{B}} - \mathbf{B}^*)\|_2^2.$$

This completes the proof of Lemma A.2. \square

Proof of Theorem 3.3.

By setting $\mathbf{B} = \mathbf{B}^*$ in (A.5) in Lemma A.2, we have that on event \mathcal{A} ,

$$\begin{aligned} \frac{1}{n} \|\mathbf{X}(\hat{\mathbf{B}} - \mathbf{B}^*)\|_2^2 &\leq 4\lambda \sum_{jk \in J_1(\mathbf{B}^*)} |\hat{\beta}_{jk} - \beta_{jk}^*| + 4 \sum_{g \in J_2(\mathbf{B}^*)} \lambda_g \|\hat{\mathbf{B}}_g - \mathbf{B}_g^*\|_2 \\ &\leq 4\lambda r^{1/2} \|(\hat{\mathbf{B}} - \mathbf{B}^*)_{J_1(\mathbf{B}^*)}\|_2 + 4 \left(\sum_{g \in J_2(\mathbf{B}^*)} \lambda_g^2 \right)^{1/2} \|(\hat{\mathbf{B}} - \mathbf{B}^*)_{J_2(\mathbf{B}^*)}\|_2. \end{aligned} \quad (\text{A.7})$$

The last inequality is by Cauchy-Schwarz. Specifically, we have

$$\begin{aligned} \left(\sum_{jk \in J_1(\mathbf{B}^*)} |\hat{\beta}_{jk} - \beta_{jk}^*| \right)^2 &= \left(\sum_{jk \in J_1(\mathbf{B}^*)} 1 \times |\hat{\beta}_{jk} - \beta_{jk}^*| \right)^2 \\ &\leq \left(\sum_{jk \in J_1(\mathbf{B}^*)} 1^2 \right) \left(\sum_{jk \in J_1(\mathbf{B}^*)} |\hat{\beta}_{jk} - \beta_{jk}^*|^2 \right) \\ &= r \|(\hat{\mathbf{B}} - \mathbf{B}^*)_{J_1(\mathbf{B}^*)}\|_2^2, \end{aligned}$$

and

$$\left(\sum_{g \in J_2(\mathbf{B}^*)} \lambda_g \|\hat{\mathbf{B}}_g - \mathbf{B}_g^*\|_2 \right)^2 \leq \left(\sum_{g \in J_2(\mathbf{B}^*)} \lambda_g^2 \right) \left(\sum_{g \in J_2(\mathbf{B}^*)} \|\hat{\mathbf{B}}_g - \mathbf{B}_g^*\|_2^2 \right).$$

Also by inequality (A.5), on event \mathcal{A} , we have

$$\begin{aligned} \lambda |\hat{\mathbf{B}} - \mathbf{B}^*|_1 + 2 \sum_{g \in \mathcal{G}} \lambda_g \|\hat{\mathbf{B}}_g - \mathbf{B}_g^*\|_2 \\ \leq 4\lambda \sum_{jk \in J_1(\mathbf{B}^*)} |\hat{\beta}_{jk} - \beta_{jk}^*| + 4 \sum_{g \in J_2(\mathbf{B}^*)} \lambda_g \|\hat{\mathbf{B}}_g - \mathbf{B}_g^*\|_2. \end{aligned} \quad (\text{A.8})$$

This is equivalent to

$$\begin{aligned} \lambda \sum_{jk \in J_1^c(\mathbf{B}^*)} |\hat{\beta}_{jk} - \beta_{jk}^*| + 2 \sum_{g \in J_2^c(\mathbf{B}^*)} \lambda_g \|\hat{\mathbf{B}}_g - \mathbf{B}_g^*\|_2 \\ \leq 3\lambda \sum_{jk \in J_1(\mathbf{B}^*)} |\hat{\beta}_{jk} - \beta_{jk}^*| + 2 \sum_{g \in J_2(\mathbf{B}^*)} \lambda_g \|\hat{\mathbf{B}}_g - \mathbf{B}_g^*\|_2. \end{aligned}$$

Thus the condition in Assumption 1 holds with $\Delta = \hat{\mathbf{B}} - \mathbf{B}^*$ and $\rho_g = \lambda_g/\lambda$. Therefore,

$$\|(\hat{\mathbf{B}} - \mathbf{B}^*)_{J_1(\mathbf{B}^*)}\|_2 \leq \frac{\|\mathbf{X}(\hat{\mathbf{B}} - \mathbf{B}^*)\|_2}{\kappa_1 n^{1/2}}, \quad \|(\hat{\mathbf{B}} - \mathbf{B}^*)_{J_2(\mathbf{B}^*)}\|_2 \leq \frac{\|\mathbf{X}(\hat{\mathbf{B}} - \mathbf{B}^*)\|_2}{\kappa_2 n^{1/2}}.$$

Plugging the above two inequalities into (A.7), we have

$$\begin{aligned} \frac{1}{n} \|\mathbf{X}(\hat{\mathbf{B}} - \mathbf{B}^*)\|_2^2 &\leq \left(\frac{4\lambda r^{1/2}}{\kappa_1 n^{1/2}} + \frac{4 \left(\sum_{g \in J_2(\mathbf{B}^*)} \lambda_g^2 \right)^{1/2}}{\kappa_2 n^{1/2}} \right) \|\mathbf{X}(\hat{\mathbf{B}} - \mathbf{B}^*)\|_2 \\ &= \left(\frac{4\lambda r^{1/2}}{\kappa_1 n^{1/2}} + \frac{4\lambda \left(\sum_{g \in J_2(\mathbf{B}^*)} \rho_g^2 \right)^{1/2}}{\kappa_2 \sqrt{n}} \right) \|\mathbf{X}(\hat{\mathbf{B}} - \mathbf{B}^*)\|_2, \end{aligned}$$

which gives

$$\frac{1}{n} \|\mathbf{X}(\hat{\mathbf{B}} - \mathbf{B}^*)\|_2^2 \leq 16\lambda^2 \left(\frac{r^{1/2}}{\kappa_1} + \frac{\left(\sum_{g \in J_2(\mathbf{B}^*)} \rho_g^2 \right)^{1/2}}{\kappa_2} \right)^2.$$

Define $\|\mathbf{A}\|_{2,1} = \sum_{g \in \mathcal{G}^2 \cup \mathcal{G}^1} \|\mathbf{A}\|_2$ for a matrix \mathbf{A} , where each coefficient in $\mathcal{G}^1 = \mathcal{G}_L$ forms a group. Hence

$$\|\hat{\mathbf{B}} - \mathbf{B}^*\|_{2,1} = |\hat{\mathbf{B}} - \mathbf{B}^*|_1 + \sum_{g \in \mathcal{G}} \|\hat{\mathbf{B}}_g - \mathbf{B}_g^*\|_2 \quad (\text{A.9})$$

$$\leq (c+1)|\hat{\mathbf{B}} - \mathbf{B}^*|_1. \quad (\text{A.10})$$

Then we have

$$\begin{aligned} (\lambda + \rho\lambda) \|\hat{\mathbf{B}} - \mathbf{B}^*\|_{2,1} &= \lambda \|\hat{\mathbf{B}} - \mathbf{B}^*\|_{2,1} + \rho\lambda \|\hat{\mathbf{B}} - \mathbf{B}^*\|_{2,1} \\ &\leq (c+1)\lambda |\hat{\mathbf{B}} - \mathbf{B}^*|_1 + \rho\lambda \|\hat{\mathbf{B}} - \mathbf{B}^*\|_{2,1} \quad (\text{by (A.10)}) \\ &= (c+1)\lambda |\hat{\mathbf{B}} - \mathbf{B}^*|_1 + \rho\lambda |\hat{\mathbf{B}} - \mathbf{B}^*|_1 + \sum_{g \in \mathcal{G}} \rho\lambda \|\hat{\mathbf{B}}_g - \mathbf{B}_g^*\|_2 \quad (\text{by (A.9)}) \\ &\leq (c+2)\lambda |\hat{\mathbf{B}} - \mathbf{B}^*|_1 + \sum_{g \in \mathcal{G}} \lambda_g \|\hat{\mathbf{B}}_g - \mathbf{B}_g^*\|_2 \\ &\leq (c+2)\lambda |\hat{\mathbf{B}} - \mathbf{B}^*|_1 + 2 \sum_{g \in \mathcal{G}} \lambda_g \|\hat{\mathbf{B}}_g - \mathbf{B}_g^*\|_2 \\ &\leq (c+2) \left[\lambda |\hat{\mathbf{B}} - \mathbf{B}^*|_1 + 2 \sum_{g \in \mathcal{G}} \lambda_g \|\hat{\mathbf{B}}_g - \mathbf{B}_g^*\|_2 \right] \end{aligned}$$

By (A.8) and the last inequality in (A.7) we obtain

$$\begin{aligned}
& \frac{1+\rho}{c+2} \lambda \|\hat{\mathbf{B}} - \mathbf{B}^*\|_{2,1} \\
& \leq \lambda |\hat{\mathbf{B}} - \mathbf{B}^*|_1 + 2 \sum_{g \in \mathcal{G}} \lambda_g \|\hat{\mathbf{B}}_g - \mathbf{B}_g^*\|_2 \\
& \leq 4\lambda \sum_{jk \in J_1(\mathbf{B}^*)} |\hat{\beta}_{jk} - \beta_{jk}^*| + 4 \sum_{g \in J_2(\mathbf{B}^*)} \lambda_g \|\hat{\mathbf{B}}_g - \mathbf{B}_g^*\|_2 \\
& \leq 4\lambda r^{1/2} \|(\hat{\mathbf{B}} - \mathbf{B}^*)_{J_1(\mathbf{B}^*)}\|_2 + 4 \left(\sum_{g \in J_2(\mathbf{B}^*)} \lambda_g^2 \right)^{1/2} \|(\hat{\mathbf{B}} - \mathbf{B}^*)_{J_2(\mathbf{B}^*)}\|_2 \\
& \leq \left(\frac{4\lambda r^{1/2}}{\kappa_1 n^{1/2}} + \frac{4 \left(\sum_{g \in J_2(\mathbf{B}^*)} \lambda_g^2 \right)^{1/2}}{\kappa_2 n^{1/2}} \right) \|\mathbf{X}(\hat{\mathbf{B}} - \mathbf{B}^*)\|_2 \\
& \leq \left(\frac{4\lambda r^{1/2}}{\kappa_1 n^{1/2}} + \frac{4\lambda \left(\sum_{g \in J_2(\mathbf{B}^*)} \rho_g^2 \right)^{1/2}}{\kappa_2 n^{1/2}} \right) 4n^{1/2} \lambda \left(\frac{r^{1/2}}{\kappa_1} + \frac{\left(\sum_{g \in J_2(\mathbf{B}^*)} \rho_g^2 \right)^{1/2}}{\kappa_2} \right) \\
& = 16\lambda^2 \left(\frac{r^{1/2}}{\kappa_1} + \frac{\left(\sum_{g \in J_2(\mathbf{B}^*)} \rho_g^2 \right)^{1/2}}{\kappa_2} \right)^2. \tag{5}
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|\hat{\mathbf{B}} - \mathbf{B}^*\|_{2,1} & \leq \frac{16(c+2)\lambda}{1+\rho} \left(\frac{r^{1/2}}{\kappa_1} + \frac{\left(\sum_{g \in J_2(\mathbf{B}^*)} \rho_g^2 \right)^{1/2}}{\kappa_2} \right)^2 \\
& = \frac{32(c+2)\sigma A}{1+\rho} \left(\frac{\log(pq)}{n} \right)^{1/2} \left(\frac{r^{1/2}}{\kappa_1} + \frac{\left(\sum_{g \in J_2(\mathbf{B}^*)} \rho_g^2 \right)^{1/2}}{\kappa_2} \right)^2.
\end{aligned}$$

It is trivial that $|\hat{\mathbf{B}} - \mathbf{B}^*|_1 \leq \|\hat{\mathbf{B}} - \mathbf{B}^*\|_{2,1}$.

From (A.6) in Lemma A.2, we obtain

$$M_1(\hat{\mathbf{B}}) \leq \frac{4}{\lambda^2 n^2} \|\mathbf{X}^\top \mathbf{X}(\hat{\mathbf{B}} - \mathbf{B}^*)\|_2^2 \leq \frac{4\psi_{\max}}{\lambda^2 n} \|\mathbf{X}(\hat{\mathbf{B}} - \mathbf{B}^*)\|_2^2,$$

where the second inequality is from

$$\begin{aligned} \|[\mathbf{X}^T \mathbf{X}(\hat{\mathbf{B}} - \mathbf{B}^*)]_{\cdot k}\|_2^2 &= (\hat{\mathbf{B}} - \mathbf{B}^*)_{\cdot k}^T \mathbf{X}^T (\mathbf{X} \mathbf{X}^T) \mathbf{X} (\hat{\mathbf{B}} - \mathbf{B}^*)_{\cdot k} \\ &\leq n \psi_{\max} \|\mathbf{X}(\hat{\mathbf{B}} - \mathbf{B}^*)_{\cdot k}\|_2^2 \end{aligned}$$

for each $1 \leq k \leq q$. By the upper bound of $\|\mathbf{X}(\hat{\mathbf{B}} - \mathbf{B}^*)\|_2^2$ we have

$$M_1(\hat{\mathbf{B}}) \leq 64 \psi_{\max} \left(\frac{r^{1/2}}{\kappa_1} + \frac{\left(\sum_{g \in J_2(\mathbf{B}^*)} \rho_g^2 \right)^{1/2}}{\kappa_2} \right)^2.$$

□

A.4 Proof of Proposition 4.1

To prove Proposition 4.1, we first show the following lemma.

Lemma A.3 *For every pair of (j, k) , the sequence of coordinate decent estimates $\{\hat{\beta}_{jk}^{(m)} : m = 0, 1, 2, \dots\}$ obtained at each step m by solving the following equation*

$$\hat{\beta}_{jk} = \frac{\text{sgn}(S_{jk}) (|S_{jk}| - n\lambda_{jk})_+}{\|\mathbf{x}_j\|_2^2 + n \sum_{\{g \in \mathcal{G}: \beta_{jk} \in \mathbf{B}_g, \|\hat{\mathbf{B}}_g\|_2 > 0\}} \lambda_g / (\|\hat{\mathbf{B}}_{g-(jk)}\|_2^2 + |\hat{\beta}_{jk}|^2)^{1/2}} \quad (\text{A.11})$$

for $\hat{\beta}_{jk}$ while fixing others converges to a global minimizer of the objective function (2) in the main text.

First, it is easy to see that the exact solution of (A.11) exists. If $\|\hat{\mathbf{B}}_{g-(jk)}\|_2 = 0$, the close form solution of (A.11) is just the lasso solution. If $\|\hat{\mathbf{B}}_{g-(jk)}\|_2 \neq 0$, then the right hand side of (A.11) is a continuous function of $\hat{\beta}_{jk}$, which is monotone when $\hat{\beta}_{jk} > 0$ or $\hat{\beta}_{jk} < 0$, bounded away from zero when $\hat{\beta}_{jk} = 0$, and bounded away from $\pm\infty$ when $\hat{\beta}_{jk}$ goes to $\pm\infty$, therefore must intersect with either $y = \hat{\beta}_{jk}$ or $y = -\hat{\beta}_{jk}$. Therefore an exact solution of (A.11) must exist.

Wu and Lange (2008) proved the convergence to a minimal point of the lasso objective function for the greedy coordinate descent algorithm. In a very similar way, one can extend the proof to the multivariate sparse group lasso objective function and the coordinate descent algorithm of iteratively solving for the exact solution of (A.11). Due to significant overlapping with Wu and Lange (2008), we omit the proof of Lemma A.3 here.

Proof of Proposition 4.1.

Denote $\{\hat{\beta}_{jk}^{(m)}\}$ the sequence of estimates of jk^{th} coordinate from the coordinate descent algorithm that solves equation (A.11) in each step indexed by m . Starting from $\hat{\mathbf{B}}^{(m-1)}$, denote $\hat{\beta}_{jk}^{\text{MCD}(m-1)}$ the one step update of the jk^{th} coordinate by the mixed coordinate descent algorithm. We prove in the following that

$$|\hat{\beta}_{jk}^{(m)}| \leq |\hat{\beta}_{jk}^{\text{MCD}(m)}| \leq |\hat{\beta}_{jk}^{(m-1)}| \quad (\text{A.12})$$

with equalities hold only when $|\hat{\beta}_{jk}^{(m)}| = |\hat{\beta}_{jk}^{(m-1)}|$.

First, if $\hat{\beta}_{jk}^{(m)}$ is updated by (3) in the main text, then (A.12) is automatically satisfied since

$$0 = |\hat{\beta}_{jk}^{(m)}| = |\hat{\beta}_{jk}^{\text{MCD}(m)}| \leq |\hat{\beta}_{jk}^{(m-1)}|$$

Otherwise, if $\hat{\beta}_{jk}^{(m)}$ is updated by (II) or (III) or (IV) in Section 4 of the main text, then it must be one of the following cases.

(i) If

$$\hat{\beta}_{jk}^{(m-1)} < \hat{\beta}_{jk}^{(m)} = \frac{-(|S_{jk}| - n\lambda_{jk})_+}{\|\mathbf{x}_j\|_2^2 + n \sum_{\{g \in \mathcal{G}: \beta_{jk} \in \mathbf{B}_g, \|\hat{\mathbf{B}}_g^{(m-1)}\|_2 > 0\}} \lambda_g / (\|\hat{\mathbf{B}}_{g-(jk)}^{(m-1)}\|_2^2 + |\hat{\beta}_{jk}^{(m)}|^2)^{1/2}} < 0,$$

then

$$\hat{\beta}_{jk}^{\text{MCD}(m)} = \frac{-(|S_{jk}| - n\lambda_{jk})_+}{\|\mathbf{x}_j\|_2^2 + n \sum_{\{g \in \mathcal{G}: \beta_{jk} \in \mathbf{B}_g, \|\hat{\mathbf{B}}_g^{(m-1)}\|_2 > 0\}} \lambda_g / (\|\hat{\mathbf{B}}_{g-(jk)}^{(m-1)}\|_2^2 + |\hat{\beta}_{jk}^{(m-1)}|^2)^{1/2}} < \hat{\beta}_{jk}^{(m)}.$$

From the proof of Theorem 3.1, $\hat{\beta}_{jk}^{(m-1)} < \hat{\beta}_{jk}^{(m)}$ if and only if

$$\left. \frac{\partial L(\mathbf{B})}{\partial \beta_{jk}} \right|_{\hat{\beta}_{jk}^{(m-1)}} = -S_{jk}/n + \|\mathbf{x}_j\|_2^2 \hat{\beta}_{jk}^{(m-1)}/n - \lambda_{jk} + \sum_{\mathcal{G}_{jk}} \lambda_g \hat{\beta}_{jk}^{(m-1)} / \|\hat{\mathbf{B}}_g^{(m-1)}\|_2 < 0.$$

Notice that the above is also the partial derivative of $L^{\text{net}}(\mathbf{B})$ w.r.t. β_{jk} taking value at $\hat{\beta}_{jk}^{(m-1)}$, where $L^{\text{net}}(\mathbf{B})$ is the elastic net objective function

$$L^{\text{net}}(\mathbf{B}) = \frac{1}{2n} \|\mathbf{Y} - \mathbf{XB}\|_2^2 + \sum_{jk} \lambda_{jk} |\beta_{jk}| + \sum_{g \in \mathcal{G}} \lambda_g \|B_g\|_2^2 / (2 \|\hat{\mathbf{B}}_g^{(m-1)}\|_2)$$

holding $\|\hat{\mathbf{B}}_g^{(m-1)}\|_2$ as constants and constraining that $\beta_{jk} < 0$.

Following exactly the same argument as in the proof of Theorem 3.1, we can prove that $\left. \frac{\partial L^{\text{net}}(\mathbf{B})}{\partial \beta_{jk}} \right|_{\hat{\beta}_{jk}^{(m-1)}} < 0$ if and only if $\hat{\beta}_{jk}^{(m-1)}$ is less than the solution of $\partial L(\mathbf{B})/\partial \beta_{jk} = 0$ with the constraint $\beta_{jk} < 0$, which is the solution of

$$-S_{jk}/n + \|\mathbf{x}_j\|_2^2 \beta_{jk}/n - \lambda_{jk} + \sum_{\mathcal{G}_{jk}} \lambda_g \beta_{jk} / \|\hat{\mathbf{B}}_g^{(m-1)}\|_2 = 0$$

under $\beta_{jk} < 0$ given by

$$\frac{-(|S_{jk}| - n\lambda_{jk})_+}{\|\mathbf{x}_j\|_2^2 + n \sum_{\{g \in \mathcal{G}: \beta_{jk} \in \mathbf{B}_g, \|\hat{\mathbf{B}}_g^{(m-1)}\|_2 > 0\}} \lambda_g / \|\hat{\mathbf{B}}^{(m-1)}\|_2} = \hat{\beta}_{jk}^{\text{MCD}(m)}.$$

Therefore, we have

$$\hat{\beta}_{jk}^{(m-1)} < \hat{\beta}_{jk}^{\text{MCD}(m)} < \hat{\beta}_{jk}^{(m)} < 0.$$

(ii) If

$$\hat{\beta}_{jk}^{(m-1)} > \hat{\beta}_{jk}^{(m)} = \frac{(|S_{jk}| - n\lambda_{jk})_+}{\|\mathbf{x}_j\|_2^2 + n \sum_{\{g \in \mathcal{G}: \beta_{jk} \in \mathbf{B}_g, \|\hat{\mathbf{B}}_g^{(m-1)}\|_2 > 0\}} \lambda_g / (\|\hat{\mathbf{B}}_{g-(jk)}^{(m-1)}\|_2^2 + |\hat{\beta}_{jk}^{(m)}|^2)^{1/2}} \geq 0,$$

with similar argument, we have that

$$\hat{\beta}_{jk}^{(m-1)} > \hat{\beta}_{jk}^{\text{MCD}(m)} > \hat{\beta}_{jk}^{(m)} \geq 0.$$

(iii) If $\hat{\beta}_{jk}^{(m-1)} = \hat{\beta}_{jk}^{(m)}$, the mixed coordinate descent algorithm will be exact update and we will have

$$\hat{\beta}_{jk}^{(m-1)} = \hat{\beta}_{jk}^{\text{MCD}(m)} = \hat{\beta}_{jk}^{(m)}.$$

In summary, we have (A.12).

Lemma A.3 shows that the sequence of estimates of jk^{th} coordinate $\{\hat{\beta}_{jk}^{(m)}\}$ iteratively updated from solving (A.11) converges to a global minimizer regardless the value of the starting point. For each term in the sequence $\{\hat{\beta}_{jk}^{\text{MCD}(l)}\}$, suppose one can construct a sequence of $\{\hat{\beta}_{jk}^{(m)}\}$ starting from $\hat{\beta}_{jk}^{\text{MCD}(l)}$, then those sequences all converge to minimizers (if the minimizer is not unique, e.g. for not strictly convex objective function) with the same minimum value. Thus from (A.12) we know that $\{\hat{\beta}_{jk}^{\text{MCD}(l)}\}$ converge to a global minimizer with the same minimum value. \square

Figure A.1 illustrates coordinate updates by the standard coordinate descent and the mixed coordinate descent algorithms on a contour surface of a two-dimensional objective function. Given the same starting values, one step update on one coordinate from the mixed coordinate descent algorithm is always bounded between the previous and current values from the standard coordinate descent algorithm.

A.5 Comparison of computing costs

The computational cost of coordinate descent algorithm with inner iterations is much higher than our mixed coordinate descent algorithm. Figure A.2 shows the comparison between these two algorithms. The group structure of the regression coefficient matrix used is set to be (b) in Figure 1 in the main text. In Figure A.2, the mixed coordinate descent algorithm converges to a minimizer after 500 iterations while the coordinate descent algorithm with inner iterations converges after 150000 iterations.

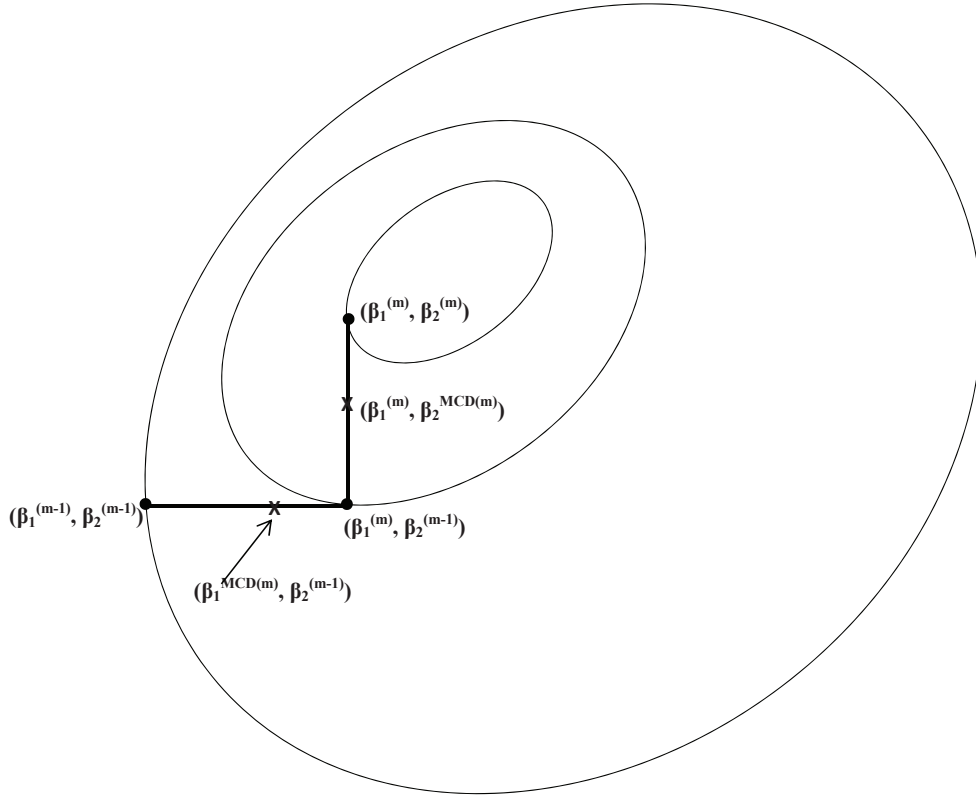


Figure A.1: Illustration of coordinate updates by the standard coordinate descent and the mixed coordinate descent algorithms on a contour surface of a two-dimensional objective function.

B Comparison between univariate and multivariate approaches

Figures B.1 and B.2 illustrate the comparisons between univariate approaches and multivariate approaches. The true regression coefficient matrix takes a $\mathcal{G}_{XY} \cup \mathcal{G}_X$ group structure. It can be seen that when different response variables have a similar sparsity to the predictors, the multiple univariate lasso (using different λ values for different response variables) and the multivariate lasso (using the same λ value for all response variables) have similar performance on variable selection. The multiple univariate sparse group lasso approach has

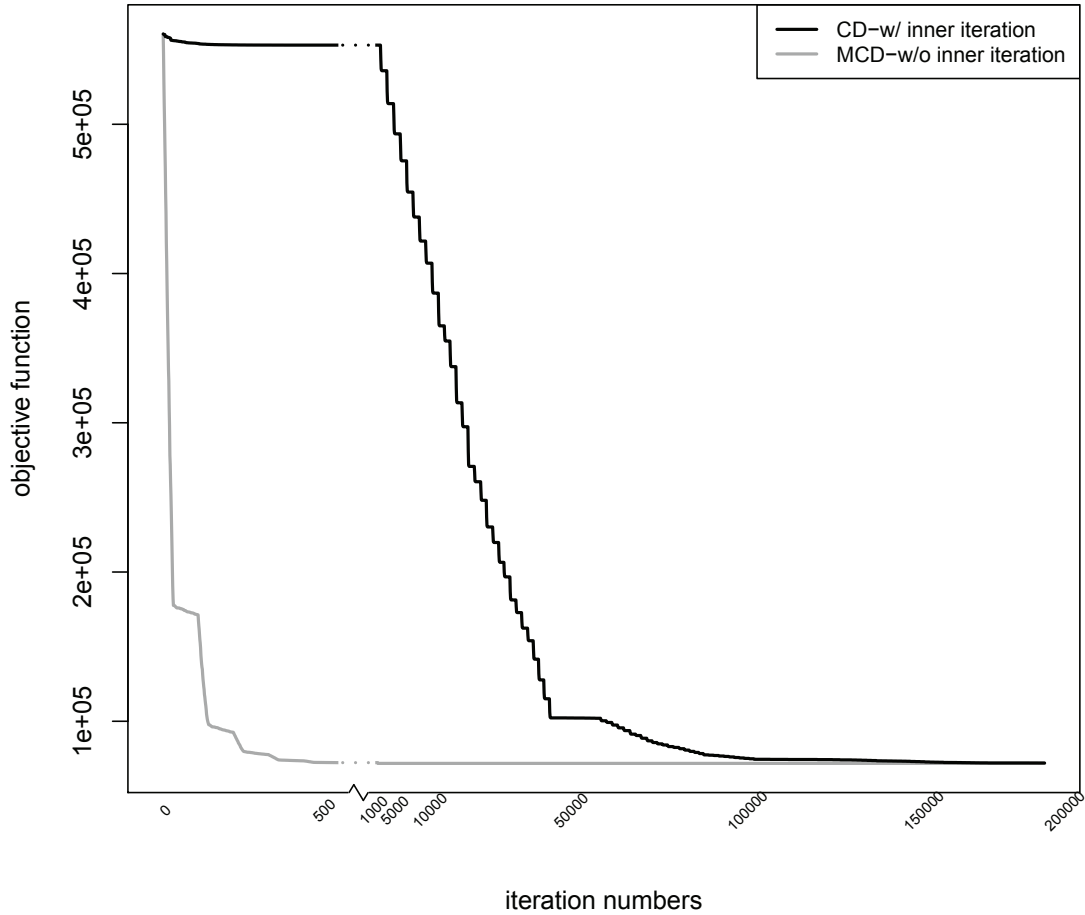


Figure A.2: Decreasing of the objective function. The gray line is for the proposed mixed coordinate descent (MCD) algorithm without inner iterations of updating (A.11) and the black line is for the coordinate descent (CD) algorithm with inner iterations.

a slightly better variable selection performance than the multiple univariate lasso. The proposed multivariate sparse group lasso yields the best variable selection result by borrowing information from other response variables within the same group. It also has the smallest prediction error.

C More simulations

Figures C.1 to C.5 show the variable selection and prediction effects in some other simulation settings, such as with different autocorrelation coefficient values or with a true “all-in-all-out”

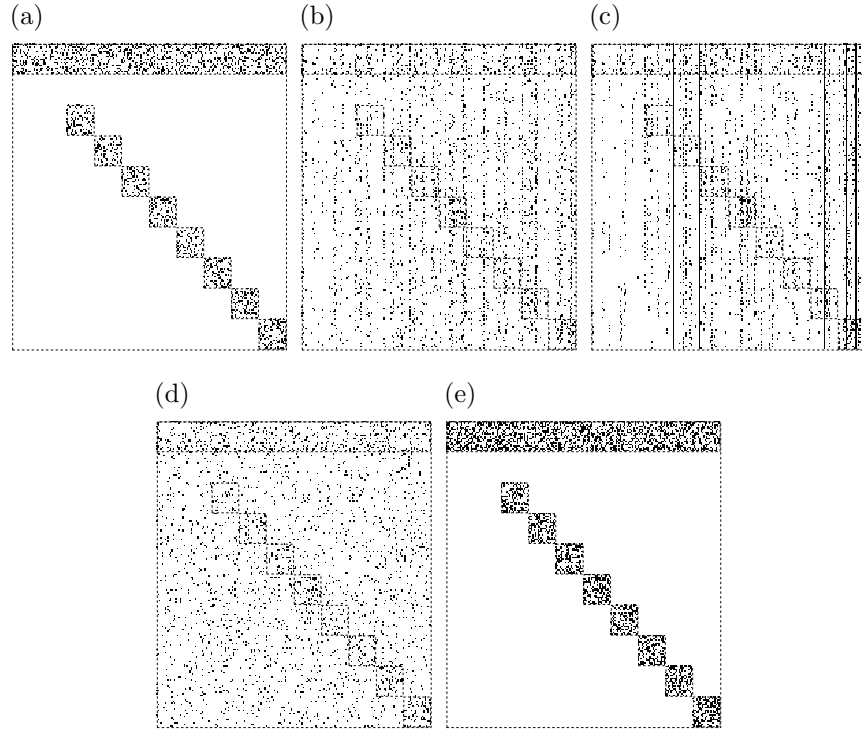


Figure B.1: Heatmaps of coefficient matrices. (a) True B^* ; (b) The multiple univariate lasso; (c) The multiple univariate sparse group lasso (d) The multivariate lasso; (e) The multivariate sparse group lasso; The true B^* has a “not all in all out” and $X+XY$ group structure with $p = q = 200$, $n = 100$, $\rho = 0.5$.

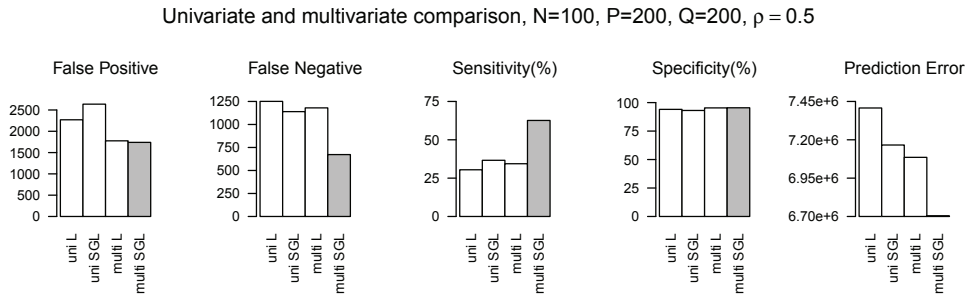


Figure B.2: Comparison between multiple-univariate and multivariate approaches from 100 simulated data sets. “uni L” – the multiple univariate lasso; “uni SGL” – the multiple univariate group lasso; “multi L” – the multivariate lasso; “multi SGL” – the multivariate sparse group lasso with an XY group structure on the coefficient matrix.

group structure.

D Network structure for the yeast eQTL data analysis

Figure D.1 shows the network constructed from the multivariate sparse group lasso method. The top association signals are highlighted in dark lines and also reported in Table 2 and 3 in the main text.

References

Wu, T. and Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *Annals of Applied Statistics*, 2:224–44.

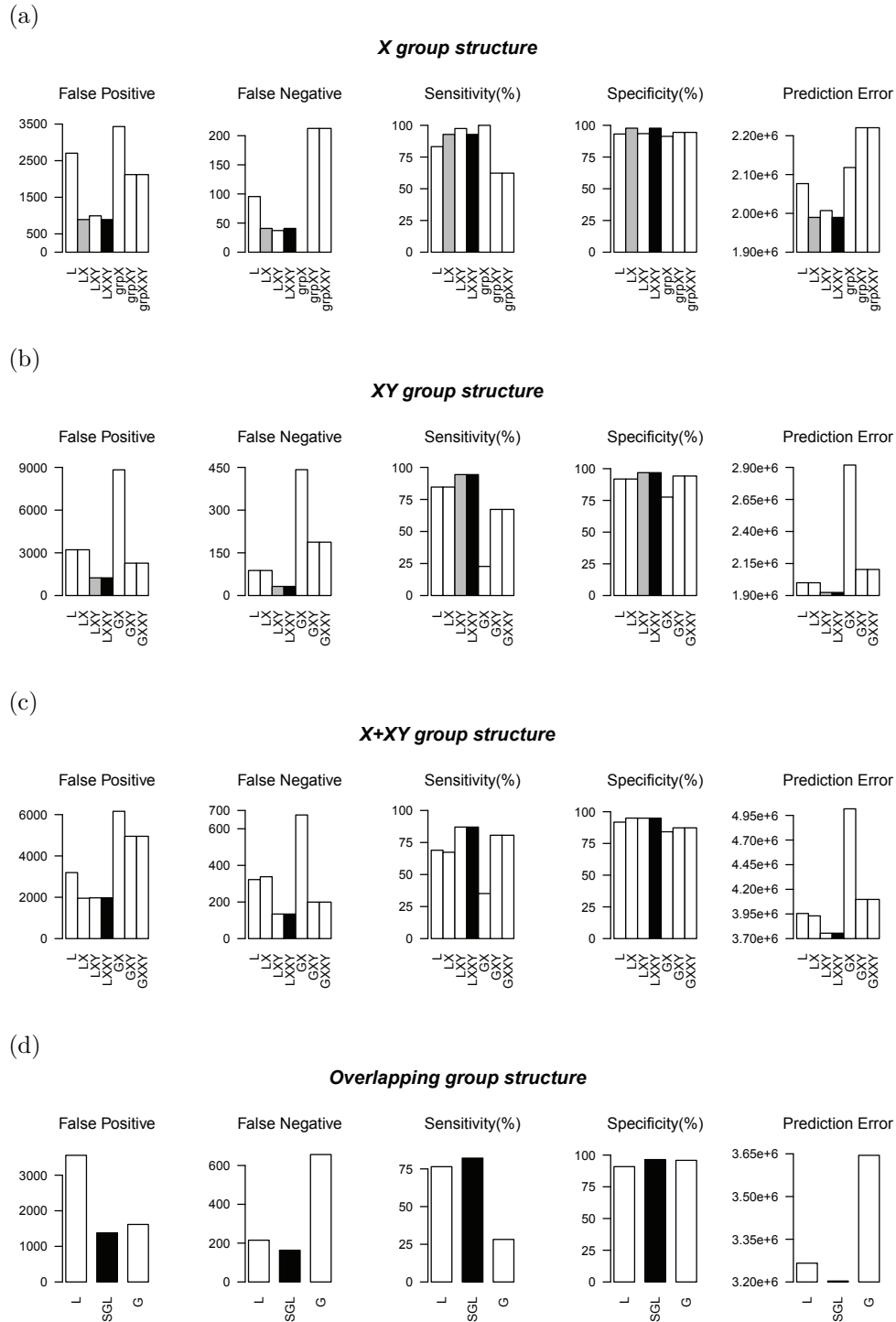


Figure C.1: More simulation results, “not all in all out” cases with $n = 150$, $p = q = 200$ and $\rho = 0.2$.

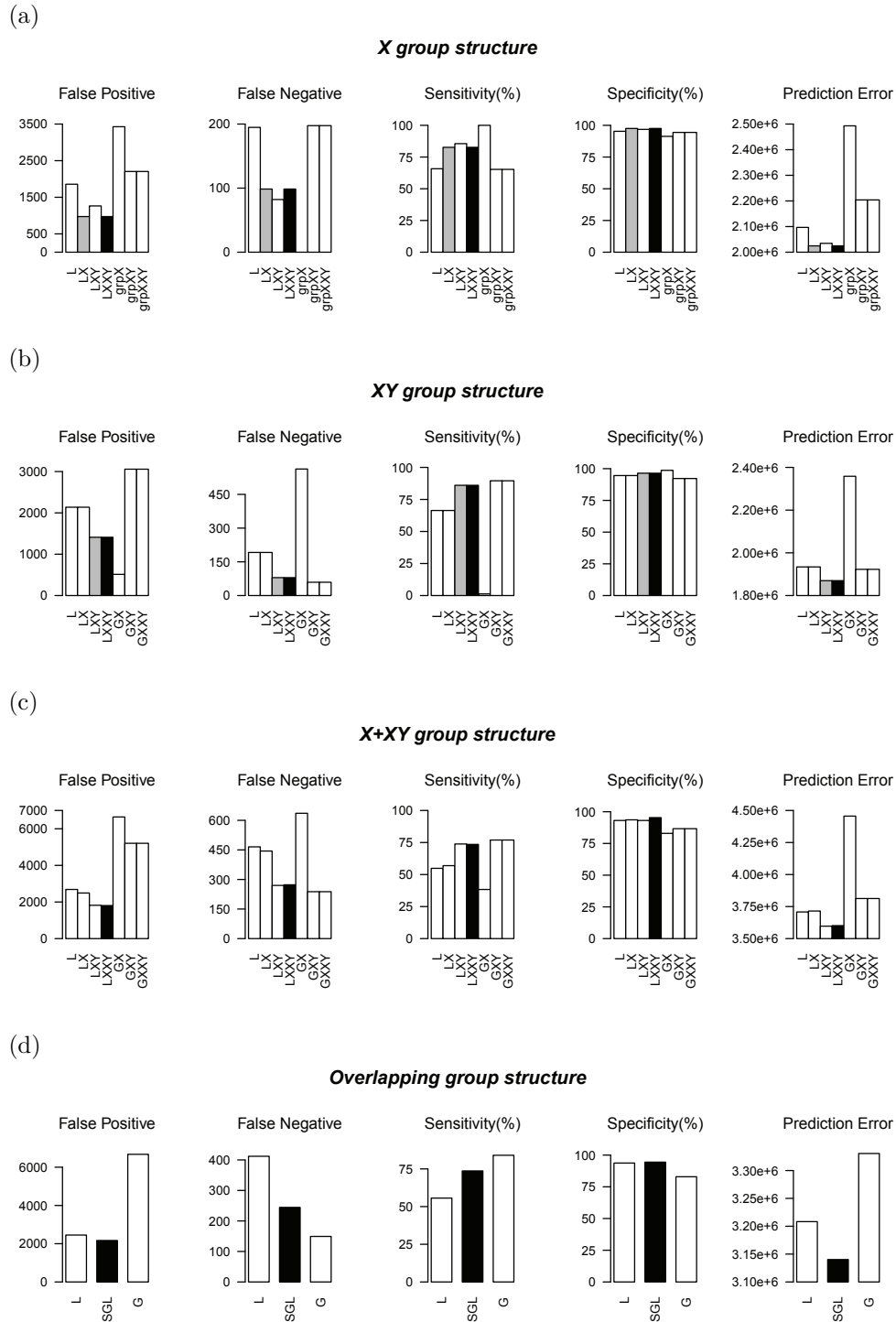


Figure C.2: More simulation results, “not all in all out” cases with $n = 150$, $p = q = 200$ and $\rho = 0.8$.

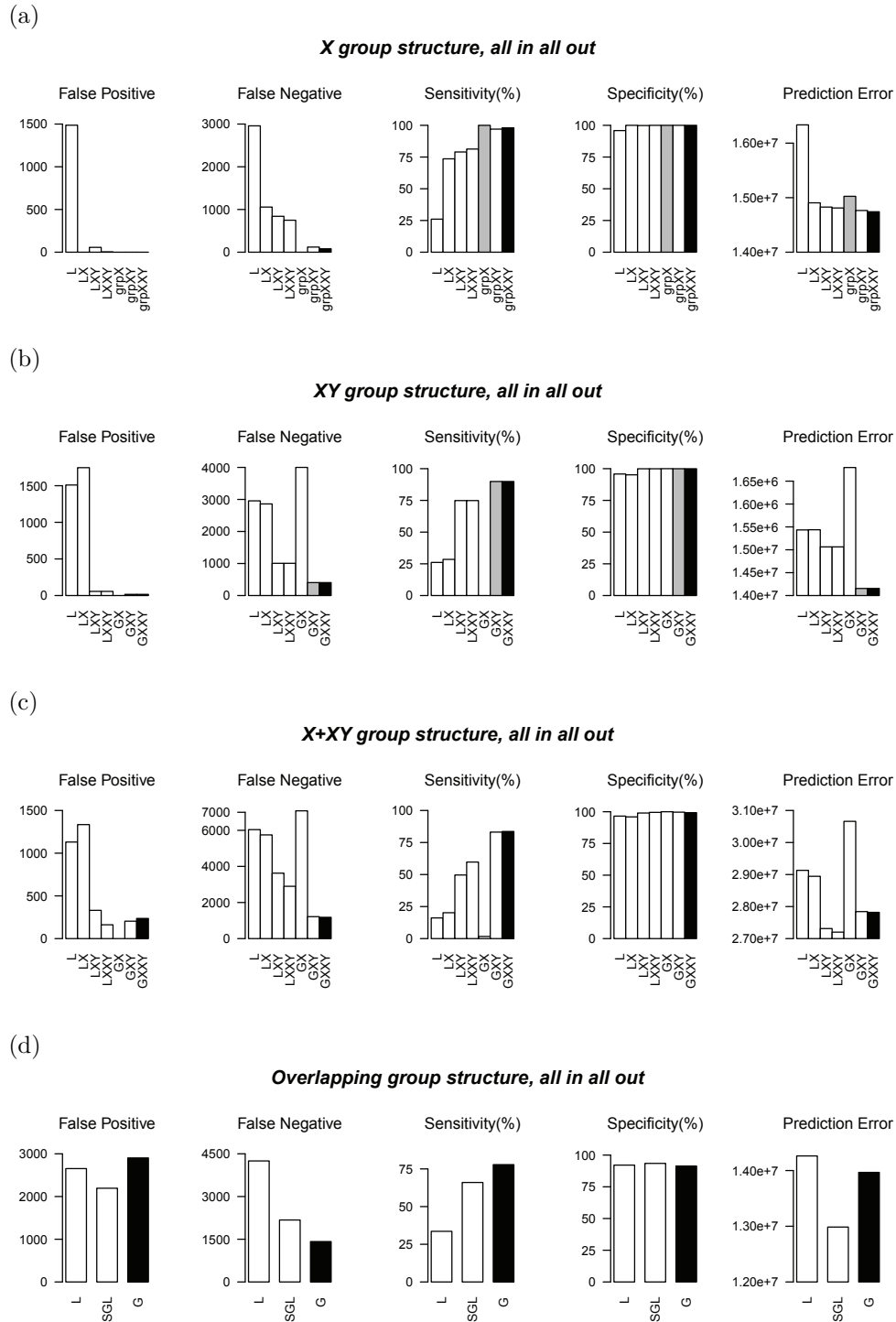


Figure C.3: More simulation results, “all in all out” cases with $n = 150$, $p = q = 200$ and $\rho = 0.5$.

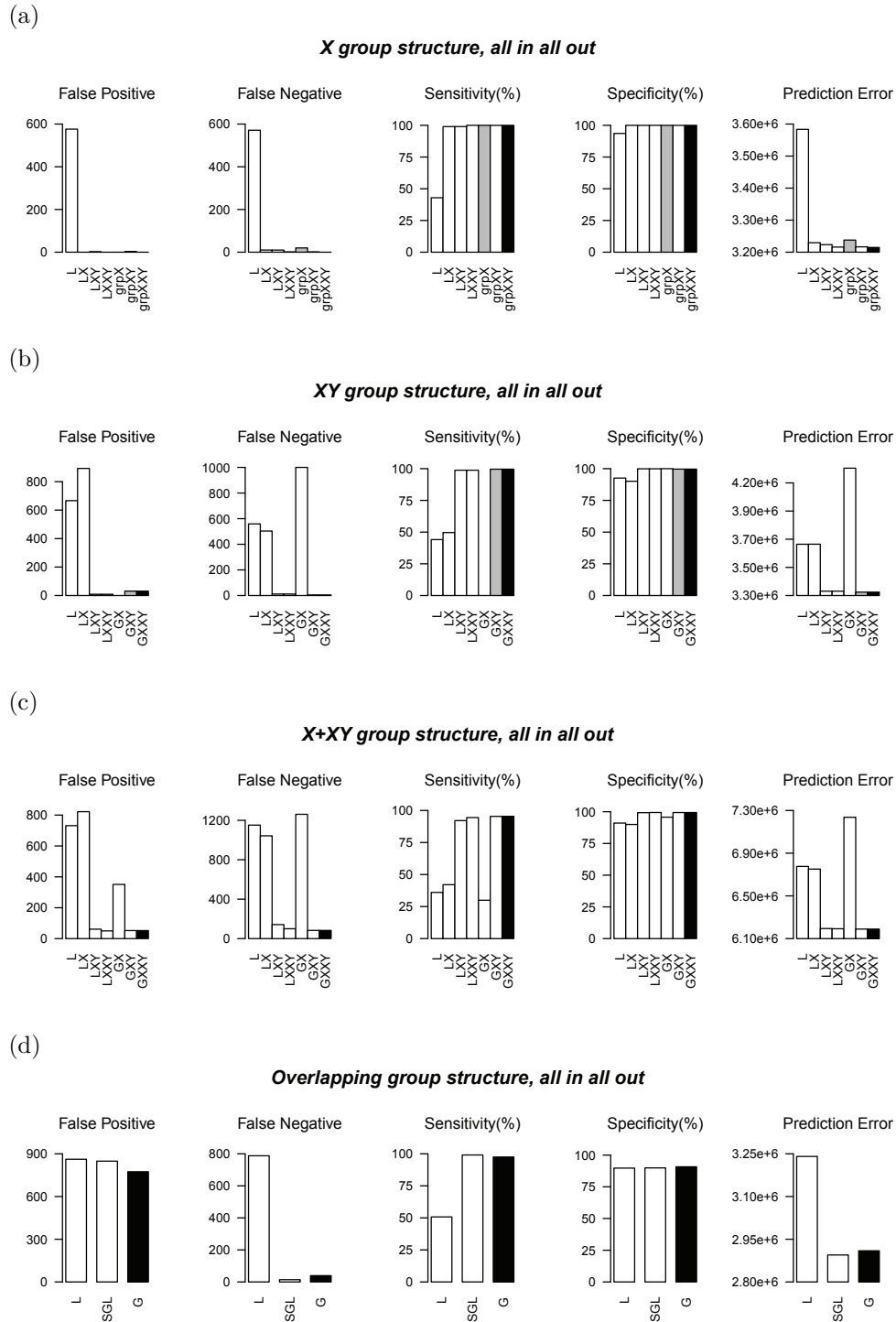


Figure C.4: More simulation results, “all in all out” cases with $n = 150$, $p = q = 100$ and $\rho = 0.5$.

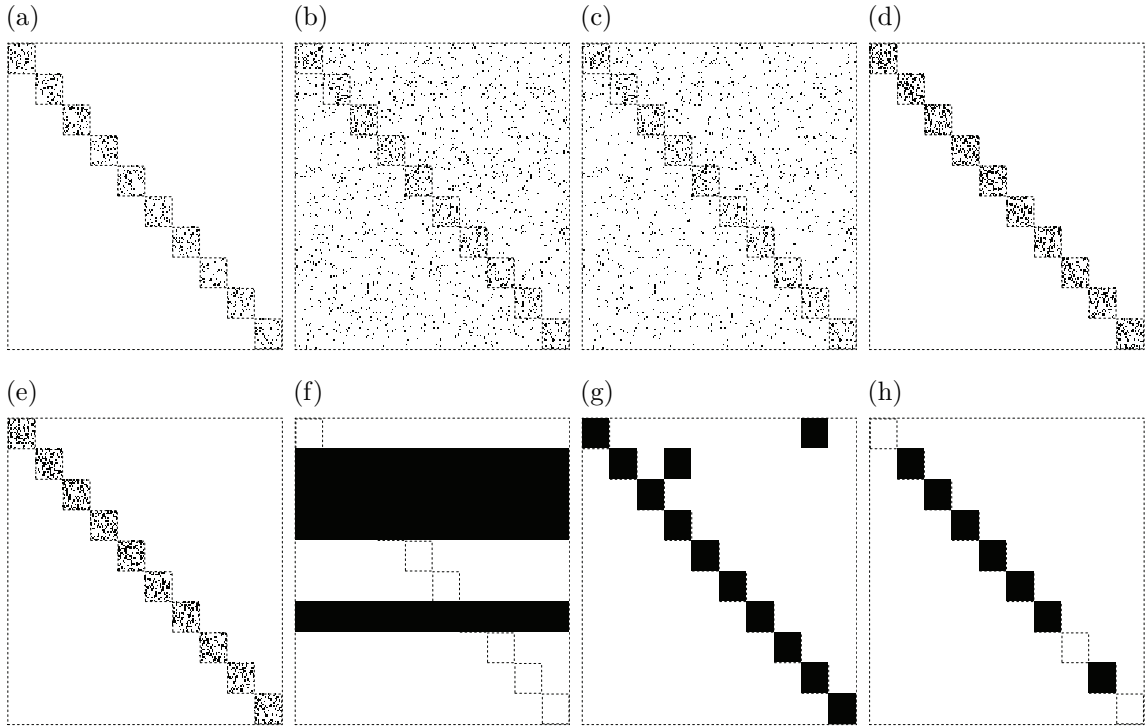


Figure C.5: Heatmaps of coefficient matrices, selection effects. “Not all in all out” XY group structure with $n = 100$, $p = 200$, $q = 200$, and $\rho = 0.5$. (a) \mathbf{B}^* ; (b) $\hat{\mathbf{B}}_L$; (c) $\hat{\mathbf{B}}_{LX}$; (d) $\hat{\mathbf{B}}_{LXY}$; (e) $\hat{\mathbf{B}}_{LXXY}$; (f) $\hat{\mathbf{B}}_{GX}$; (g) $\hat{\mathbf{B}}_{GXY}$; (h) $\hat{\mathbf{B}}_{GXXY}$.

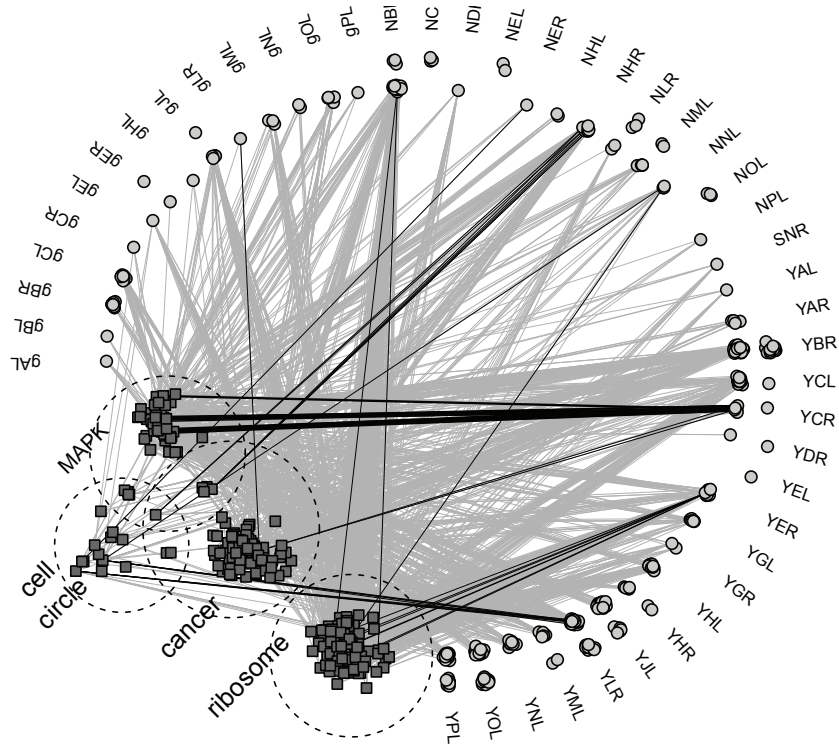


Figure D.1: Network constructed from the multivariate sparse group lasso method. Network structure is between gene expressions grouped in *mitogen-activated protein kinases (MAPK)*, *cell cycle*, *cancer*, *ribosome* pathways and markers grouped in 45 gene groups. Gray lines connect expression-marker pairs with non-zero $\hat{\beta}_{jk}$. Dark lines are for the top 10 associations in each pathways. The strength of these top associations are indicated by the width of the dark lines. The dotted circles indicate the overlapping pathway group structure.