Supplementary Materials for "REGULATION-INCORPORATED GENE EXPRESSION NETWORK-BASED HETEROGENEITY ANALYSIS"

Rong Li[1], Qingzhao Zhang[2], Shuangge Ma[1]

[1] *Yale University*

[2] *Xiamen University*

S1 Additional Details on Methods

Small example in Section 1 It is first noted that this small example serves to provide a simple and intuitive presentation of the proposed method. All definitive conclusions are based on the simulation results in Section 3, which have larger sizes, more realistic settings, and more replicates. For this small example, we set $p = 8$ and $q = 1$. The one regulator is set to be associated with all eight gene expressions. We first consider when the sample grouping is known. Then the network structures, without and with accounting for the regulations, can be directly generated and plotted. This result is provided in Figure 1 of the main text.

We further simulate 100 replicates and conduct clustering analysis with the proposed and various alternatives. The Rand index results are summarized in Table S1. The proposed approach is observed to be superior. This finding is in line with those

in the main text, and we refer to the main text for more discussions.

Table S1: Simulation results for the small example in Section 1. In each cell, mean Rand index.

	$K = 2$	$K = 4$
K-means	0.441	0.236
nonp-mix	0.441	0.249
HeteroGGM	0.248	0.285
Proposed	0.883	0.874

Comparison of different methods The proposed approach shares certain technical components with Hao et al. (2018) and Radchenko and Mukherjee (2017). Their key differences are briefly summarized in Table S2. More discussions are provided in the main text.

Table S2: Comparison of different methods.

	Component	Known number of components	Technique
Hao et al. (2018)	GGM	Yes	mixture model
Radchenko and Mukherjee (2017)	Mean	No	fusion
Proposed	CGGM	No	mixture model + fusion

Flowchart The steps and rationale of the proposed approach are described in detail in the main text. In Figure S1, we present a brief flowchart which may assist a better understanding.

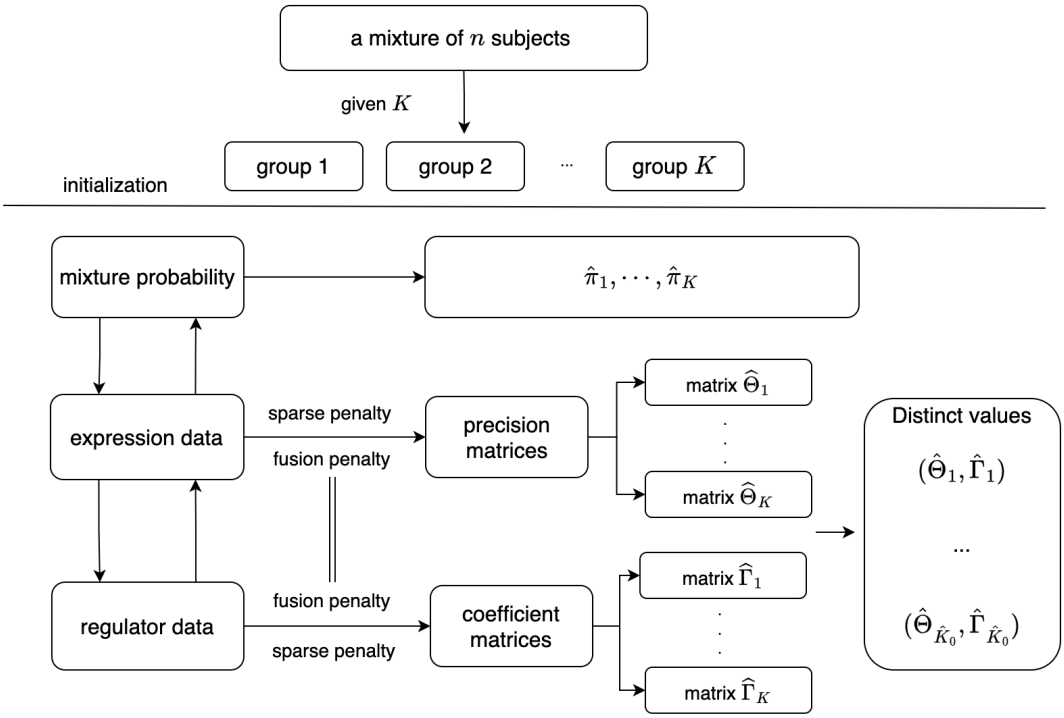

Figure S1: Flowchart of the proposed method.

S2 Additional Details on Computation

We show the detailed calculations to update the estimate of $\boldsymbol{\Theta}^{(t)}$ and $\boldsymbol{\Gamma}^{(t)}$, solving for the optimal solution of equations (2.6) and (2.7) in the t-th maximization step in the EM algorithm.

Update of $\{\boldsymbol{\Gamma}\}$ in the EM algorithm

Recall that maximizing the conditional expectation with respect to $\boldsymbol{\Gamma}$ in the t-th step is equivalent to solving:

$$\begin{aligned}\left\{\boldsymbol{\Gamma}^{(t)}\right\} = \arg\min_{\boldsymbol{\Gamma}} \Bigg(& \frac{1}{2n}\sum_{i=1}^{n}\sum_{l=1}^{K} L_{il}^{(t)} \left\{ (\boldsymbol{y}_i - \boldsymbol{\Gamma}_l \boldsymbol{x}_i)^T \boldsymbol{\Theta}_l^{(t-1)} (\boldsymbol{y}_i - \boldsymbol{\Gamma}_l \boldsymbol{x}_i) \right\} \\ & + \sum_{l=1}^{K}\sum_{j=1}^{p}\sum_{m=1}^{q+1} p(|\gamma_{jm,l}|, \lambda_2) + \sum_{l<l\prime} p\left((\|\boldsymbol{\Theta}_l^{(t-1)} - \boldsymbol{\Theta}_{l\prime}^{(t-1)}\|_F^2 + \|\boldsymbol{\Gamma}_l - \boldsymbol{\Gamma}_{l\prime}\|_F^2)^{1/2}, \lambda_3 \right) \Bigg).\end{aligned}$$

With the local quadratic approximation, it can be rewritten as:

$$\begin{aligned}\left\{\boldsymbol{\Gamma}^{(t)}\right\} = \arg\min_{\boldsymbol{\Gamma}} \Bigg(& \frac{1}{2n}\sum_{i=1}^{n}\sum_{l=1}^{K} L_{il}^{(t)} \left\{ (\boldsymbol{y}_i - \boldsymbol{\Gamma}_l \boldsymbol{x}_i)^T \boldsymbol{\Theta}_l^{(t-1)} (\boldsymbol{y}_i - \boldsymbol{\Gamma}_l \boldsymbol{x}_i) \right\} \\ & + \sum_{l=1}^{K}\sum_{j=1}^{p}\sum_{m=1}^{q+1} \frac{1}{2}\frac{p'(|\gamma_{jm,l}^{(t-1)}|, \lambda_2)}{|\gamma_{jm,l}^{(t-1)}|}\gamma_{jm,l}^2 + \sum_{l<l\prime}\frac{1}{2}\frac{p'(\tau_{ll\prime}^{(t-1,t-1)}, \lambda_3)}{\tau_{ll\prime}^{(t-1,t-1)}}\|\boldsymbol{\Gamma}_l - \boldsymbol{\Gamma}_{l\prime}\|_F^2) \Bigg).\end{aligned}$$

The update of $\boldsymbol{\Gamma}^{(t)}$ is as follows. For $j = 1, \ldots, p$, $q = 1, \ldots, q+1$ and $l = 1 \ldots, K$,

$$\gamma_{jm,l}^{(t)} = \frac{h_{jm,l}^{(t-1)} + n\tilde{v}_{jm,l}^{(t-1)}}{n\hat{v}_{jm,l}^{(t-1)} + n_l\theta_{jj,l}^{(t-1)}C_{xl,mm}^{(t-1)} + np'\left(|\gamma_{jm,l}^{(t-1)}|, \lambda_2\right)/|\gamma_{jm,l}^{(t-1)}|},$$

where

$$\begin{aligned}h_{jm,l}^{(t-1)} = & -n_l^{(t)}\left[\sum_{g\neq j}^{p}\theta_{gj,l}^{(t-1)}\left(\sum_{h=1}^{q}C_{xl,mh}^{(t-1)}\gamma_{gh,l}^{(t-1)}\right) + \theta_{jj,l}^{(t-1)}\left(\sum_{h\neq m}^{q}C_{xl,mh}^{(t-1)}\gamma_{gh,l}^{(t-1)}\right)\right] \\ & + n_l\boldsymbol{e}_m^T\left(\mathbf{C}_{yx,l}^{(t-1)T}\boldsymbol{\Theta}_l^{(t-1)}\right)\boldsymbol{e}_j,\end{aligned}$$

$$\tilde{v}_{jm,l}^{(t-1)} = \sum_{l<l'} \frac{p'(\tau_{ll'}^{(t-1,t-1)}, \lambda_3)}{\tau_{ll'}^{(t-1,t-1)}} \gamma_{jm,l}^{(t-1)},$$

$$\hat{v}_{jm,l}^{(t-1)} = \sum_{l<l'} \frac{p'(\tau_{ll'}^{(t-1,t-1)}, \lambda_3)}{\tau_{ll'}^{(t-1,t-1)}},$$

$$\tau_{ll'}^{(t-1,t-1)} = \left(\|\boldsymbol{\Theta}_l^{(t-1)} - \boldsymbol{\Theta}_{l'}^{(t-1)}\|_F^2 + \|\boldsymbol{\Gamma}_l^{(t-1)} - \boldsymbol{\Gamma}_{l'}^{(t-1)}\|_F^2 \right)^{1/2},$$

the weighted covariance matrices $\mathbf{C}_{yl}^{(t-1)} = \sum_{i=1}^n L_{il}^{(t-1)} \boldsymbol{y}_i \boldsymbol{y}_i^T / \sum_{i=1}^n L_{il}^{(t-1)}$, $\mathbf{C}_{yx,l}^{(t-1)} = \sum_{i=1}^n L_{il}^{(t-1)} \boldsymbol{y}_i \boldsymbol{x}_i^T / \sum_{i=1}^n L_{il}^{(t-1)}$, $\mathbf{C}_{xl}^{(t-1)} = \sum_{i=1}^n L_{il}^{(t-1)} \boldsymbol{x}_i \boldsymbol{x}_i^T / \sum_{i=1}^n L_{il}^{(t-1)}$ with $C_{xl,mh}^{(t-1)}$ as its mh-th entry, and $n_l^{(t)} = \sum_{i=1}^n L_{il}^{(t)}$. $\boldsymbol{e}_m$ denotes a $(q+1) \times 1$ vector with the m-th entry being 1 and the others being 0, and $\boldsymbol{e}_j$ denotes a $p \times 1$ vector with the j-th entry being 1 and the others being 0.

Update of $\{\boldsymbol{\Theta}\}$ in the EM algorithm

Recall that maximizing the conditional expectation with respect to $\boldsymbol{\Theta}$ in the t-th step is equivalent to solving:

$$\begin{aligned}\left\{\boldsymbol{\Theta}^{(t)}\right\} = \arg\min_{\boldsymbol{\Theta}} \Bigg\{ & \sum_{l=1}^K n_l^{(t)} \left[-\log\det(\boldsymbol{\Theta}_l) + \operatorname{tr}(\mathbf{S}_{\Gamma l}^{(t)} \boldsymbol{\Theta}_l) \right] \\ & + \sum_{l=1}^K \sum_{j \neq m} p(|\theta_{jm,l}|, \lambda_1) + \sum_{l<l'} p\left((\|\boldsymbol{\Theta}_l - \boldsymbol{\Theta}_{l'}\|_F^2 + \|\boldsymbol{\Gamma}_l^{(t)} - \boldsymbol{\Gamma}_{l'}^{(t)}\|_F^2)^{1/2}, \lambda_3 \right) \Bigg\},\end{aligned}$$

where $\mathbf{S}_{\Gamma l}^{(t)} = \mathbf{C}_{yl}^{(t)} - \mathbf{C}_{yx,l}^{(t)} \boldsymbol{\Gamma}_l^{(t)T} - \boldsymbol{\Gamma}_l^{(t)} \mathbf{C}_{yx,l}^{(t)T} + \boldsymbol{\Gamma}_l^{(t)} \mathbf{C}_{xl}^{(t)} \boldsymbol{\Gamma}_l^{(t)T}$, and $\mathbf{C}_{yl}^{(t)}$, $\mathbf{C}_{xl}^{(t)}$, and $\mathbf{C}_{yx,l}^{(t)}$ are the weighted covariance matrices based on $L_{il}^{(t)}$. This can be solved using the ADMM algorithm (Danaher et al., 2014; Wang and Jiang, 2020), which can be reformulated

as:

$$\arg\min_{\boldsymbol{\Theta},\boldsymbol{\Xi}} \left\{ \sum_{l=1}^{K} n_l^{(t)} \left[-\log\det(\boldsymbol{\Theta}_l) + \mathrm{tr}(\mathbf{S}_{\Gamma l}^{(t)}\boldsymbol{\Theta}_l) \right] + \sum_{l=1}^{K}\sum_{j\neq m} p(|\xi_{jm,l}|,\lambda_1) \right.$$
$$\left. + \sum_{l<l'} p\left((\|\boldsymbol{\Xi}_l - \boldsymbol{\Xi}_{l'}\|_F^2 + \|\boldsymbol{\Gamma}_l^{(t)} - \boldsymbol{\Gamma}_{l'}^{(t)}\|_F^2)^{1/2}, \lambda_3 \right) \right\}, \text{s.t.} \boldsymbol{\Xi}_l = \boldsymbol{\Theta}_l, l = 1,\ldots,K,$$

where $\{\boldsymbol{\Xi}\} = (\boldsymbol{\Xi}_1, \cdots, \boldsymbol{\Xi}_K)$ and $\boldsymbol{\Xi}_k = (\xi_{jm,k})_{1\leq j,m\leq p}$. The scale augmented Lagrangian form for this problem is:

$$\arg\min_{\boldsymbol{\Theta},\boldsymbol{\Xi},\boldsymbol{\Psi}} \left\{ \sum_{l=1}^{K} n_l^{(t)} \left[-\log\det(\boldsymbol{\Theta}_l) + \mathrm{tr}(\mathbf{S}_{\Gamma l}^{(t)}\boldsymbol{\Theta}_l) \right] + \sum_{l=1}^{K}\sum_{j\neq m} p(|\xi_{jm,l}|,\lambda_1) \right.$$
$$\left. + \sum_{l<l'} p\left((\|\boldsymbol{\Xi}_l - \boldsymbol{\Xi}_{l'}\|_F^2 + \|\boldsymbol{\Gamma}_l^{(t)} - \boldsymbol{\Gamma}_{l'}^{(t)}\|_F^2)^{1/2}, \lambda_3 \right) + \frac{\kappa}{2}\sum_{l=1}^{K} \|\boldsymbol{\Theta}_l - \boldsymbol{\Xi}_l + \boldsymbol{\Psi}_l\|_F^2 - \frac{\kappa}{2}\sum_{l=1}^{K}\|\boldsymbol{\Psi}_l\|_F^2 \right\},$$

where $\boldsymbol{\Psi} = (\boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_K)$ are dual variables. κ is the penalty parameter, and we set $\kappa = 1$. We update $\boldsymbol{\Theta}$, $\boldsymbol{\Xi}$ and $\boldsymbol{\Psi}$ iteratively with initial values $\boldsymbol{\Theta}_l^{(0)} = \mathbf{I}$, $\boldsymbol{\Xi}_l^{(0)} = \boldsymbol{\Psi}_l^{(0)} = \mathbf{0}$.

First, update $\boldsymbol{\Theta}_l^{(m)}$ for $l = 1,\ldots,K$ by solving:

$$\arg\min_{\boldsymbol{\Theta}} \left(\sum_{l=1}^{K} n_l^{(t)} \left[-\log\det(\boldsymbol{\Theta}_l) + \mathrm{tr}(\mathbf{S}_{\Gamma l}^{(t)}\boldsymbol{\Theta}_l) \right] + \frac{\kappa}{2}\sum_{l=1}^{K} \|\boldsymbol{\Theta}_l - \boldsymbol{\Xi}_l^{(m-1)} + \boldsymbol{\Psi}_l^{(m-1)}\|_F^2 \right).$$

The closed-form solution is given by:

$$\boldsymbol{\Theta}_l^{(m)} = \boldsymbol{U}\tilde{\boldsymbol{D}}\boldsymbol{U}^T,$$

where $\boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T$ is the eigen-decomposition of $\mathbf{S}_{\Gamma l}^{(t)} - \kappa\boldsymbol{\Xi}_l^{(m-1)}/n_l^{(t)} + \kappa\boldsymbol{\Psi}_l^{(m-1)}/n_l^{(t)}$, $\tilde{\boldsymbol{D}}$ is a diagonal matrix with the j-th diagonal element $n_l^{(t)}\left[-D_{jj} + (D_{jj}^2 + 4\kappa/n_l^{(t)})^{1/2}\right]/2\kappa$, and D_{jj} is the j-th diagonal element of $\boldsymbol{D}$. Note that $\boldsymbol{\Theta}_l^{(m)}$'s are not necessarily

symmetric. They can be symmetrized as:

$$\theta_{jm,l}^{(m)} = \theta_{jm,l}^{(m)} I(|\theta_{jm,l}^{(m)}| \leq \theta_{mj,l}^{(m)}) + \theta_{mj,l}^{(m)} I(|\theta_{jm,l}^{(m)}| > \theta_{mj,l}^{(m)}).$$

Second, update $\boldsymbol{\Xi}_l^{(m)}$ for $l = 1, \ldots, K$ by solving:

$$\arg\min_{\boldsymbol{\Xi}} \left(\frac{\kappa}{2} \sum_{l=1}^{K} \|\boldsymbol{\Xi}_l - \boldsymbol{Z}_l\|_F^2 + \sum_{l=1}^{K} \sum_{j \neq m} p(|\xi_{jm,l}|, \lambda_1) \right.$$
$$\left. + \sum_{l<l'} p\left((\|\boldsymbol{\Xi}_l - \boldsymbol{\Xi}_{l'}\|_F^2 + \|\boldsymbol{\Gamma}_l^{(t)} - \boldsymbol{\Gamma}_{l'}^{(t)}\|_F^2)^{1/2}, \lambda_3 \right) \right),$$

where $\boldsymbol{Z}_l = \boldsymbol{\Theta}_l^{(m)} - \boldsymbol{\Psi}_l^{(m-1)}$. This can be done using the sparse alternating minimization algorithm (S-AMA), which has reformulation:

$$\arg\min_{\boldsymbol{\Xi}} \left(\frac{\kappa}{2} \sum_{j=1}^{p^2} \|\boldsymbol{\xi}_{(j)} - \boldsymbol{z}_{(j)}\|_2^2 + \sum_{j=1}^{p^2} \sum_{l=1}^{K} p(|\xi_{jl}|, \lambda_1) \cdot I(j \in \mathcal{O}) + \sum_{r \in \mathcal{E}} p((\eta_r^{(t)} + \|\boldsymbol{v}_r\|_2^2)^{1/2}, \lambda_3) \right),$$
$$\text{s.t. } \text{vec}\boldsymbol{\Xi}_l - \text{vec}\boldsymbol{\Xi}_{l'} - \boldsymbol{v}_r = 0,$$

where $\boldsymbol{\xi}_{(j)}, \boldsymbol{z}_{(j)} \in \mathbb{R}^K$ are the j-th columns of $(\text{vec}\boldsymbol{\Xi}_1, \cdots, \text{vec}\boldsymbol{\Xi}_K)^T$ and $(\text{vec}\boldsymbol{Z}_1, \cdots, \text{vec}\boldsymbol{Z}_K)^T$, respectively, $j = 1, \ldots, p^2$, and ξ_{jl} is the l-th element of $\boldsymbol{\xi}_{(j)}$. $\mathcal{O} = \{j : j \neq d(p+1), d = 0, 1, \ldots, p-1\}$ is the index set of the off-diagonal elements of the precision matrices. $\mathcal{E} = \{(l, l') : 1 \leq l, l' \leq K\}$, and $\eta_r^{(t)} = \|\boldsymbol{\Gamma}_l^{(t)} - \boldsymbol{\Gamma}_{l'}^{(t)}\|_F^2$. The augmented Largrangian form for this problem is:

$$\arg\min_{\boldsymbol{\Xi}, \boldsymbol{V}, \boldsymbol{\Delta}} \frac{\kappa}{2} \sum_{j=1}^{p^2} \|\boldsymbol{\xi}_{(j)} - \boldsymbol{z}_{(j)}\|_2^2 + \sum_{j=1}^{p^2} \sum_{l=1}^{K} p(|\xi_{jl}|, \lambda_1) \cdot I(j \in \mathcal{O}) + \sum_{r \in \mathcal{E}} p((\eta_r^{(t)} + \|\boldsymbol{v}_r\|_2^2)^{1/2}, \lambda_3)$$
$$\sum_{r \in \mathcal{E}} \langle \boldsymbol{\delta}_r, \boldsymbol{v}_r - \text{vec}\boldsymbol{\Xi}_k + \text{vec}\boldsymbol{\Xi}_{k'} \rangle + \frac{\kappa'}{2} \sum_{r \in \mathcal{E}} \|\boldsymbol{v}_r - \text{vec}\boldsymbol{\Xi}_k + \text{vec}\boldsymbol{\Xi}_{k'}\|_2^2,$$

where $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{|\mathcal{E}|})$, and the dual variables $\boldsymbol{\Delta} = (\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_{|\mathcal{E}|})$. κ' is a penalty parameter, and we set $\kappa' = 0$ when solving for $\boldsymbol{\Xi}$. S-AMA solves one block of variables at a time and updates iteratively.

Updating $\boldsymbol{\Xi}$ requires solving p^2 individual regularization problems:

$$\min_{\boldsymbol{\xi}_{(j)}} \frac{\kappa}{2}\|\boldsymbol{\xi}_{(j)} - \boldsymbol{u}_{(j)}\|_2^2 + \sum_{l=1}^{K} p(|\xi_{jl}|, \lambda_1) \cdot I(j \in \mathcal{O}),$$

where $\boldsymbol{u}_{(j)} = \boldsymbol{z}_{(j)} + \sum_{r\in\mathcal{E}} \delta_{rj}(\boldsymbol{e}_l - \boldsymbol{e}_{l'})$, δ_{rj} is the j-th element of $\boldsymbol{\delta}_r$, and $\boldsymbol{e}_k$ is a K-dimensional vector with the k-th element being 1 and the other elements being 0. This has a closed-form solution with the MCP penalty:

$$\hat{\xi}_{jl} = I(j \in \mathcal{O}) \cdot \begin{cases} \frac{S(u_{lj}, \lambda_2/\kappa)}{1-1/(a\kappa)} & \text{if} |u_{lj}| \le a\lambda_2 \\ u_{lj}, & \text{if} |u_{lj}| > a\lambda_2 \end{cases} + I(j \notin \mathcal{O}) \cdot u_{lj},$$

where $S(t, \lambda) = (|t| - \lambda)_+ \cdot \text{sign}(t)$.

Updating $\boldsymbol{V}$ requires solving the separate regularization problems:

$$\hat{\boldsymbol{v}}_r = \arg\min_{\boldsymbol{v}_r} \frac{\kappa'}{2}\|\boldsymbol{v}_r - \boldsymbol{w}_r\|_2^2 + p((\eta_r^{(t)} + \|\boldsymbol{v}_r\|_2^2)^{1/2}, \lambda_3),$$

where $\boldsymbol{w}_r = \text{vec}\boldsymbol{\Xi}_l - \text{vec}\boldsymbol{\Xi}_{l'} - \boldsymbol{\delta}_r/\kappa'$. This solution has a closed-form with the group MCP penalty:

$$\hat{\boldsymbol{v}}_r = \frac{\boldsymbol{w}_r}{(\eta_r^{(t)} + \|\boldsymbol{w}_r\|_2^2)^{1/2}} \cdot \begin{cases} \frac{S((\eta_r^{(t)} + \|\boldsymbol{w}_r\|_2^2)^{1/2}, \lambda_3/\kappa')}{1-1/(a\kappa')} & \text{if} (\eta_r^{(t)} + \|\boldsymbol{w}_r\|_2^2)^{1/2} \le a\lambda_3 \\ (\eta_r^{(t)} + \|\boldsymbol{w}_r\|_2^2)^{1/2} & \text{if} (\eta_r^{(t)} + \|\boldsymbol{w}_r\|_2^2)^{1/2} > a\lambda_3. \end{cases}$$

Update $\boldsymbol{\Delta}$ by $\boldsymbol{\delta}_r \leftarrow \boldsymbol{\delta}_r + \kappa'(\boldsymbol{v}_r - \text{vec}\boldsymbol{\Xi}_l + \text{vec}\boldsymbol{\Xi}_{l'}) \cdot I(\|\boldsymbol{v}_r\|_2 > 0),\ r \in \mathcal{E}$.

Lastly, update $\mathbf{\Psi}_l^{(m)}$ for $l = 1, \ldots, K$ by $\mathbf{\Psi}_l^{(m)} = \mathbf{\Psi}_l^{(m-1)} + \mathbf{\Theta}_l^{(m)} - \mathbf{\Xi}_l^{(m)}$.

We iteratively update $\mathbf{\Theta}$, $\mathbf{\Xi}$ and $\mathbf{\Psi}$ until $\frac{\|\mathbf{\Theta}_l^{(m)} - \mathbf{\Theta}_l^{(m-1)}\|_F}{\|\mathbf{\Theta}_l^{(m-1)}\|_F} < 10^{-2}$ and then conclude convergence.

S3 Proofs

Proof. Denote the index set of the diagonal components in the K_0 precision matrices as $\mathcal{G} = \bigcup_{k=1}^{K_0} \mathcal{G}_k$, $\mathcal{G}_k = \{k[p(q+1)+1], k[p(q+1)+p+2], \ldots, k[p(q+1)+p^2]\}$, $k = 1, \ldots, K_0$, and the index set of the nonzero components of $\boldsymbol{\Upsilon}^*$ as $\mathcal{M} = \bigcup_{k=1}^{K_0} \mathcal{M}_k$ and $\mathcal{M}_k = \mathcal{M}_{\mathcal{D}_k} \cup \mathcal{M}_{\mathcal{S}_k} \cup \mathcal{G}_k$, where:

$$\mathcal{M}_{\mathcal{D}_k} = \{(k-1)[p(q+1)+p^2] + (j-1)(q+1) + m : (j,m) \in \mathcal{D}_k\},$$

$$\mathcal{M}_{\mathcal{S}_k} = \{(k-1)[p(q+1)+p^2] + p(q+1) + (j-1)p + m : (j,m) \in \mathcal{S}_k\}.$$

Denote $\boldsymbol{\Omega}^* = (\boldsymbol{\Omega}_1^*, \ldots, \boldsymbol{\Omega}_K^*)$ with some true group annotation $\mathcal{T}_k^* = \{l : \boldsymbol{\Omega}_l^* = \boldsymbol{\Upsilon}_k^*, 1 \leq l \leq K\}$, $k = 1, \ldots, K_0$. Define $|\mathcal{T}_{\min}| = \min_{1 \leq k \leq K_0} |\mathcal{T}_k^*|$ and $|\mathcal{T}_{\max}| = \max_{1 \leq k \leq K_0} |\mathcal{T}_k^*|$, where $|\mathcal{T}_k^*|$ is the cardinality of $\mathcal{T}_k^*$. Let

$$\Lambda_{\mathcal{T}^*} = \left\{ \boldsymbol{\Omega} \in \mathbb{R}^{K(p(q+1)+p^2)} : \boldsymbol{\Omega}_l = \boldsymbol{\Omega}_{l'}, \forall l, l' \in \mathcal{T}_k^*, 1 \leq k \leq K_0 \right\}.$$

First, we define the oracle estimator for $\boldsymbol{\Omega}$, $\hat{\boldsymbol{\Omega}}^o \in \mathbb{R}^{K(p(q+1)+p^2)}$, for which the true grouping structure $\{\mathcal{T}_1^*, \ldots, \mathcal{T}_{K_0}^*\}$ is known:

$$\begin{aligned} \hat{\boldsymbol{\Omega}}^o = \arg \max_{\boldsymbol{\Omega} \in \Lambda_{\mathcal{T}^*}} \Bigg\{ & \frac{1}{n} \sum_{i=1}^{n} \log \left(\sum_{l=1}^{K} \pi_l f_l(\mathbf{y}_i | \boldsymbol{x}_i, \boldsymbol{\Gamma}_l, \boldsymbol{\Theta}_l) \right) \\ & - \sum_{l=1}^{K} \sum_{j \neq m} p(|\theta_{jm,l}|, \lambda_1) - \sum_{l=1}^{K} \sum_{j=1}^{p} \sum_{m=1}^{q+1} p(|\gamma_{jm,l}|, \lambda_2) \Bigg\}, \end{aligned} \tag{S3.1}$$

and the corresponding distinct values as $\hat{\boldsymbol{\Upsilon}}^o = (\hat{\boldsymbol{\Upsilon}}_1^o, \ldots, \hat{\boldsymbol{\Upsilon}}_{K_0}^o)$. To prove the theorem, it is sufficient to establish the following Result 1 and Result 2.

Result 1: Under the assumed conditions, the oracle estimator satisfies:

$$\|\hat{\boldsymbol{\Upsilon}}^o - \boldsymbol{\Upsilon}^*\|_2 = O\left(K_0^2\sqrt{\frac{(d+s+p)(\log p + \log q)}{n}}\right), \tag{S3.2}$$

and $\hat{\mathcal{S}}_k^o = \mathcal{S}_k$ as well as $\hat{\mathcal{D}}_k^o = \mathcal{D}_k$ for $k = 1, \ldots, K_0$.

Proof of Result 1: We first define $\boldsymbol{\Upsilon}_0 = (\boldsymbol{\Upsilon}_{01}, \cdots, \boldsymbol{\Upsilon}_{0,K_0})$, where $\boldsymbol{\Upsilon}_{0k} = \text{vec}(\boldsymbol{\Gamma}_{\mathcal{D}_k}, \boldsymbol{\Theta}_k)$ with $\boldsymbol{\Gamma}_{\mathcal{D}_k} = \{\gamma_{jm,k} : (j,m) \in \mathcal{D}_k\}$ contains the corresponding coefficients of the true active ones. Then there exists a local maximizer $\hat{\boldsymbol{\Upsilon}}_0$ of objective function:

$$Q_n(\boldsymbol{\Upsilon}_0) = \frac{1}{n}\sum_{i=1}^{n}\log\left(\sum_{k=1}^{K_0}\pi_k f_k(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\Upsilon}_{0k})\right) - \mathcal{P}_1(\boldsymbol{\Upsilon}_0), \tag{S3.3}$$

where $\mathcal{P}_1(\boldsymbol{\Upsilon}_0) = \sum_{k=1}^{K_0}\sum_{j\neq m}|\mathcal{T}_k^*|\rho(|\theta_{jm,k}|, \lambda_1)$ and $\rho(t,\lambda) = \lambda^{-1}p(t,\lambda)$. Denote $\boldsymbol{\omega} = (\omega_{ik})_{n\times K_0}$, where $\omega_{ik} = I(\boldsymbol{y}_i \in \mathcal{A}_k)$ is the latent indicator variable designating the component membership of the i-th observation in the mixture, and $\mathcal{A}_k$ is the k-th group. If ω_{ik} is available, objective function (S3.3) can be written as:

$$Q_n(\boldsymbol{\Upsilon}_0|\boldsymbol{\omega}, \boldsymbol{y}, \boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K_0}\omega_{ik}\left[\log\pi_k + \log f_k(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\Upsilon}_{0k})\right] - \mathcal{P}_1(\boldsymbol{\Upsilon}_0). \tag{S3.4}$$

ω_{ik} is a latent Bernoulli variable with expectation $\mathbb{E}(\omega_{ik}|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\Upsilon}) = P(\omega_{ik} = 1|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\Upsilon})$, which is denoted as $L_k(\boldsymbol{y}_i; \boldsymbol{x}_i, \boldsymbol{\Upsilon}) = \frac{\pi_k f_k(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\Upsilon})}{\sum_{k=1}^{K_0}\pi_k f_k(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\Upsilon})}$, and its population version is $\mathbb{E}(L_k(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{\Upsilon})) = \int L_k(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{\Upsilon}) f(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\Upsilon}) d\boldsymbol{y}$. In the t-th step of the EM algorithm, it is needed to maximize the conditional expectation of (S3.4), which is denoted as $\tilde{H}_n(\boldsymbol{\Upsilon}_0|\boldsymbol{\Upsilon}_0^{(t-1)})$:

$$\tilde{H}_n(\boldsymbol{\Upsilon}_0|\boldsymbol{\Upsilon}_0^{(t-1)}) = \mathbb{E}_{\boldsymbol{\omega},\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\Upsilon}_0^{(t-1)}}[Q_n(\boldsymbol{\Upsilon}_0|\boldsymbol{\omega}, \boldsymbol{y}, \boldsymbol{x})] = H_n(\boldsymbol{\Upsilon}_0|\boldsymbol{\Upsilon}_0^{(t-1)}) - \mathcal{P}_1(\boldsymbol{\Upsilon}_0), \tag{S3.5}$$

where

$$H_n(\boldsymbol{\Upsilon}_0|\boldsymbol{\Upsilon}_0^{(t-1)}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K_0} L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}_0^{(t-1)})\left[\log\pi_k + \log f_k(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\Upsilon}_0)\right].$$

Here $L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}_0^{(t-1)}) = \frac{\pi_k^{(t-1)} f_k(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\Upsilon}_0^{(t-1)})}{\sum_{k=1}^{K_0}\pi_k^{(t-1)} f_k(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\Upsilon}_0^{(t-1)})}$, which depends on $\pi_k^{(t-1)}$ and $\boldsymbol{\Upsilon}_0^{(t-1)}$. $\boldsymbol{\Upsilon}_0^{(t-1)}$ is the estimate from the $(t-1)$-th iteration of the EM algorithm, and $\boldsymbol{\Upsilon}_0^{(t-1)} = \mathrm{vec}(\boldsymbol{\Gamma}_{\mathcal{D}}^{(t-1)},\boldsymbol{\Theta}^{(t-1)}) = \arg\max_{\boldsymbol{\Upsilon}_0}\tilde{H}_n(\boldsymbol{\Upsilon}_0|\boldsymbol{\Upsilon}_0^{(t-2)})$.

STEP 1: Define $\boldsymbol{\Upsilon}_0^* = (\boldsymbol{\Upsilon}_{01}^*,\cdots,\boldsymbol{\Upsilon}_{0,K_0}^*)$, where $\boldsymbol{\Upsilon}_{0k}^* = \mathrm{vec}(\boldsymbol{\Gamma}_{\mathcal{D}_k}^*,\boldsymbol{\Theta}_k^*)$ with $\boldsymbol{\Gamma}_{\mathcal{D}_k}^* = \{\gamma_{jm,k}^* : (j,m)\in\mathcal{D}_k\}$ is the true value of the nonzero parameters in the coefficient matrices. We need to show that if $\boldsymbol{\Upsilon}_0^{(t-1)}\in\mathcal{B}_\alpha(\boldsymbol{\Upsilon}_0^*)$, then $\|\boldsymbol{\Upsilon}_0^{(t)}-\boldsymbol{\Upsilon}_0^*\|_2\le\chi$ with probability tending to 1, where $\chi = \frac{4\epsilon}{\varrho}+\iota\|\boldsymbol{\Upsilon}_0^{(t-1)}-\boldsymbol{\Upsilon}_0^*\|_2$, $1/6\le\iota<1$. We set $\alpha = O((d+s+p)\sqrt{(\log p+\log q)/n})$ and $\epsilon = O(\sqrt{(d+s+p)(\log p+\log q)/n})$.

It suffices to show that, for

$$q(\mathbf{v}) = \tilde{H}_n(\boldsymbol{\Upsilon}_0^*+\mathbf{v}|\boldsymbol{\Upsilon}_0^{(t-1)}) - \tilde{H}_n(\boldsymbol{\Upsilon}_0^*|\boldsymbol{\Upsilon}_0^{(t-1)}), \tag{S3.6}$$

$P(\sup_{\mathbf{v}\in C(\chi)}q(\mathbf{v})<0)\to 1$, where $C(\chi)=\{\mathbf{v}:\|\mathbf{v}\|_2=\chi\}$. Note that for a sufficiently large n, $\chi\le 2\alpha$.

Next, we first show that for any $\boldsymbol{\Upsilon}_0$, with probability at least $1-\delta$, for all $\boldsymbol{\Upsilon}_0' = \mathrm{vec}(\boldsymbol{\Gamma}_{\mathcal{D}}',\boldsymbol{\Theta}')\in\{\boldsymbol{\Upsilon}_0' : \|\boldsymbol{\Upsilon}_0'-\boldsymbol{\Upsilon}_0^*\|_2\le\alpha_0\}$,

$$H_n(\boldsymbol{\Upsilon}_0'|\boldsymbol{\Upsilon}_0) - H_n(\boldsymbol{\Upsilon}_0^*|\boldsymbol{\Upsilon}_0) \le \langle\nabla H_n(\boldsymbol{\Upsilon}_0^*|\boldsymbol{\Upsilon}_0),\boldsymbol{\Upsilon}_0'-\boldsymbol{\Upsilon}_0^*\rangle - \frac{\varrho}{2}\|\boldsymbol{\Upsilon}_0'-\boldsymbol{\Upsilon}_0^*\|_2^2, \tag{S3.7}$$

with a sufficiently large n, where $\varrho = c\cdot\min(\beta_1 C_0,(\beta_2+\alpha_0)^{-2}/4)$ and the gradient $\nabla H_n(\boldsymbol{\Upsilon}_0^*|\boldsymbol{\Upsilon}_0)$ is taken with respect to the first variable in $H_n(\cdot|\cdot)$.

For $k = 1, \ldots, K_0$, we define:

$$\nabla_{\boldsymbol{\Upsilon}'_{0k}} H_n(\boldsymbol{\Upsilon}'_{0k}|\boldsymbol{\Upsilon}_0) = \left([\mathrm{vec}(\nabla_{\boldsymbol{\Gamma}'_{\mathcal{D}_k}} H_n(\boldsymbol{\Upsilon}'_{0k}|\boldsymbol{\Upsilon}_0))]^T, [\mathrm{vec}(\nabla_{\boldsymbol{\Theta}'_k} H_n(\boldsymbol{\Upsilon}'_{0k}|\boldsymbol{\Upsilon}_0))]^T\right)^T,$$

$\nabla_{\boldsymbol{\Gamma}'_{\mathcal{D}_k}} H_n(\boldsymbol{\Upsilon}'_{0k}|\boldsymbol{\Upsilon}_0) = \frac{1}{n}\sum_{i=1}^n \left[L_k(\mathbf{y}_i; \boldsymbol{x}_i, \boldsymbol{\Upsilon}_0)\boldsymbol{\Theta}'_k(\mathbf{y}_i - \boldsymbol{\Gamma}'_{\mathcal{D}_k}\mathbf{x}_{i\mathcal{D}_k})\mathbf{x}_{i\mathcal{D}_k}^T\right]$, and $\nabla_{\boldsymbol{\Theta}'_k} H_n(\boldsymbol{\Upsilon}'_{0k}|\boldsymbol{\Upsilon}_0) = \frac{1}{2n}\sum_{i=1}^n \left[L_k(\mathbf{y}_i; \boldsymbol{x}_i, \boldsymbol{\Upsilon}_0)\boldsymbol{\Theta}_k'^{-1}\right] - \frac{1}{2n}\sum_{i=1}^n \left[L_k(\mathbf{y}_i; \boldsymbol{x}_i, \boldsymbol{\Upsilon}_0)(\mathbf{y}_i - \boldsymbol{\Gamma}'_{\mathcal{D}_k}\mathbf{x}_{i\mathcal{D}_k})(\mathbf{y}_i - \boldsymbol{\Gamma}'_{\mathcal{D}_k}\mathbf{x}_{i\mathcal{D}_k})^T\right]$,

where $\boldsymbol{x}_{i\mathcal{D}_k} = \{x_{ij}, j \in \mathcal{D}_k\}$. Note that:

$$H_n(\boldsymbol{\Upsilon}'_{0k}|\boldsymbol{\Upsilon}_0) - H_n(\boldsymbol{\Upsilon}^*_{0k}|\boldsymbol{\Upsilon}_0) = \frac{1}{n}\sum_{i=1}^n \left\{ L_k(\mathbf{y}_i; \boldsymbol{x}_i, \boldsymbol{\Upsilon}_0)\left[\frac{1}{2}\log\det(\boldsymbol{\Theta}'_k) - \frac{1}{2}\log\det(\boldsymbol{\Theta}^*_k) + \frac{1}{2}(\mathbf{y}_i - \boldsymbol{\Gamma}^*_{\mathcal{D}_k}\mathbf{x}_{i\mathcal{D}_k})^T\boldsymbol{\Theta}^*_k(\mathbf{y}_i - \boldsymbol{\Gamma}^*_{\mathcal{D}_k}\mathbf{x}_{i\mathcal{D}_k}) - \frac{1}{2}(\mathbf{y}_i - \boldsymbol{\Gamma}'_{\mathcal{D}_k}\mathbf{x}_{i\mathcal{D}_k})^T\boldsymbol{\Theta}'_k(\mathbf{y}_i - \boldsymbol{\Gamma}'_{\mathcal{D}_k}\mathbf{x}_{i\mathcal{D}_k})\right]\right\}.$$

If we define $h(\boldsymbol{\Gamma}, \boldsymbol{\Theta}) = \frac{1}{2}(\mathbf{y}_i - \boldsymbol{\Gamma}\mathbf{x}_{i\mathcal{D}_k})^T\boldsymbol{\Theta}(\mathbf{y}_i - \boldsymbol{\Gamma}\mathbf{x}_{i\mathcal{D}_k})$, then we can show that:

$$H_n(\boldsymbol{\Upsilon}'_{0k}|\boldsymbol{\Upsilon}_0) - H_n(\boldsymbol{\Upsilon}^*_{0k}|\boldsymbol{\Upsilon}_0) - \left\langle \nabla_{\boldsymbol{\Upsilon}'_{0k}} H_n(\boldsymbol{\Upsilon}^*_{0k}|\boldsymbol{\Upsilon}_0), \boldsymbol{\Upsilon}'_{0k} - \boldsymbol{\Upsilon}^*_{0k}\right\rangle = H_1 + H_2,$$

with

$$H_1 = \frac{1}{n}\sum_{i=1}^n \left\{L_k(\boldsymbol{y}_i; \boldsymbol{x}_i, \boldsymbol{\Upsilon}_0)\left[h(\boldsymbol{\Gamma}^*_{\mathcal{D}_k}, \boldsymbol{\Theta}^*_k) - h(\boldsymbol{\Gamma}'_{\mathcal{D}_k}, \boldsymbol{\Theta}^*_k)\right]\right\} - [\mathrm{vec}(\boldsymbol{\Gamma}'_{\mathcal{D}_k} - \boldsymbol{\Gamma}^*_{\mathcal{D}_k})]^T\nabla_{\boldsymbol{\Gamma}^*_{\mathcal{D}_k}} H_n(\boldsymbol{\Upsilon}^*_{0k}|\boldsymbol{\Upsilon}_0),$$

$$H_2 = \frac{1}{n}\sum_{i=1}^n \left\{ L_k(\boldsymbol{y}_i; \boldsymbol{x}_i, \boldsymbol{\Upsilon}_0)\left[\frac{1}{2}\log\det(\boldsymbol{\Theta}'_k) - \frac{1}{2}\log\det(\boldsymbol{\Theta}^*_k) + h(\boldsymbol{\Gamma}'_{\mathcal{D}_k}, \boldsymbol{\Theta}^*_k) - h(\boldsymbol{\Gamma}'_{\mathcal{D}_k}, \boldsymbol{\Theta}'_k)\right]\right\} - [\mathrm{vec}(\boldsymbol{\Theta}'_k - \boldsymbol{\Theta}^*_k)]^T\nabla_{\boldsymbol{\Theta}^*_k} H_n(\boldsymbol{\Upsilon}^*_{0k}|\boldsymbol{\Upsilon}_0).$$

As for H_1, if we define $g_1(\boldsymbol{\Gamma}_{\mathcal{D}_k}) = -\frac{1}{n}\sum_{i=1}^n L_k(\boldsymbol{y}_i; \boldsymbol{x}_i, \boldsymbol{\Upsilon}_0)h(\boldsymbol{\Gamma}_{\mathcal{D}_k}, \boldsymbol{\Theta}^*_k)$ and note that

$$\begin{aligned} H_1 &= g_1(\boldsymbol{\Gamma}'_{\mathcal{D}_k}) - g_1(\boldsymbol{\Gamma}^*_{\mathcal{D}_k}) - \left\langle \mathrm{vec}(\nabla g_1(\boldsymbol{\Gamma}^*_{\mathcal{D}_k})), \mathrm{vec}(\boldsymbol{\Gamma}'_{\mathcal{D}_k} - \boldsymbol{\Gamma}^*_{\mathcal{D}_k})\right\rangle \\ &= \frac{1}{2}[\mathrm{vec}(\boldsymbol{\Gamma}'_{\mathcal{D}_k} - \boldsymbol{\Gamma}^*_{\mathcal{D}_k})]^T\nabla^2 g_1(\mathbf{Z})[\mathrm{vec}(\boldsymbol{\Gamma}'_{\mathcal{D}_k} - \boldsymbol{\Gamma}^*_{\mathcal{D}_k})] \end{aligned}$$

according to Taylor's expansion, with $\mathbf{Z} = t\mathbf{\Gamma}'_{\mathcal{D}_k} + (1-t)\mathbf{\Gamma}^*_{\mathcal{D}_k}$ and $t \in [0,1]$, then:

$$-\nabla^2 g_1(\mathbf{Z}) = \frac{1}{n}\sum_{i=1}^{n} L_k(\boldsymbol{y}_i; \boldsymbol{x}_i, \mathbf{\Upsilon}_0)\mathbf{\Theta}^*_k \otimes (\boldsymbol{x}_{i\mathcal{D}_k}\boldsymbol{x}^T_{i\mathcal{D}_k}),$$

$$\begin{aligned}\lambda_{\min}(-\nabla^2 g_1(\mathbf{Z})) &\geq \lambda_{\min}(\mathbf{\Theta}^*_k)\lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^{n} L_k(\boldsymbol{y}_i; \boldsymbol{x}_i, \mathbf{\Upsilon}_0)\boldsymbol{x}_{i\mathcal{D}_k}\boldsymbol{x}^T_{i\mathcal{D}_k}\right) \\ &\geq \lambda_{\min}(\mathbf{\Theta}^*_k)\left[\lambda_{\min}(\frac{1}{n}\mathbf{X}^T_{\mathcal{D}_k}\mathbf{G}_k\mathbf{X}_{\mathcal{D}_k}) - \left\|\frac{1}{n}\sum_{i=1}^{n} L_k(\boldsymbol{y}_i; \boldsymbol{x}_i, \mathbf{\Upsilon}_0)\boldsymbol{x}_{i\mathcal{D}_k}\boldsymbol{x}^T_{i\mathcal{D}_k} - \frac{1}{n}\mathbf{X}^T_{\mathcal{D}_k}\mathbf{G}_k\mathbf{X}_{\mathcal{D}_k}\right\|_F\right],\end{aligned} \tag{S3.8}$$

where $\mathbf{G} = \mathrm{diag}(\mathbb{E}(L_k(\boldsymbol{y}_i; \boldsymbol{x}_i, \mathbf{\Upsilon}_0))$ is a diagonal matrix. Note that $L_{ik}(\boldsymbol{y}_i; \boldsymbol{x}_i, \mathbf{\Upsilon}_0)\boldsymbol{x}_{i\mathcal{D}_k}\boldsymbol{x}^T_{i\mathcal{D}_k}$ is bounded by some matrix $\mathbf{A}$, and define $M^2 = \|\mathbf{A}\|^2$, where $\|\cdot\|$ is the spectral norm. According to matrix Hoeffding's inequality, we have:

$$P\left\{\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^{n} L_k(\boldsymbol{y}_i; \boldsymbol{x}_i, \mathbf{\Upsilon}_0)\boldsymbol{x}_{i\mathcal{D}_k}\boldsymbol{x}^T_{i\mathcal{D}_k} - \frac{1}{n}\mathbf{X}^T_{\mathcal{D}_k}\mathbf{G}_k\mathbf{X}_{\mathcal{D}_k}\right) \geq t\right\} \geq 1 - d_k\exp(-nt^2/8M^2), \tag{S3.9}$$

where $d_k = |\mathcal{D}_k|$. If we let $t = \sqrt{\frac{8M^2}{n}\log\frac{d_k K_0}{\delta}}$, then

$$\left\|\frac{1}{n}\sum_{i=1}^{n} L_k(\boldsymbol{y}_i; \boldsymbol{x}_i, \mathbf{\Upsilon}_0)\boldsymbol{x}_{i\mathcal{D}_k}\boldsymbol{x}^T_{i\mathcal{D}_k} - \frac{1}{n}\mathbf{X}^T_{\mathcal{D}_k}\mathbf{G}_k\mathbf{X}_{\mathcal{D}_k}\right\|_F \leq \sqrt{\frac{8d_k M^2}{n}\log\frac{d_k K_0}{\delta}}$$

with probability at least $1-\delta/K_0$, and $\sqrt{\frac{8d_k M^2}{n}\log\frac{d_k K_0}{\delta}} = o(1)$. Therefore, by Condition (C1) and (C3), we have $H_1 \leq -\beta_1 C_0\|\mathrm{vec}(\mathbf{\Gamma}'_{\mathcal{D}_k} - \mathbf{\Gamma}^*_{\mathcal{D}_k})\|_2^2/2$ with probability at least $1-\delta/K_0$.

As for H_2, if we define $g_2(\mathbf{\Theta}_k) = \frac{1}{n}\sum_{i=1}^{n}\left[L_k(\boldsymbol{y}_i; \boldsymbol{x}_i, \mathbf{\Upsilon}_0)\left\{\frac{1}{2}\log\det(\mathbf{\Theta}_k) - h(\mathbf{\Gamma}'_{\mathcal{D}_k}, \mathbf{\Theta}_k)\right\}\right]$, then

$$H_2 = g_2(\mathbf{\Theta}'_k) - g_2(\mathbf{\Theta}^*_k) - \langle\mathrm{vec}(\nabla g_2(\mathbf{\Theta}^*_k)), \mathrm{vec}(\mathbf{\Theta}'_k - \mathbf{\Theta}^*_k)\rangle = \frac{1}{2}[\mathrm{vec}(\mathbf{\Theta}'_k - \mathbf{\Theta}^*_k)]^T\nabla^2 g_2(\mathbf{Z})[\mathrm{vec}(\mathbf{\Theta}'_k - \mathbf{\Theta}^*_k)],$$

according to Taylor's expansion, where $\mathbf{Z} = t\boldsymbol{\Theta}'_k + (1-t)\boldsymbol{\Theta}^*_k$ with $t \in [0,1]$. In other words, if we define $\boldsymbol{\Delta}' = \boldsymbol{\Theta}'_k - \boldsymbol{\Theta}^*_k$, then $\mathbf{Z} = \boldsymbol{\Theta}^*_k + t\boldsymbol{\Delta}'$. According to the proof of Lemma 9 in Hao et al. (2018), under Condition (C1),

$$\lambda_{\min}(-\nabla^2 g_2(\mathbf{Z})) = \frac{1}{2}L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}_0)\|\boldsymbol{\Theta}^*_k + t\boldsymbol{\Delta}'\|_2^{-2} \geq \frac{1}{2}L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}_0)[\|\boldsymbol{\Theta}^*_k\|_2 + \|t\boldsymbol{\Delta}'\|_2]^{-2}$$
$$\geq \frac{1}{2}L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}_0)(\beta_2+\alpha_0)^{-2}.$$

Therefore, it can be obtained that:

$$H_2 \leq -\frac{1}{2n}\sum_{i=1}^{n} L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}_0)\frac{(\beta_2+\alpha_0)^{-2}}{2}\|\mathrm{vec}(\boldsymbol{\Theta}'_k - \boldsymbol{\Theta}^*_k)\|_2^2.$$

Since $0 \leq L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}_0) \leq 1$, according to Hoeffding's inequality,

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}[L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}_0) - \mathbb{E}(L_k(\boldsymbol{y};\boldsymbol{x},\boldsymbol{\Upsilon}_0))]\right| \leq t\right) \geq 1 - 2\exp(-2nt^2).$$

If we let $t = \sqrt{\frac{1}{2n}\log\frac{2K_0}{\delta}}$, then

$$\left|\frac{1}{n}\sum_{i=1}^{n}[L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}_0) - \mathbb{E}(L_k(\boldsymbol{y};\boldsymbol{x},\boldsymbol{\Upsilon}_0))]\right| \leq \sqrt{\frac{1}{2n}\log\frac{2K_0}{\delta}}, \tag{S3.10}$$

with probability at least $1-\delta/K_0$. Since $\sqrt{\frac{1}{2n}\log\frac{2K_0}{\delta}} = o(1)$ and $0 \leq \mathbb{E}(L_k(\boldsymbol{y};\boldsymbol{x},\boldsymbol{\Upsilon}_0)) \leq 1$, there exists some constant c such that $-\frac{1}{n}\sum_{i=1}^{n} L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}_0) \leq -c$ when n is large enough. Then, we have $H_2 \leq -\frac{1}{4}c(\beta_2+\alpha_0)^{-2}\|\mathrm{vec}(\boldsymbol{\Theta}'_k - \boldsymbol{\Theta}^*_k)\|_2^2$ with probability at least $1-\delta/K_0$.

Combining with the upper bound of H_1, we have:

$$H_1 + H_2 \leq -c\cdot\frac{\min(\beta_1 C_0, 0.5(\beta_2+\alpha_0)^{-2})}{2}\|\boldsymbol{\Upsilon}'_{0k} - \boldsymbol{\Upsilon}^*_{0k}\|_2^2,$$

with probability at least $1-\delta/K_0$. It further implies that (S3.7) holds with probability at least $1-\delta$, if we define $\varrho = c \cdot \min(\beta_1 C_0, 0.5(\beta_2+\alpha_0)^{-2})$.

Therefore, it can be obtained that:

$$\begin{aligned} q(\mathbf{v}) &= \tilde{H}_n(\boldsymbol{\Upsilon}_0^* + \mathbf{v}|\boldsymbol{\Upsilon}_0^{(t-1)}) - \tilde{H}_n(\boldsymbol{\Upsilon}_0^*|\boldsymbol{\Upsilon}_0^{(t-1)}) \\ &= H_n(\boldsymbol{\Upsilon}_0^* + \mathbf{v}|\boldsymbol{\Upsilon}_0^{(t-1)}) - H_n(\boldsymbol{\Upsilon}_0^*|\boldsymbol{\Upsilon}_0^{(t-1)}) - \lambda_1[\mathcal{P}_1(\boldsymbol{\Upsilon}_0^* + \mathbf{v}) - \mathcal{P}_1(\boldsymbol{\Upsilon}_0^*)] \\ &\leq \left\langle \nabla H_n(\boldsymbol{\Upsilon}_0^*|\boldsymbol{\Upsilon}_0^{(t-1)}), \mathbf{v} \right\rangle - \lambda_1 \left[\mathcal{P}_1(\boldsymbol{\Upsilon}_0^* + \mathbf{v}) - \mathcal{P}_1(\boldsymbol{\Upsilon}_0^*)\right] \\ &= \left\langle \nabla H_n(\boldsymbol{\Upsilon}_0^*|\boldsymbol{\Upsilon}_0^{(t-1)}) - \nabla H(\boldsymbol{\Upsilon}_0^*|\boldsymbol{\Upsilon}_0^{(t-1)}) + \nabla H(\boldsymbol{\Upsilon}_0^*|\boldsymbol{\Upsilon}_0^{(t-1)}) - \nabla H(\boldsymbol{\Upsilon}_0^*|\boldsymbol{\Upsilon}_0^*), \mathbf{v} \right\rangle \\ &\quad - \lambda_1 \left[\mathcal{P}_1(\boldsymbol{\Upsilon}_0^* + \mathbf{v}) - \mathcal{P}_1(\boldsymbol{\Upsilon}_0^*)\right] - \frac{\varrho}{2}\|\mathbf{v}\|_2^2 \\ &= q_1 + q_2 + q_3, \end{aligned} \tag{S3.11}$$

where

$$q_1 = \left\langle \nabla_{\boldsymbol{\Upsilon}^*_{\mathcal{S}^C}} H_n(\boldsymbol{\Upsilon}_0^*|\boldsymbol{\Upsilon}_0^{(t-1)}) - \nabla_{\boldsymbol{\Upsilon}^*_{\mathcal{S}^C}} H(\boldsymbol{\Upsilon}_0^*|\boldsymbol{\Upsilon}_0^{(t-1)}), \mathbf{v}_{\mathcal{S}^C} \right\rangle - \lambda_1 \left[\mathcal{P}_1(\boldsymbol{\Upsilon}^*_{\mathcal{S}^C} + \mathbf{v}_{\mathcal{S}^C}) - \mathcal{P}_1(\boldsymbol{\Upsilon}^*_{\mathcal{S}^C})\right],$$

$$q_2 = \left\langle \nabla_{\boldsymbol{\Upsilon}^*_{\mathcal{M}}} H_n(\boldsymbol{\Upsilon}_0^*|\boldsymbol{\Upsilon}_0^{(t-1)}) - \nabla_{\boldsymbol{\Upsilon}^*_{\mathcal{M}}} H(\boldsymbol{\Upsilon}_0^*|\boldsymbol{\Upsilon}_0^{(t-1)}), \mathbf{v}_{\mathcal{M}} \right\rangle - \lambda_1 \left[\mathcal{P}_1(\boldsymbol{\Upsilon}^*_{\mathcal{M}} + \mathbf{v}_{\mathcal{M}}) - \mathcal{P}_1(\boldsymbol{\Upsilon}^*_{\mathcal{M}})\right],$$

$$q_3 = \left\langle \nabla H(\boldsymbol{\Upsilon}_0^*|\boldsymbol{\Upsilon}_0^{(t-1)}) - \nabla H(\boldsymbol{\Upsilon}_0^*|\boldsymbol{\Upsilon}_0^*), \mathbf{v} \right\rangle - \frac{\varrho}{2}\|\mathbf{v}\|_2^2.$$

With respect to q_1, if $|v_{kj}| \leq a\lambda_1$, then $\lambda_1|\mathcal{T}_k^*|\rho(|v_{kj}|, \lambda_1) \geq \lambda_1 C_{\lambda_1}|\mathcal{T}_{\min}||v_{kj}|$ for a constant $C_{\lambda_1} > 0$. And if $|v_{kj}| > a\lambda_1$, then $\lambda_1|\mathcal{T}_k^*|\rho(|v_{kj}|, \lambda_1) \geq \frac{1}{2}a\lambda_1^2|\mathcal{T}_{\min}|$. Thus,

$$\begin{aligned} q_1 \leq &\sum_{k=1}^{K_0} \sum_{j \in \mathcal{S}_k^C} \left(\|\nabla H_n(\boldsymbol{\Upsilon}_0^*|\boldsymbol{\Upsilon}_0^{(t-1)}) - \nabla H(\boldsymbol{\Upsilon}_0^*|\boldsymbol{\Upsilon}_0^{(t-1)})\|_\infty - \lambda_1 C_{\lambda_1}|\mathcal{T}_{\min}| \right) |v_{kj}| \cdot I(|v_{kj}| \leq a\lambda_1) \\ &+ \sum_{k=1}^{K_0} \sum_{j \in \mathcal{S}_k^C} \left(\|\nabla H_n(\boldsymbol{\Upsilon}_0^*|\boldsymbol{\Upsilon}_0^{(t-1)}) - \nabla H(\boldsymbol{\Upsilon}_0^*|\boldsymbol{\Upsilon}_0^{(t-1)})\|_\infty \|\mathbf{v}\|_\infty - \frac{1}{2}a\lambda_1^2|\mathcal{T}_{\min}| \right) \cdot I(|v_{kj}| > a\lambda_1). \end{aligned}$$

According to Lemma 1, $\|\nabla H_n(\boldsymbol{\Upsilon}_0^*|\boldsymbol{\Upsilon}_0^{(t-1)}) - \nabla H(\boldsymbol{\Upsilon}_0^*|\boldsymbol{\Upsilon}_0^{(t-1)})\|_\infty = O_p(\sqrt{(\log p + \log q)/n})$. Note that $\|\mathbf{v}\|_\infty \leq \frac{4\epsilon}{\varrho} + \iota\alpha$, $\alpha = O((d+s+p)\sqrt{(\log p + \log q)/n})$, and $\lambda_1 \gg O(\sqrt{(d+s+p)(\log p + \log q)/n})$. Therefore, $q_1 < 0$.

Now consider q_2. Denote $u_{\mathcal{M}}(\boldsymbol{\Upsilon}_0^*|\boldsymbol{\Upsilon}_0^{(t-1)}) = \nabla_{\boldsymbol{\Upsilon}^*_{\mathcal{M}}} H_n(\boldsymbol{\Upsilon}_0^*|\boldsymbol{\Upsilon}_0^{(t-1)}) - \nabla_{\boldsymbol{\Upsilon}^*_{\mathcal{M}}} H(\boldsymbol{\Upsilon}_0^*|\boldsymbol{\Upsilon}_0^{(t-1)})$.

$$\left\langle u_{\mathcal{M}}(\boldsymbol{\Upsilon}_0^*|\boldsymbol{\Upsilon}_0^{(t-1)}), \mathbf{v}_{\mathcal{M}} \right\rangle \leq \|u_{\mathcal{M}}(\boldsymbol{\Upsilon}_0^*|\boldsymbol{\Upsilon}_0^{(t-1)})\|_\infty \|\mathbf{v}_{\mathcal{M}}\|_2 \leq \epsilon_1 \sqrt{K_0(d+s+p)}\|\mathbf{v}\|_2, \tag{S3.12}$$

where $\epsilon_1 = C\sqrt{(\log p + \log q)/n}$ according to Lemma 1. In addition,

$$-\lambda_1 \left[\mathcal{P}_1(\boldsymbol{\Upsilon}^*_{\mathcal{M}} + \mathbf{v}_{\mathcal{M}}) - \mathcal{P}_1(\boldsymbol{\Upsilon}^*_{\mathcal{M}})\right] \leq \lambda_1 C'_{p1} |\nabla \mathcal{P}_1(\boldsymbol{\Upsilon}(\mathcal{M})^*)^T \mathbf{v}_{\mathcal{M}}| = 0, \tag{S3.13}$$

for a positive constant C'_{p1} according to the minimal signal condition of the true parameters and Condition (C7). Combining Equation (S3.12) and (S3.13), we have:

$$q_2 \leq \epsilon_1 \sqrt{K_0(d+s+p)}\|\mathbf{v}\|_2. \tag{S3.14}$$

With respect to q_3, according to Lemma 7 in Ren et al. (2022), under Condition (C6), we have:

$$q_3 \leq -\frac{\varrho}{2}\|\mathbf{v}\|_2^2 + \tau \cdot \|\boldsymbol{\Upsilon}_0^{(t-1)} - \boldsymbol{\Upsilon}_0^*\|_2 \|\mathbf{v}\|_2, \tag{S3.15}$$

with probability at least $1 - (26K_0^2 + 8K_0 + 1)\delta$. Combining the upper bounds of q_1, q_2 and q_3, an upper bound of $q(\mathbf{v})$ can be obtained as:

$$q(\mathbf{v}) < -\frac{\varrho}{2}\|\mathbf{v}\|_2^2 + \left(\epsilon_1 \sqrt{K_0(d+s+p)} + \tau \cdot \|\boldsymbol{\Upsilon}_0^{(t-1)} - \boldsymbol{\Upsilon}_0^*\|_2\right) \|\mathbf{v}\|_2, \tag{S3.16}$$

with probability at least $1 - (26K_0^2 + 8K_0 + 1)/(p+q)$ and $\delta = 1/(p+q)$.

Note that $\epsilon = CK_0^2\sqrt{\frac{(d+s+p)(\log p+\log q)}{n}}$. Thus for a properly chosen positive constant C, an upper bound of $q(\mathbf{v})$ can be obtained as:

$$q(\mathbf{v}) < -\frac{\varrho}{2}\|\mathbf{v}\|_2^2 + \left(\epsilon + \tau\cdot\|\boldsymbol{\Upsilon}_0^{(t-1)} - \boldsymbol{\Upsilon}_0^*\|_2\right)\|\mathbf{v}\|_2,$$

with probability at least $1-(26K_0^2+8K_0+1)/(p+q)$. It can be obtained that, when $\|\mathbf{v}\|_2 > \frac{2\epsilon}{\varrho} + \frac{2\tau}{\varrho}\|\boldsymbol{\Upsilon}_0^{(t-1)} - \boldsymbol{\Upsilon}_0^*\|_2$, $q(\mathbf{v}) < 0$. Note that $\|\mathbf{v}\|_2 = \chi = \frac{4\epsilon}{\varrho} + \iota\|\boldsymbol{\Upsilon}_0^{(t-1)} - \boldsymbol{\Upsilon}_0^*\|_2$, $1/6 \le \iota < 1$, and $\tau \le \varrho/12$. Therefore, there is a local maximizer $\boldsymbol{\Upsilon}_0^{(t)}$ that follows $\boldsymbol{\Upsilon}_0^{(t-1)}$ and satisfies: if $\boldsymbol{\Upsilon}_0^{(t-1)} \in \mathcal{B}_\alpha(\boldsymbol{\Upsilon}_0^*)$, then $\|\boldsymbol{\Upsilon}_0^{(t)} - \boldsymbol{\Upsilon}_0^*\|_2 \le \chi$ with probability at least $1-(26K_0^2+8K_0+1)/(p+q)$.

STEP 2: We can show that, if $\boldsymbol{\Upsilon}_0^{(0)} \in \mathcal{B}_\alpha(\boldsymbol{\Upsilon}_0^*)$, for any $t \ge 1$,

$$\begin{aligned}\|\boldsymbol{\Upsilon}_0^{(t)} - \boldsymbol{\Upsilon}_0^*\|_2 &\le \frac{4\epsilon}{\varrho}(\iota^0 + \iota^2 + \cdots + \iota^{t-1}) + \iota^t\|\boldsymbol{\Upsilon}_0^{(0)} - \boldsymbol{\Upsilon}_0^*\|_2 \\ &= \frac{1-\iota^t}{1-\iota}\frac{4\epsilon}{\varrho} + \iota^t\|\boldsymbol{\Upsilon}_0^{(0)} - \boldsymbol{\Upsilon}_0^*\|_2 \\ &\le \frac{8\epsilon}{\varrho} + \iota^t\|\boldsymbol{\Upsilon}_0^{(0)} - \boldsymbol{\Upsilon}_0^*\|_2,\end{aligned} \tag{S3.17}$$

with probability at least $1-t(26K_0^2+8K_0+1)/(p+q)$, where $1/6 \le \iota < 1$. When t is sufficiently large, i.e., $t \ge T = \log_{1/\iota}\left(\frac{\varrho\|\boldsymbol{\Upsilon}_0^{(0)} - \boldsymbol{\Upsilon}_0^*\|_2}{8\epsilon}\right)$, $\iota^t\|\boldsymbol{\Upsilon}_0^{(0)} - \boldsymbol{\Upsilon}_0^*\|_2$ is dominated by $8\epsilon/\varrho$. So the error of $\hat{\boldsymbol{\Upsilon}}_0$ can be bounded as:

$$\|\hat{\boldsymbol{\Upsilon}}_0 - \boldsymbol{\Upsilon}_0^*\|_2 = O\left(\sqrt{\frac{(d+s+p)(\log p+\log q)}{n}})\right), \tag{S3.18}$$

with probability at least $1-T(26K_0^2+8K_0+1)/(p+q)$, which goes to 1 as p, q and n diverge.

STEP 3: We show that $\hat{\boldsymbol{\Upsilon}}^o = (\hat{\boldsymbol{\Upsilon}}_0, \hat{\boldsymbol{\Gamma}}_{\mathcal{D}^C}) = (\hat{\boldsymbol{\Upsilon}}_0, \mathbf{0})$ is a local maximizer of objective function (S3.1), which can also be rewritten as:

$$\tilde{Q}_n(\boldsymbol{\Upsilon}) = \frac{1}{n}\sum_{i=1}^{n} \log\left(\sum_{k=1}^{K_0} \pi_k f_k(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\Upsilon}_k)\right) - \mathcal{P}_1(\boldsymbol{\Upsilon}) - \mathcal{P}_2(\boldsymbol{\Upsilon}) = Q_n(\boldsymbol{\Upsilon}) - \mathcal{P}_2(\boldsymbol{\Upsilon}), \tag{S3.19}$$

where $\mathcal{P}_1(\boldsymbol{\Upsilon}) = \sum_{k=1}^{K_0}\sum_{j\neq m}|\mathcal{T}_k^*|\rho(|\theta_{jm,k}|, \lambda_1)$, $\mathcal{P}_2(\boldsymbol{\Upsilon}) = \sum_{k=1}^{K_0}\sum_{j=1}^{p}\sum_{m=1}^{q+1}|\mathcal{T}_k^*|\rho(|\gamma_{jm,k}|, \lambda_2)$, and $\rho(t,\lambda) = \lambda^{-1}p(t,\lambda)$. This indicates that $\|\hat{\boldsymbol{\Upsilon}}^o - \boldsymbol{\Upsilon}^*\|_2 = O\left(\sqrt{\frac{(d+s+p)(\log p + \log q)}{n}}\right)$ in Result 1 holds due to the conclusion drawn in STEP 2. Following Theorem 1 in Fan and Lv (2011), it suffices to show that $(\lambda_2)^{-1}\|\mathbf{z}\|_\infty \leq \rho'(0+)$, where $\mathbf{z} = \nabla_{\boldsymbol{\Gamma}_{\mathcal{D}^C}} Q_n(\hat{\boldsymbol{\Upsilon}}^o)$. For each $k = 1, \ldots, K_0$, note that:

$$\begin{pmatrix} \nabla_{\boldsymbol{\Upsilon}_0} Q_n(\hat{\boldsymbol{\Upsilon}}_k^o) \\ \nabla_{\boldsymbol{\Gamma}_{\mathcal{D}_k^C}} Q_n(\hat{\boldsymbol{\Upsilon}}_k^o) \end{pmatrix} = \begin{pmatrix} \nabla_{\boldsymbol{\Upsilon}_0} Q_n(\boldsymbol{\Upsilon}_k^*) \\ \nabla_{\boldsymbol{\Gamma}_{\mathcal{D}_k^C}} Q_n(\boldsymbol{\Upsilon}_k^*) \end{pmatrix} + \begin{pmatrix} \nabla^2_{\boldsymbol{\Upsilon}_0\boldsymbol{\Upsilon}_0} Q_n(\boldsymbol{\Upsilon}_k^*) & \nabla^2_{\boldsymbol{\Upsilon}_0\boldsymbol{\Gamma}_{\mathcal{D}_k^C}} Q_n(\boldsymbol{\Upsilon}_k^*) \\ \nabla^2_{\boldsymbol{\Gamma}_{\mathcal{D}_k^C}\boldsymbol{\Upsilon}_0} Q_n(\boldsymbol{\Upsilon}_k^*) & \nabla^2_{\boldsymbol{\Gamma}_{\mathcal{D}_k^C}\boldsymbol{\Gamma}_{\mathcal{D}_k^C}} Q_n(\boldsymbol{\Upsilon}_k^*) \end{pmatrix}$$
$$\begin{pmatrix} \hat{\boldsymbol{\Upsilon}}_{0k} - \boldsymbol{\Upsilon}_{0k}^* \\ \mathbf{0} \end{pmatrix} + \begin{pmatrix} R(\Delta)_0 \\ R(\Delta)_{\mathcal{D}_k^C} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{z}_k \end{pmatrix}, \tag{S3.20}$$

where $R(\Delta) = \left(R(\Delta)_0, R(\Delta)_{\mathcal{D}_k^C}\right)$ is the residual of the first order Taylor's expansion of the gradient. From Lemma 3 in Wytock and Kolter (2013), $\|R(\Delta)\|_\infty$ is bounded

by $\|\Delta\|_\infty = \|\hat{\boldsymbol{\Upsilon}}_k^o - \boldsymbol{\Upsilon}_k^*\|_\infty^2$. According to (S3.20),

$$\begin{aligned}
\mathbf{z}_k &= \nabla_{\boldsymbol{\Gamma}_{\mathcal{D}_k^C}} Q_n(\boldsymbol{\Upsilon}_k^*) + \nabla^2_{\boldsymbol{\Gamma}_{\mathcal{D}_k^C}\boldsymbol{\Upsilon}_0} Q_n(\boldsymbol{\Upsilon}_k^*)(\hat{\boldsymbol{\Upsilon}}_{0k} - \boldsymbol{\Upsilon}_{0k}^*) + R(\Delta)_{\mathcal{D}_k^C} \\
&= \frac{1}{n}\sum_{i=1}^n L_k(\boldsymbol{y}_i; \boldsymbol{x}_i, \boldsymbol{\Upsilon}^*)\boldsymbol{\Theta}_k^*(\boldsymbol{y}_i - \boldsymbol{\Gamma}_k^*\boldsymbol{x}_i)\boldsymbol{x}_{i\mathcal{D}_k^C}^T + \\
&\quad \frac{1}{n}\sum_{i=1}^n L_k(\boldsymbol{y}_i; \boldsymbol{x}_i, \boldsymbol{\Upsilon}^*)\boldsymbol{\Theta}_k^* \otimes \boldsymbol{x}_{i\mathcal{D}_k^C}\boldsymbol{x}_{i\mathcal{D}_k}^T(\hat{\boldsymbol{\Upsilon}}_{0k} - \boldsymbol{\Upsilon}_{0k}^*) + R(\Delta)_{\mathcal{D}_k^C} \\
&= \boldsymbol{\xi}_1 + \boldsymbol{\xi}_2 + R(\Delta)_{\mathcal{D}_k^C},
\end{aligned}$$

$$\begin{aligned}
\|\boldsymbol{\xi}_1\|_\infty &\le \frac{1}{n}\sum_{i=1}^n \mathbb{E}(L_k(\boldsymbol{y}_i; \boldsymbol{x}_i, \boldsymbol{\Upsilon}^*)) \left\| \frac{1}{n}\sum_{i=1}^n \boldsymbol{\Theta}_k^*(\boldsymbol{y}_i - \boldsymbol{\Gamma}_k^*\boldsymbol{x}_i)\boldsymbol{x}_{i\mathcal{D}_k^C}^T \right\|_\infty \\
&\quad + \frac{1}{n}\sum_{i=1}^n [L_k(\boldsymbol{y}_i; \boldsymbol{x}_i, \boldsymbol{\Upsilon}^*) - \mathbb{E}(L_k(\boldsymbol{y}_i; \boldsymbol{x}_i, \boldsymbol{\Upsilon}^*))] \left\| \boldsymbol{\Theta}_k^*(\boldsymbol{y}_i - \boldsymbol{\Gamma}_k^*\boldsymbol{x}_i)\boldsymbol{x}_{i\mathcal{D}_k^C}^T \right\|_\infty \\
&\le \frac{1}{n}\sum_{i=1}^n \|\boldsymbol{\Theta}_k^*\|_\infty \left\| \boldsymbol{\epsilon}_i \boldsymbol{x}_{i\mathcal{D}_k^C} \right\|_\infty + \frac{1}{n}\sum_{i=1}^n [L_k(\boldsymbol{y}_i; \boldsymbol{x}_i, \boldsymbol{\Upsilon}^*) - \mathbb{E}(L_k(\boldsymbol{y}_i; \boldsymbol{x}_i, \boldsymbol{\Upsilon}^*))] \left\| \boldsymbol{\Theta}_k^* \boldsymbol{x}_{i\mathcal{D}_k^C}(\boldsymbol{y}_i - \boldsymbol{\Gamma}_k^*\boldsymbol{x}_i) \right\|_\infty .
\end{aligned}$$

In the first term, $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{ip})$ and $\epsilon_{ij} = (\boldsymbol{y}_i - \boldsymbol{\Gamma}_k^*\boldsymbol{x}_i)_j,\ j = 1, \ldots, p$. According to the Hoeffding's bound,

$$\begin{aligned}
P\left(\sum_{i=1}^n \epsilon_{ij} x_{im} \ge \sqrt{(\log p + \log q)n} \right) &= 1 - P\left(\sum_{i=1}^n \epsilon_{ij} x_{im} \le \sqrt{(\log p + \log q)n} \right) \\
&\ge 1 - 2\exp\left(-\frac{n(\log p + \log q)}{2\max_{1\le j\le q} \|\mathbf{X}_j\|_2^2} \right).
\end{aligned}$$

This probability approaches 1, as $\|\mathbf{X}_j\|_2 = O(\sqrt{n})$ and p, q diverge. In the second term, from (S3.10),

$$P\left(\left| \frac{1}{n}\sum_{i=1}^n [L_k(\boldsymbol{y}_i; \boldsymbol{x}_i, \boldsymbol{\Upsilon}_k^*) - \mathbb{E}(L_k(\boldsymbol{y}_i; \boldsymbol{x}_i, \boldsymbol{\Upsilon}_k^*))] \right| \le \sqrt{\frac{\log p + \log q}{n}} \right),$$

with probability approaching 1 when setting $\delta = 1/pq$. With respect to ξ_2, according

to (S3.20),

$$\hat{\boldsymbol{\Upsilon}}_{0k} - \boldsymbol{\Upsilon}^*_{0k} = -\left[\nabla^2_{\boldsymbol{\Upsilon}_0\boldsymbol{\Upsilon}_0} Q_n(\boldsymbol{\Upsilon}^*_k)\right]^{-1}\left[\nabla_{\boldsymbol{\Upsilon}_0} Q_n(\boldsymbol{\Upsilon}^*_k) + R(\Delta)_0\right].$$

$$\begin{aligned}
\boldsymbol{\xi}_2 &= -\frac{1}{n}\sum_{i=1}^n L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}^*)\boldsymbol{\Theta}^*_k \otimes \boldsymbol{x}_{i\mathcal{D}_k^C}\boldsymbol{x}^T_{i\mathcal{D}_k}\left[\nabla^2_{\boldsymbol{\Upsilon}_0\boldsymbol{\Upsilon}_0} Q_n(\boldsymbol{\Upsilon}^*_k)\right]^{-1}\left[\nabla_{\boldsymbol{\Upsilon}_0} Q_n(\boldsymbol{\Upsilon}^*_k) + R(\Delta)_0\right] \\
&= \left[\frac{1}{n}\sum_{i=1}^n L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}^*)\boldsymbol{\Theta}^*_k \otimes \boldsymbol{x}_{i\mathcal{D}_k^C}\boldsymbol{x}^T_{i\mathcal{D}_k}\right]\left[\frac{1}{n}\sum_{i=1}^n L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}^*)\boldsymbol{\Theta}^*_k \otimes \boldsymbol{x}_{i\mathcal{D}_k}\boldsymbol{x}^T_{i\mathcal{D}_k}\right]^{-1} \\
&\qquad \left(\frac{1}{n}\sum_{i=1}^n L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}^*)\boldsymbol{\Theta}^*_k(\boldsymbol{y}_i - \boldsymbol{\Gamma}^*_k\boldsymbol{x}_i)\boldsymbol{x}^T_{i\mathcal{D}_k} + R(\Delta)_0\right) \\
&= \mathbf{I}_p \otimes \left\{\left[\frac{1}{n}\sum_{i=1}^n L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}^*)\boldsymbol{x}_{i\mathcal{D}_k^C}\boldsymbol{x}^T_{i\mathcal{D}_k}\right]\left[\frac{1}{n}\sum_{i=1}^n L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}^*)\boldsymbol{x}_{i\mathcal{D}_k}\boldsymbol{x}^T_{i\mathcal{D}_k}\right]^{-1}\right\} \\
&\qquad \left(\frac{1}{n}\sum_{i=1}^n L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}^*)\boldsymbol{\Theta}^*_k(\boldsymbol{y}_i - \boldsymbol{\Gamma}^*_k\boldsymbol{x}_i)\boldsymbol{x}^T_{i\mathcal{D}_k} + R(\Delta)_0\right).
\end{aligned}$$

Similar to the proof in (S3.8) and (S3.9), with the matrix Hoeffding inequality, we have:

$$\begin{aligned}
\|\boldsymbol{\xi}_2\|_\infty \leq &\left\|(\mathbf{X}^T_{\mathcal{D}_k^C}\mathbf{G}_k\mathbf{X}_{\mathcal{D}_k})(\mathbf{X}^T_{\mathcal{D}_k}\mathbf{G}_k\mathbf{X}_{\mathcal{D}_k})^{-1}\right\|_{2,\infty} \cdot \\
&\left\|\frac{1}{n}\sum_{i=1}^n L_{ik}(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}^*)\boldsymbol{\Theta}^*_k(\boldsymbol{y}_i - \boldsymbol{\Gamma}^*_k\boldsymbol{x}_i)\boldsymbol{x}^T_{i\mathcal{D}_k} + R(\Delta)_0\right\|_\infty,
\end{aligned} \tag{S3.21}$$

with probability tending to 1. By Condition (C3), we can bound $\|\boldsymbol{\xi}_2\|_\infty$ in the same way. Therefore, with probability tending to 1,

$$\|\mathbf{z}\|_\infty \leq \sqrt{\frac{\log p + \log q}{n}} + n^{\alpha_1}\left(\sqrt{\frac{\log p + \log q}{n}} + \|\hat{\boldsymbol{\Upsilon}}_0 - \boldsymbol{\Upsilon}^*_0\|^2_\infty\right) + \|\hat{\boldsymbol{\Upsilon}}_0 - \boldsymbol{\Upsilon}^*_0\|^2_\infty \ll \lambda_2.$$

STEP 4: Here we establish selection consistency for the precision matrices. That is, $\hat{\mathcal{S}}^o_k \supseteq \mathcal{S}_k$ and $\mathcal{S}_k \supseteq \hat{\mathcal{S}}^o_k$.

First, it is sufficient to show that, for any $(i,j) \in \mathcal{S}_k$ and $k = 1, \ldots, K_0$, $\hat{\theta}_{ij,k} \neq 0$. Note that:

$$|\hat{\theta}_{ij,k}| \geq |\theta^*_{ij,k}| - |\hat{\theta}_{ij,k} - \theta^*_{ij,k}| \geq |\theta^*_{ij,k}| - \sqrt{\sum_{1 \leq i,j \leq p} (\hat{\theta}_{ij,k} - \theta_{ij,k})^2} \geq |\theta^*_{ij,k}| - \|\hat{\boldsymbol{\Upsilon}}^o - \boldsymbol{\Upsilon}^*\|_2.$$

According to the results in STEP 3 and the minimal signal Condition (C4), $|\hat{\theta}_{ij,k}| > 0$, which implies that $\hat{\mathcal{S}}^o_k \supseteq \mathcal{S}_k$.

Second, we show that for any $(i,j) \in \mathcal{S}^c_k$ and $k = 1, \ldots, K_0$, $\hat{\theta}_{ij,k} = 0$. Consider a local maximizer $\hat{\boldsymbol{\Upsilon}}$ that satisfies (S3.18) and optimizes objective function (S3.19). The derivative of the objective function with respect to $\theta_{jm,k}$ for $(j,m) \in \mathcal{S}^C_k, k = 1, \ldots, K_0$ is:

$$\frac{\partial \tilde{Q}_n(\hat{\boldsymbol{\Upsilon}})}{\partial \theta_{jm,k}} = \mathcal{R}(\hat{\boldsymbol{\Upsilon}}_k) - \lambda_1 \rho'(|\hat{\theta}_{jm,k}|)\text{sgn}(\hat{\theta}_{jm,k}),$$

where $\mathcal{R}(\hat{\boldsymbol{\Upsilon}}_k) = \frac{1}{2n} \sum_{i=1}^n L_k(\boldsymbol{y}_i; \boldsymbol{x}_i, \hat{\boldsymbol{\Upsilon}}_k)[\hat{\sigma}_{jm,k} - (y_{ij} - (\hat{\boldsymbol{\Gamma}}_k \boldsymbol{x}_i)_j)(y_{im} - (\hat{\boldsymbol{\Gamma}}_k \boldsymbol{x}_i)_m)]$, and $L_k(\boldsymbol{y}_i; \boldsymbol{x}_i, \hat{\boldsymbol{\Upsilon}}_k) = \frac{\hat{\pi}_k f_k(\boldsymbol{y}_i | \boldsymbol{x}_i, \hat{\boldsymbol{\Upsilon}}_k)}{\sum_{k=1}^{K_0} \hat{\pi}_k f_k(\boldsymbol{y}_i | \boldsymbol{x}_i, \hat{\boldsymbol{\Upsilon}}_k)}$, $\hat{\sigma}_{jm,k}$ is the (j,m)-th element of $\hat{\boldsymbol{\Theta}}_k^{-1}$, and $\text{sgn}(\hat{\theta}_{jm,k})$ denotes the sign of $\hat{\theta}_{jm,k}$. It can be decomposed that $\mathcal{R}(\hat{\boldsymbol{\Upsilon}}_k) \leq |\hat{\sigma}_{jm,k} - \sigma^*_{jm,k}| + \mathcal{R}^*(\hat{\boldsymbol{\Upsilon}}_k)$, where $\sigma^*_{jm,k}$ denotes the true value of $\sigma_{jm,k}$. Note that $|\hat{\sigma}_{jm,k} - \sigma^*_{jm,k}| \leq \|\boldsymbol{\Theta}^*_k - \hat{\boldsymbol{\Theta}}_k\|_2 \leq \|\hat{\boldsymbol{\Upsilon}} - \boldsymbol{\Upsilon}^*\|_2$.

$$\mathcal{R}^*(\hat{\boldsymbol{\Upsilon}}_k) = \frac{1}{2n} \sum_{i=1}^n L_k(\boldsymbol{y}_i; \boldsymbol{x}_i, \hat{\boldsymbol{\Upsilon}}_k)[\sigma^*_{jm,k} - (y_{ij} - (\hat{\boldsymbol{\Gamma}}_k \boldsymbol{x}_i)_j)(y_{im} - (\hat{\boldsymbol{\Gamma}}_k \boldsymbol{x}_i)_m)],$$

which can be bounded as $|\mathcal{R}^*(\hat{\boldsymbol{\Upsilon}}_k)| \leq \frac{1}{2}|\mathcal{R}^*_1| + \frac{1}{2}|\mathcal{R}^*_2|$, where

$$\mathcal{R}^*_1 = \frac{1}{n} \sum_{i=1}^n L_k(\boldsymbol{y}_i; \boldsymbol{x}_i, \boldsymbol{\Upsilon}^*_k)[\sigma^*_{jm,k} - (y_{ij} - \boldsymbol{\Gamma}^*_k \boldsymbol{x}_i)_j)(y_{im} - (\boldsymbol{\Gamma}^*_k \boldsymbol{x}_i)_m)],$$

$$\mathcal{R}_2^* = \frac{1}{n}\sum_{i=1}^{n}[L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\hat{\boldsymbol{\Upsilon}}_k) - L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}_k^*)][(y_{ij}-(\hat{\boldsymbol{\Gamma}}_k\boldsymbol{x}_i)_j)(y_{im}-(\hat{\boldsymbol{\Gamma}}_k\boldsymbol{x}_i)_m)]$$
$$+\frac{1}{n}\sum_{i=1}^{n}L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}_k^*)[(y_{ij}-(\boldsymbol{\Gamma}_k^*\boldsymbol{x}_i)_j)((\boldsymbol{\Gamma}_k^*\boldsymbol{x}_i)_j-(\hat{\boldsymbol{\Gamma}}_k\boldsymbol{x}_i)_j)]$$
$$+\frac{1}{n}\sum_{i=1}^{n}L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}_k^*)[(y_{im}-(\boldsymbol{\Gamma}_k^*\boldsymbol{x}_i)_m)((\boldsymbol{\Gamma}_k^*\boldsymbol{x}_i)_m-(\hat{\boldsymbol{\Gamma}}_k\boldsymbol{x}_i)_m)]$$
$$+\frac{1}{n}\sum_{i=1}^{n}L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}_k^*)[((\boldsymbol{\Gamma}_k^*\boldsymbol{x}_i)_j-(\hat{\boldsymbol{\Gamma}}_k\boldsymbol{x}_i)_j)((\boldsymbol{\Gamma}_k^*\boldsymbol{x}_i)_m-(\hat{\boldsymbol{\Gamma}}_k\boldsymbol{x}_i)_m)].$$

Note that $L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\cdot)$ is continuous. Then,

$$\mathcal{R}_2^* = O\left\{\sup_{i,k}[L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\hat{\boldsymbol{\Upsilon}}_k) - L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}_k^*)] + \sup_k\|\hat{\boldsymbol{\Gamma}}_k-\boldsymbol{\Gamma}_k^*\|_\infty\right\} \lesssim \|\hat{\boldsymbol{\Upsilon}}-\boldsymbol{\Upsilon}^*\|_2,$$

where $a \lesssim b$ if $a \le Db$ for some positive constant D. As for $\mathcal{R}_1^*$, we have

$$\frac{1}{2n}\sum_{i=1}^{n}\mathbb{E}[L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}_k^*)](\boldsymbol{\Theta}_k^{*-1}) - \frac{1}{2n}\sum_{i=1}^{n}\mathbb{E}[L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}_k^*)(\boldsymbol{y}-\boldsymbol{\Gamma}_k^*\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{\Gamma}_k^*\boldsymbol{x})^T] = 0.$$

Thus, $\mathcal{R}_1^*$ can be rewritten as:

$$\mathcal{R}_1^* = \frac{1}{n}\sum_{i=1}^{n}\left(L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}_k^*)\sigma_{jm,k}^* - \mathbb{E}[L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}_k^*)]\sigma_{jm,k}^*\right)$$
$$-\frac{1}{n}\sum_{i=1}^{n}\left(L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}_k^*)[(y_{ij}-\boldsymbol{\Gamma}_k^*\boldsymbol{x}_i)_j)(y_{im}-(\boldsymbol{\Gamma}_k^*\boldsymbol{x}_i)_m)]\right.$$
$$\left.-\mathbb{E}[L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}_k^*)(\boldsymbol{y}-\boldsymbol{\Gamma}_k^*\boldsymbol{x})_j(\boldsymbol{y}-\boldsymbol{\Gamma}_k^*\boldsymbol{x})_m^T]\right),$$

which is the same as the bound of (S3.34) in the proof of Lemma 1, and we have $|\mathcal{R}_1^*| = O(\sqrt{\log p/n})$. In summary, it can be obtained that:

$$|\mathcal{R}(\hat{\boldsymbol{\Upsilon}}_k)| \lesssim \|\hat{\boldsymbol{\Upsilon}}_0-\boldsymbol{\Upsilon}^*\|_2 \lesssim \sqrt{\frac{(d+s+p)(\log p+\log q)}{n}}.$$

By Conditions (C5) and (C7), we have:

$$\lambda_1\rho'(|\hat{\theta}_{jm,k}|) \gg C_{\lambda_1}\sqrt{\frac{(d+s+p)(\log p+\log q)}{n}},$$

for $\hat{\theta}_{jm,k}$ in a small neighborhood of 0 and some positive constant C_{λ_1}. Therefore, if $\hat{\theta}_{jm,k}$ lies in a small neighborhood of 0, the sign of $\frac{\partial Q_n(\hat{\boldsymbol{\Upsilon}})}{\partial \theta_{jm,k}}$ only depends on $\text{sgn}(\hat{\theta}_{jm,k})$, with probability tending to 1. Then, variable selection consistency of the precision matrix estimators can be proved.

Result 2: Assume that $\lambda_3 \gg \sqrt{\frac{(s+p)(\log p+\log q)}{n}}$, $b = \min_{1\le k\neq k'\le K_0} \|\boldsymbol{\Upsilon}_k^* - \boldsymbol{\Upsilon}_{k'}^*\|_2 > a\lambda_3$, and the conditions in Result 1 hold. Then there exists $\hat{\boldsymbol{\Omega}}$, a local maximizer of $\mathcal{L}(\boldsymbol{\Omega}, \boldsymbol{\pi}|\mathbf{Y})$ defined in equation (2.1) that satisfies:

$$P(\hat{\boldsymbol{\Omega}} = \hat{\boldsymbol{\Omega}}^o) \to 1. \tag{S3.22}$$

Proof of Result 2: Denote $\hat{\boldsymbol{\Omega}}^o$ as the maximizer of (S3.1). According to Result 1, we have:

$$\|\hat{\boldsymbol{\Omega}}^o - \boldsymbol{\Omega}^*\|_2 = O_p\left(K\|\hat{\boldsymbol{\Upsilon}}^o - \boldsymbol{\Upsilon}^*\|_2\right) = O_p(\epsilon_n),$$

where $\epsilon_n = \sqrt{(d+s+p)(\log p+\log q)/n}$. Denote the locations of the non-zero coefficients in $\boldsymbol{\Omega}_l^*$ as $\mathcal{W}_l, l = 1,\ldots,K$. Consider two neighborhood sets of $\boldsymbol{\Omega}^*$:

$$\mathcal{C} = \left\{\boldsymbol{\Omega} \in \mathbb{R}^{K(p(q+1)+p^2)} : \sup_l \|\boldsymbol{\Omega}_l - \boldsymbol{\Omega}_l^*\|_2 \le \epsilon_n\right\},$$

$$\mathcal{C}_0 = \left\{\boldsymbol{\Omega} \in \mathbb{R}^{K(p(q+1)+p^2)} : \sup_l \|\boldsymbol{\Omega}_l - \boldsymbol{\Omega}_l^*\|_2 \le \epsilon_n, \Omega_{lj} = 0, j \notin \mathcal{W}_l, l = 1,\ldots,K\right\}.$$

By Result 1, there exists an event E_1 in which $\sup_l \|\hat{\boldsymbol{\Omega}}_l^o - \boldsymbol{\Omega}_l^*\|_2 \le \epsilon_n, \hat{\Omega}_{lj}^o = 0, j \notin \mathcal{W}_l, l = 1,\ldots,K$, and $P(E_1^C) \to 0$. Thus $\hat{\boldsymbol{\Omega}}^o \in \mathcal{C}_0$. Let $G : \Lambda_{\mathcal{T}^*} \to \mathbb{R}^{K_0(p(q+1)+p^2)}$ be the mapping such that $G(\boldsymbol{\Omega})$ is the $K_0(p(q+1)+p^2)$ vector consisting of the K_0 groups with each group having dimension $p(q+1)+p^2$, and its l-th vector component

equal to the common value of $\boldsymbol{\Omega}_l$ for $l \in \mathcal{T}_k^*$. Let $\breve{G} : \mathbb{R}^{K(p(q+1)+p^2)} \to \mathbb{R}^{K_0(p(q+1)+p^2)}$ be the mapping such that $\breve{G}(\boldsymbol{\Omega}) = \{|\mathcal{T}_k^*|^{-1} \sum_{l \in \mathcal{T}_k^*} \boldsymbol{\Omega}_l^T, l = 1, \ldots, K\}^T$. For any $\boldsymbol{\Omega}$, denote $\breve{\boldsymbol{\Omega}} = G^{-1}\left(\breve{G}(\boldsymbol{\Omega})\right) \in \Lambda_{\mathcal{T}^*}$. For any $\boldsymbol{\Omega}^{(0)}$, define $\boldsymbol{\Omega}$ with $\Omega_{lj} = \Omega_{lj}^{(0)}$ for $j \in \mathcal{W}_l$ and $\Omega_{lj} = 0$ for $j \notin \mathcal{W}_l$, $l = 1, \ldots, K$. Clearly, if $\boldsymbol{\Omega}^{(0)} \in \mathcal{C}$, then $\boldsymbol{\Omega} \in \mathcal{C}_0$.

With objective function (S3.1), remove the constraint of oracle group membership $\boldsymbol{\Omega} \in \Lambda_{\mathcal{T}^*}$:

$$\mathcal{Q}(\boldsymbol{\Omega}) = \frac{1}{n} \sum_{i=1}^{n} \log \left(\sum_{l=1}^{K} \pi_l f_l(\mathbf{y}_i; \boldsymbol{x}_i, \boldsymbol{\Gamma}_l, \boldsymbol{\Theta}_l) \right) - \sum_{l=1}^{K} \sum_{j \neq m} p(|\theta_{jm,l}|, \lambda_1) - \sum_{l=1}^{K} \sum_{j=1}^{p} \sum_{m=1}^{q+1} p(|\gamma_{jm,l}|, \lambda_2),$$

$$\mathcal{P}_3(\boldsymbol{\Omega}) = \sum_{1 \leq l < l' \leq K} p(\|\boldsymbol{\Omega}_l - \boldsymbol{\Omega}_{l'}\|_2, \lambda_3).$$

Then, $\mathcal{L}(\boldsymbol{\Omega}) = \mathcal{Q}(\boldsymbol{\Omega}) - \mathcal{P}_3(\boldsymbol{\Omega})$.

If we can show the following two results, then we have that $\hat{\boldsymbol{\Omega}}^o$ is a strict local maximizer of objective function (2.1) with probability converging to 1. With results (i) and (ii), for any $\boldsymbol{\Omega}^{(0)} \in \mathcal{C} \cap \mathcal{C}_n$ and $\boldsymbol{\Omega}^{(0)} \neq \hat{\boldsymbol{\Omega}}^o$ on $\mathcal{C} \cap \mathcal{C}_n$, where $\mathcal{C}_n$ is the neighborhood of $\hat{\boldsymbol{\Omega}}^o$, we have $\mathcal{L}(\boldsymbol{\Omega}^{(0)}) \leq \mathcal{L}(\hat{\boldsymbol{\Omega}}^o)$. So $\hat{\boldsymbol{\Omega}}^o$ is a strict local maximizer of objective function (2.1) on event E_1 with $P(E_1) \to 1$ and a sufficiently large n.

(i) On event E_1, for and $\breve{\boldsymbol{\Omega}} \in \mathcal{C}_0$ and $\breve{\boldsymbol{\Omega}} \neq \hat{\boldsymbol{\Omega}}^o$, $\mathcal{L}(\breve{\boldsymbol{\Omega}}) < \mathcal{L}(\hat{\boldsymbol{\Omega}}^o)$.

(ii) On event E_1, there is a neighborhood of $\hat{\boldsymbol{\Omega}}^o$, denoted as $\mathcal{C}_n$, such that $\mathcal{L}(\boldsymbol{\Omega}^{(0)}) \leq \mathcal{L}(\boldsymbol{\Omega}) \leq \mathcal{L}(\breve{\boldsymbol{\Omega}})$, for any $\boldsymbol{\Omega}^{(0)} \in \mathcal{C} \cap \mathcal{C}_n$ and a sufficiently large n.

By $\lambda_3 \gg \sqrt{\frac{(d+s+p)(\log p + \log q)}{n}}$ and $b > a\lambda_3$, the penalty function $\sum_{1 \leq l < l' \leq K} p(\|\boldsymbol{\Omega}_l - \boldsymbol{\Omega}_{l'}\|_2, \lambda_3)$ is a constant that does not depend on $\boldsymbol{\Omega}$. We further impose the constraint

$\boldsymbol{\Omega} \in \Lambda_{\mathcal{T}^*}$ on $\mathcal{Q}(\boldsymbol{\Omega})$ and $\mathcal{P}_3(\boldsymbol{\Omega})$, denoted as $\mathcal{Q}^T(\boldsymbol{\Omega})$ and $\mathcal{P}_3^T(\boldsymbol{\Omega})$, respectively. Define $\mathcal{L}^T(\boldsymbol{\Omega}) = \mathcal{Q}^T(\boldsymbol{\Omega}) - \mathcal{P}_3^T(\boldsymbol{\Omega})$. When $\breve{\boldsymbol{\Omega}} \in \Lambda_{\mathcal{T}^*}$, we have $\mathcal{L}(\breve{\boldsymbol{\Omega}}) = \mathcal{L}^T(\breve{\boldsymbol{\Omega}})$. Since $\mathcal{P}_3(\boldsymbol{\Omega})$ is a constant and $\hat{\boldsymbol{\Omega}}^o$ is the unique maximizer of $\mathcal{Q}^T(\boldsymbol{\Omega})$, for any $\breve{\boldsymbol{\Omega}} \in \mathcal{C}_0$, $\mathcal{Q}^T(\breve{\boldsymbol{\Omega}}) < \mathcal{Q}^T(\hat{\boldsymbol{\Omega}}^o)$. Therefore, $\mathcal{L}(\breve{\boldsymbol{\Omega}}) < \mathcal{L}(\hat{\boldsymbol{\Omega}}^o)$, and the result in (i) is proved.

Next, we prove result (ii). First, we show that $\mathcal{L}(\boldsymbol{\Omega}^{(0)}) \leq \mathcal{L}(\boldsymbol{\Omega})$. Given a positive sequence ϕ_n, consider:

$$\mathcal{C}_n = \left\{ \boldsymbol{\Omega} \in \mathbb{R}^{K(p(q+1)+p^2)} : \sup_l \|\boldsymbol{\Omega}_l - \hat{\boldsymbol{\Omega}}_l^o|_\infty \leq \phi_n \right\}.$$

For any $\boldsymbol{\Omega}^{(0)} \in \mathcal{C} \cap \mathcal{C}_n$ and the corresponding $\boldsymbol{\Omega} \in \mathcal{C}_0 \cap \mathcal{C}_n$, $\|\boldsymbol{\Omega}_l^{(0)} - \boldsymbol{\Omega}_{l'}^{(0)}\|_2 \geq \|\boldsymbol{\Omega}_l - \boldsymbol{\Omega}_{l'}\|_2$. Therefore, $-\mathcal{P}_3(\boldsymbol{\Omega}^{(0)}) \leq -\mathcal{P}_3(\boldsymbol{\Omega})$. Moreover, for the non-zero part $\mathcal{W}$, $\boldsymbol{\Omega}^{(0)} = \boldsymbol{\Omega}$, and the difference between $\boldsymbol{\Omega}^{(0)}$ and $\boldsymbol{\Omega}$ lies in the zero entries of $\boldsymbol{\Omega}$ in $\mathcal{W}^C$. And the corresponding values of $\boldsymbol{\Omega}^{(0)}$ are small enough and can be controlled by ϕ_n. According to the proof of selection consistency, when $\hat{\theta}_{jm,l}$ or $\hat{\gamma}_{jm,l}$ lies in a small neighborhood of 0, $\partial \mathcal{Q}(\hat{\boldsymbol{\Upsilon}})/\partial \theta_{jm,l} < 0$, which indicates that $\mathcal{Q}(\hat{\boldsymbol{\Upsilon}})$ is a decreasing function in terms of $\hat{\theta}_{jm,l}$ and $\hat{\gamma}_{jm,l}$ in the neighborhood of 0. Thus $\mathcal{Q}(\boldsymbol{\Omega}^{(0)}) \leq \mathcal{Q}(\boldsymbol{\Omega})$. Therefore, $\mathcal{L}(\boldsymbol{\Omega}^{(0)}) \leq \mathcal{L}(\boldsymbol{\Omega})$.

Second, we show that $\mathcal{L}(\boldsymbol{\Omega}) \leq \mathcal{L}(\breve{\boldsymbol{\Omega}})$. For $\boldsymbol{\Omega} \in \mathcal{C}_0 \cap \mathcal{C}_n$ and $\breve{\boldsymbol{\Omega}} \in \mathcal{C}_0$, by Taylor's expansion,

$$\mathcal{L}(\boldsymbol{\Omega}) - \mathcal{L}(\breve{\boldsymbol{\Omega}}) = l_1 - l_2,$$

where

$$l_1 = \sum_{l=1}^{K} \mathbf{D}_l^T(\mathbf{\Omega}_l - \breve{\mathbf{\Omega}}_l), \mathbf{D}_l^T = (\mathbf{D}_{1l}^T, \mathbf{D}_{2l}^T)^T \mathcal{I}_{\mathcal{W}_l},$$

$$\mathbf{D}_{1l} = \frac{1}{n}\sum_{i=1}^{n} L_l(\mathbf{y}_i; \boldsymbol{x}_i, \tilde{\mathbf{\Omega}}_l)\tilde{\mathbf{\Theta}}_l(\mathbf{y}_i - \tilde{\mathbf{\Gamma}}_l\mathbf{x}_i)\mathbf{x}_i - \lambda_2 \mathbf{s}_2(\tilde{\mathbf{\Gamma}}_l),$$

$$\mathbf{D}_{2l} = \frac{1}{n}\sum_{i=1}^{n} L_l(\mathbf{y}_i; \boldsymbol{x}_i, \tilde{\mathbf{\Omega}}_l)\frac{1}{2}\left[\mathrm{vec}(\tilde{\mathbf{\Theta}}_l)^{-1} - \mathrm{vec}(\mathbf{y}_i - \tilde{\mathbf{\Gamma}}_l\mathbf{x}_i)(\mathbf{y}_i - \tilde{\mathbf{\Gamma}}_l\mathbf{x}_i)\right] - \lambda_1 \mathbf{s}_1(\tilde{\mathbf{\Theta}}_l),$$

$$l_2 = \sum_{l=1}^{K} \frac{\partial \mathcal{P}_3(\tilde{\mathbf{\Omega}})}{\partial \mathbf{\Omega}_l^T}(\mathbf{\Omega}_l - \breve{\mathbf{\Omega}}_l),$$

and $\mathcal{I}_{\mathcal{W}_l}$ is a diagonal matrix with the j-th diagonal element $I(j \in \mathcal{W}_l), j = 1, \ldots, p(q+1) + p^2$, and $L_l(\mathbf{y}_i; \boldsymbol{x}_i, \tilde{\mathbf{\Omega}}_l) = \frac{\pi_l f_l(\mathbf{y}_i; \boldsymbol{x}_i, \tilde{\mathbf{\Omega}})}{\sum_{l=1}^{K} \pi_l f_l(\mathbf{y}_i; \boldsymbol{x}_i, \tilde{\mathbf{\Omega}})}$. $\mathbf{s}_1(\tilde{\mathbf{\Theta}}_l) \in \mathbb{R}^{p^2}$ such that the corresponding element is $\rho'(|\tilde{\theta}_{jm,l}|)\mathrm{sgn}(\tilde{\theta}_{jm,l})I(j \neq m)$, $\mathbf{s}_2(\tilde{\mathbf{\Gamma}}_l) \in \mathbb{R}^{p(q+1)}$ such that the corresponding element is $\rho'(|\tilde{\gamma}_{jm,l}|)\mathrm{sgn}(\tilde{\gamma}_{jm,l})I(j \neq m)$, and $\tilde{\mathbf{\Omega}} = \varsigma\mathbf{\Omega} + (1-\varsigma)\breve{\mathbf{\Omega}}$ for some constant $\varsigma \in (0, 1)$.

For l_2, we can obtain:

$$l_2 \geq \sum_{k=1}^{K_0} \sum_{\{l,l' \in \mathcal{T}_k^*, l<l'\}} \lambda_3 \rho'(4\phi_n)\|\mathbf{\Omega}_l - \mathbf{\Omega}_{l'}\|_2. \tag{S3.23}$$

For l_1, we can obtain :

$$l_1 \leq \sum_{k=1}^{K_0} \sum_{\{l,l' \in \mathcal{T}_k^*, l<l'\}} |\mathcal{T}_{\min}|^{-1} \sup_l \|\mathbf{D}_l - \mathbf{D}_{l'}\|_2 \|\mathbf{\Omega}_l - \mathbf{\Omega}_{l'}\|_2. \tag{S3.24}$$

It can be shown that:

$$\sup_l \|\mathbf{D}_l - \mathbf{D}_{l'}\|_2 \leq \sqrt{d \cdot \sup_l \|\mathbf{D}_{1l} - \mathbf{D}_{1l'}\|_\infty^2 + s \cdot \sup_l \|\mathbf{D}_{2l} - \mathbf{D}_{2l'}\|_\infty^2}. \tag{S3.25}$$

According to the minimal signal condition and Condition (C7), we have $\mathbf{s}_1(\tilde{\mathbf{\Theta}}_l)\mathcal{I}_{\mathcal{W}_l} = \mathbf{0}$ and $\mathbf{s}_2(\tilde{\mathbf{\Gamma}}_l)\mathcal{I}_{\mathcal{W}_l} = \mathbf{0}$. Note that:

$$\begin{aligned}
\|\mathbf{D}_{1l} - \mathbf{D}_{1l'}\|_\infty &\leq \sup_i \|L_l(\mathbf{y}_i; \boldsymbol{x}_i, \tilde{\mathbf{\Omega}}_l) m_1(\mathbf{y}_i; \mathbf{x}_i, \tilde{\mathbf{\Omega}}_l) - L_l(\mathbf{y}_i; \boldsymbol{x}_i, \tilde{\mathbf{\Omega}}_{l'}) m_1(\mathbf{y}_i; \mathbf{x}_i, \tilde{\mathbf{\Omega}}_{l'})\|_\infty \\
&= \sup_i \|L_l(\mathbf{y}_i; \boldsymbol{x}_i, \tilde{\mathbf{\Omega}}_l) \left[m_1(\mathbf{y}_i; \mathbf{x}_i, \tilde{\mathbf{\Omega}}_l) - m_1(\mathbf{y}_i; \mathbf{x}_i, \tilde{\mathbf{\Omega}}_{l'}) \right] \\
&\quad + \left[L_l(\mathbf{y}_i; \boldsymbol{x}_i, \tilde{\mathbf{\Omega}}_l) - L_l(\mathbf{y}_i; \boldsymbol{x}_i, \tilde{\mathbf{\Omega}}_{l'}) \right] m_1(\mathbf{y}_i; \mathbf{x}_i, \tilde{\mathbf{\Omega}}_l)\|_\infty \\
&\leq \sup_i \{L_l(\mathbf{y}_i; \boldsymbol{x}_i, \tilde{\mathbf{\Omega}}_l)\} \sup_i \|m_1(\mathbf{y}_i; \mathbf{x}_i, \tilde{\mathbf{\Omega}}_l) - m_1(\mathbf{y}_i; \mathbf{x}_i, \tilde{\mathbf{\Omega}}_{l'})\|_\infty \\
&\quad + \sup_i \{L_l(\mathbf{y}_i; \boldsymbol{x}_i, \tilde{\mathbf{\Omega}}_l) - L_l(\mathbf{y}_i; \boldsymbol{x}_i, \tilde{\mathbf{\Omega}}_{l'})\} \sup_i \|m_1(\mathbf{y}_i; \mathbf{x}_i, \tilde{\mathbf{\Omega}}_{l'})\|_\infty,
\end{aligned}$$

where $m_1(\mathbf{y}_i; \mathbf{x}_i, \tilde{\mathbf{\Omega}}_l) = \tilde{\mathbf{\Theta}}_l(\mathbf{y}_i - \tilde{\mathbf{\Gamma}}_l \mathbf{x}_i)\mathbf{x}_i$. In addition, $L_l(\mathbf{y}_i; \boldsymbol{x}_i, \cdot)$ and $m_1(\mathbf{y}_i; \mathbf{x}_i, \cdot)$ are continuous. Then, there exists a constant $C' > 0$,

$$\begin{aligned}
\sup_i \|m_1(\mathbf{y}_i; \mathbf{x}_i, \tilde{\mathbf{\Omega}}_l) - m_1(\mathbf{y}_i; \mathbf{x}_i, \tilde{\mathbf{\Omega}}_{l'})\|_\infty &\leq C' \|\tilde{\mathbf{\Omega}}_l - \tilde{\mathbf{\Omega}}_{l'}\|_\infty, \\
\sup_i \{L_l(\mathbf{y}_i; \boldsymbol{x}_i, \tilde{\mathbf{\Omega}}_l) - L_l(\mathbf{y}_i; \boldsymbol{x}_i, \tilde{\mathbf{\Omega}}_{l'})\} &\leq C' \|\tilde{\mathbf{\Omega}}_l - \tilde{\mathbf{\Omega}}_{l'}\|_\infty.
\end{aligned}$$

Note that, for $l, l' \in \mathcal{T}_k^*$, $\mathbf{\Omega}_l, \mathbf{\Omega}_{l'} \in \mathcal{C}_n$. Then,

$$\|\tilde{\mathbf{\Omega}}_l - \tilde{\mathbf{\Omega}}_{l'}\|_\infty = \varsigma \|\mathbf{\Omega}_l - \mathbf{\Omega}_{l'}\|_\infty \leq 2\phi_n.$$

Thus, we have:

$$\sup_l \|\mathbf{D}_{1l} - \mathbf{D}_{1l'}\|_\infty = O(\phi_n).$$

Similarly,

$$\begin{aligned}
\|\mathbf{D}_{2l} - \mathbf{D}_{2l'}\|_\infty &\leq \sup_i \{L_l(\mathbf{y}_i; \boldsymbol{x}_i, \tilde{\mathbf{\Omega}}_l)\} \sup_i \|m_2(\mathbf{y}_i; \mathbf{x}_i, \tilde{\mathbf{\Omega}}_l) - m_2(\mathbf{y}_i; \mathbf{x}_i, \tilde{\mathbf{\Omega}}_{l'})\|_\infty \\
&\quad + \sup_i \{L_l(\mathbf{y}_i; \boldsymbol{x}_i, \tilde{\mathbf{\Omega}}_l) - L_l(\mathbf{y}_i; \boldsymbol{x}_i, \tilde{\mathbf{\Omega}}_{l'})\} \sup_i \|m_2(\mathbf{y}_i; \mathbf{x}_i, \tilde{\mathbf{\Omega}}_{l'})\|_\infty,
\end{aligned}$$

where $m_2(\mathbf{y}_i;\mathbf{x}_i,\tilde{\mathbf{\Omega}}_l)=\frac{1}{2}\left[\text{vec}(\tilde{\mathbf{\Theta}}_l)^{-1}-\text{vec}(\mathbf{y}_i-\tilde{\mathbf{\Gamma}}_l\mathbf{x}_i)(\mathbf{y}_i-\tilde{\mathbf{\Gamma}}_l\mathbf{x}_i)\right]$. In addition, $m_2(\mathbf{y}_i;\mathbf{x}_i,\cdot)$ is continuous. Thus, we have:

$$\sup_l \|\mathbf{D}_{2l}-\mathbf{D}_{2l'}\|_\infty = O(\phi_n).$$

Further, we can obtain that:

$$\mathcal{L}(\mathbf{\Omega})-\mathcal{L}(\breve{\mathbf{\Omega}})\leq\sum_{k=1}^{K_0}\sum_{\{l,l'\in\mathcal{T}_k^*,l<l'\}}\left[\phi_n|\mathcal{T}_{\min}|^{-1}-\lambda_3\rho'(4\phi_n)\right]\|\mathbf{\Omega}_l-\mathbf{\Omega}_{l'}\|_2. \tag{S3.26}$$

Let $\phi_n=\epsilon_n$. Then $\rho'(4\phi_n)\to 1$ according to Condition (C7). Since $\lambda_3\gg\epsilon_n$ and $|\mathcal{T}_{\min}|\geq 1$, we have $|\mathcal{T}_{\min}|\lambda_3\gg\phi_n$, and then $\mathcal{L}(\mathbf{\Omega})-\mathcal{L}(\breve{\mathbf{\Omega}})\leq 0$. This completes the proof of result (ii) and thus Result 2.

□

Lemma 1: Under Conditions (C2) and (C3), define $L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\mathbf{\Upsilon})=\frac{\phi_k f_k(\boldsymbol{y}_i|\boldsymbol{x}_i,\mathbf{\Upsilon}_k)}{\sum_{k=1}^{K_0}\pi_k f_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\mathbf{\Upsilon}_k)}$,

$$H_n(\mathbf{\Upsilon}^*|\mathbf{\Upsilon}^{(t-1)})=\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K_0}L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\mathbf{\Upsilon}^{(t-1)})\left[\log\pi_k+\log f_k(\boldsymbol{y}_i|\boldsymbol{x}_i,\mathbf{\Upsilon}_k^*)\right],$$

and the population version of the log-likelihood function:

$$H(\mathbf{\Upsilon}^*|\mathbf{\Upsilon}^{(t-1)})=\mathbb{E}\left(\sum_{k=1}^{K_0}L_k(\boldsymbol{y};\boldsymbol{x},\mathbf{\Upsilon}^{(t-1)})\left[\log\pi_k+\log f_k(\boldsymbol{y}|\boldsymbol{x},\mathbf{\Upsilon}_k^*)\right]\right).$$

We have

$$\|\nabla H_n(\mathbf{\Upsilon}^*|\mathbf{\Upsilon}^{(t-1)})-\nabla H(\mathbf{\Upsilon}^*|\mathbf{\Upsilon}^{(t-1)})\|_\infty=O_p(\sqrt{(\log p+\log q)/n}). \tag{S3.27}$$

Proof. Recall that:

$$\|\nabla H_n(\boldsymbol{\Upsilon}^*|\boldsymbol{\Upsilon}) - \nabla H(\boldsymbol{\Upsilon}^*|\boldsymbol{\Upsilon})\|_\infty \leq \max_{1\leq k\leq K_0} \|\nabla_{\boldsymbol{\Gamma}_k^*} H_n(\boldsymbol{\Upsilon}^*|\boldsymbol{\Upsilon}) - \nabla_{\boldsymbol{\Gamma}_k^*} H(\boldsymbol{\Upsilon}^*|\boldsymbol{\Upsilon})\|_\infty + \max_{1\leq k\leq K_0} \|\nabla_{\boldsymbol{\Theta}_k^*} H_n(\boldsymbol{\Upsilon}^*|\boldsymbol{\Upsilon}) - \nabla_{\boldsymbol{\Theta}_k^*} H(\boldsymbol{\Upsilon}^*|\boldsymbol{\Upsilon})\|_\infty. \tag{S3.28}$$

Define $h_{\boldsymbol{\Gamma}_k^*}(\boldsymbol{\Upsilon}^*) = \nabla_{\boldsymbol{\Gamma}_k^*} H_n(\boldsymbol{\Upsilon}^*|\boldsymbol{\Upsilon}) - \nabla_{\boldsymbol{\Gamma}_k^*} H(\boldsymbol{\Upsilon}^*|\boldsymbol{\Upsilon})$. We have:

$$\begin{aligned}
\|h_{\boldsymbol{\Gamma}_k^*}(\boldsymbol{\Upsilon}^*)\|_\infty &= \left\| \frac{1}{n}\sum_{i=1}^n L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}^*)\boldsymbol{\Theta}_k^*(\boldsymbol{y}_i - \boldsymbol{\Gamma}_k^*\boldsymbol{x}_i)\boldsymbol{x}_i^T - \mathbb{E}[L_k(\boldsymbol{y};\boldsymbol{x},\boldsymbol{\Upsilon}^*)\boldsymbol{\Theta}_k^*(\boldsymbol{y}-\boldsymbol{\Gamma}_k^*\boldsymbol{x})\boldsymbol{x}^T] \right\|_\infty \\
&\leq \|\boldsymbol{\Theta}_k^*\|_\infty \left\| \frac{1}{n}\sum_{i=1}^n L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}^*)\boldsymbol{y}_i\boldsymbol{x}_i^T - \mathbb{E}[L_k(\boldsymbol{y};\boldsymbol{x},\boldsymbol{\Upsilon}^*)\boldsymbol{y}\boldsymbol{x}^T] \right\|_\infty \\
&\quad + \|\boldsymbol{\Theta}_k^*\|_\infty \|\boldsymbol{\Gamma}_k^*\|_\infty \left\| \frac{1}{n}\sum_{i=1}^n L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}^*)\boldsymbol{x}_i\boldsymbol{x}_i^T - \mathbb{E}[L_k(\boldsymbol{y};\boldsymbol{x},\boldsymbol{\Upsilon}^*)\boldsymbol{x}\boldsymbol{x}^T] \right\|_\infty \\
&\triangleq \|\boldsymbol{\Theta}_k^*\|_\infty \|I_1\|_\infty + \|\boldsymbol{\Theta}_k^*\|_\infty \|\boldsymbol{\Gamma}_k^*\|_\infty \|I_2\|_\infty.
\end{aligned} \tag{S3.29}$$

Define $\{c_i, i = 1,\ldots,n\}$, which are independent copies of random variable c. Here $c = k', k' = 1,\ldots,K_0$, indicates that $\boldsymbol{y}$ is generated by the k'-th mixture component, that is, $\boldsymbol{y}|c = k' \sim N(\boldsymbol{\Gamma}_{k'}^*\boldsymbol{x}, \boldsymbol{\Theta}_{k'}^{*-1})$. The j-th coordinate of $\boldsymbol{y}_i$, y_{ij}, can be rewritten as:

$$y_{ij} = \sum_{k'=1}^{K_0} I(c_i = k')[(\boldsymbol{\Gamma}_{k'}^*\boldsymbol{x}_i)_j + V_{k'j}],\ j = 1,\ldots,p,$$

where $V_{k'j} \sim N(0, \boldsymbol{\Sigma}_{k'jj}^*)$. Define:

$$\zeta_{jj'} = \frac{1}{n}\sum_{i=1}^n L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}^*)y_{ij}x_{ij'} - \mathbb{E}[L_k(\boldsymbol{y};\boldsymbol{x},\boldsymbol{\Upsilon}^*)y_j x_{j'}], j = 1,\ldots,p, j' = 1,\ldots,q.$$

Note that:

$$y_{ij}x_{ij'} = \sum_{k'=1}^{K_0} I(c_i = k') \left((\mathbf{\Gamma}^*_{k'}\boldsymbol{x}_i)_j x_{ij'} + V_{k'j}x_{ij'} \right).$$

Define $\zeta_{jj',1} = \frac{1}{n}\sum_{i=1}^n L_k(\boldsymbol{y}_i; \boldsymbol{x}_i, \mathbf{\Upsilon}^*) I(c_i = k')(\mathbf{\Gamma}^*_{k'}\boldsymbol{x}_i)_j x_{ij'} - \mathbb{E}[L_k(\boldsymbol{y}; \boldsymbol{x}, \mathbf{\Upsilon}^*) I(c_i = k')(\mathbf{\Gamma}^*_{k'}\boldsymbol{x})_j x_{j'}]$. $L_k(\boldsymbol{y}_i; \boldsymbol{x}_i, \mathbf{\Upsilon}^*) I(c_i = k')(\mathbf{\Gamma}^*_{k'}\boldsymbol{x}_i)_j x_{ij'}$ is a sub-gaussian random variable with $\|L_k(\boldsymbol{y}_i; \boldsymbol{x}_i, \mathbf{\Upsilon}^*) I(c_i = k')(\mathbf{\Gamma}^*_{k'}\boldsymbol{x}_i)_j x_{ij'}\|_{\psi_2} \leq m\|\mathbf{\Gamma}^*_{k'}\|_\infty$ by the boundedness condition $|x_{ij}| \leq m$, where $\|\cdot\|_{\psi_2}$ denotes the sub-gaussian norm. According to Lemma S.5 in Hao et al. (2018), we have:

$$\|L_k(\boldsymbol{y}_i; \boldsymbol{x}_i, \mathbf{\Upsilon}^*) I(c_i = k')(\mathbf{\Gamma}^*_{k'}\boldsymbol{x}_i)_j x_{ij'} - \mathbb{E}[L_k(\boldsymbol{y}; \boldsymbol{x}, \mathbf{\Upsilon}^*) I(c_i = k')(\mathbf{\Gamma}^*_{k'}\boldsymbol{x})_j x_{j'}]\|_{\psi_2} \leq 2m\|\mathbf{\Gamma}^*_{k'}\|_\infty.$$

Thus, for any $t > 0$ and some constant D_1, according to Lemma S.6 in Hao et al. (2018),

$$P(|\zeta_{jj',1}| \leq t) \geq 1 - e\exp\left(-\frac{D_1 n t^2}{4m^2\|\mathbf{\Gamma}^*_{k'}\|^2_\infty}\right).$$

Let t take a suitable value. Taking a union bound over pq coordinates, we have:

$$\sup_{j\in[p], j'\in[q]} |\zeta_{jj',1}| \leq \sqrt{\frac{4}{D_1}} m\|\mathbf{\Gamma}^*_{k'}\|_\infty \sqrt{\frac{\log p + \log q + \log(e/\delta)}{n}}, \tag{S3.30}$$

with probability at least $1 - \delta$.

Define $\zeta_{jj',2} = \frac{1}{n}\sum_{i=1}^n L_k(\boldsymbol{y}_i; \boldsymbol{x}_i, \mathbf{\Upsilon}^*) I(c_i = k') x_{ij'} V_{k',j} - \mathbb{E}[L_{ik}(\boldsymbol{y}) I(c_i = k') x_{j'} V_{k',j}]$. Since $L_k(\boldsymbol{y}_i; \boldsymbol{x}_i, \mathbf{\Upsilon}^*) I(c_i = k') x_{ij'} V_{k',j}$ is sub-exponential with norm $(m\|\mathbf{\Sigma}^*_{k'}\|_\infty)^{1/2}$, for some constant D_2,

$$\sup_{j\in[p], j'\in[q]} |\zeta_{jj',2}| \leq \sqrt{\frac{4}{D_2}} (m\|\mathbf{\Sigma}^*_{k'}\|_\infty)^{1/2} \sqrt{\frac{\log p + \log q + \log(2/\delta)}{n}}, \tag{S3.31}$$

with probability at least $1-\delta$ for a sufficient large n.

Note that:

$$\sup_{j\in[p],j'\in[q]}|\zeta_{jj'}|\leq\sum_{k'=1}^{K_0}\left(\sup_{j\in[p],j'\in[q]}|\zeta_{jj',1}|\right)+\sum_{k'=1}^{K_0}\left(\sup_{j\in[p],j'\in[q]}|\zeta_{jj',2}|\right).$$

Based on (S3.30) and (S3.31), we can bound $\zeta_{jj'}$ and further $\|I_1\|_\infty$ in (S3.29):

$$\|I_1\|_\infty\leq\sqrt{\frac{1}{D_3}}\phi_{K_0}\sqrt{\frac{\log p+\log q+\log(e/\delta)}{n}},\tag{S3.32}$$

with probability at least $1-4K_0\delta$, where $\phi_{K_0}=\sum_{k'=1}^{K_0}m\left(\|\boldsymbol{\Gamma}_{k'}^*\|_\infty+(\|\boldsymbol{\Sigma}_{k'}^*\|_\infty)^{1/2}\right)^2$ and $D_3=\min(D_1,D_2)$. According to (S3.10), I_2 is bounded by:

$$\|I_2\|_\infty\leq m\sqrt{\frac{1}{2n}\log(2/\delta)},\tag{S3.33}$$

with probability at least $1-\delta$.

Define $h_{\boldsymbol{\Theta}_k^*}(\boldsymbol{\Upsilon}^*)=\nabla_{\boldsymbol{\Theta}_k^*}H_n(\boldsymbol{\Upsilon}^*|\boldsymbol{\Upsilon})-\nabla_{\boldsymbol{\Theta}_k^*}H(\boldsymbol{\Upsilon}^*|\boldsymbol{\Upsilon})$. Then,

$$\begin{aligned}
\|h_{\boldsymbol{\Theta}_k^*}(\boldsymbol{\Upsilon}^*)\|_\infty=&\left\|\frac{1}{2n}\sum_{i=1}^n L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}^*)\boldsymbol{\Sigma}_k^*-\frac{1}{2n}\sum_{i=1}^n L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}^*)(\boldsymbol{y}_i-\boldsymbol{\Gamma}_k^*\boldsymbol{x}_i)(\boldsymbol{y}_i-\boldsymbol{\Gamma}_k^*\boldsymbol{x}_i)^T\right.\\
&\left.-\frac{1}{2}\mathbb{E}[L_k(\boldsymbol{y};\boldsymbol{x},\boldsymbol{\Upsilon}^*)]\boldsymbol{\Sigma}_k^*+\frac{1}{2}\mathbb{E}\left[L_k(\boldsymbol{y};\boldsymbol{x},\boldsymbol{\Upsilon}^*)(\boldsymbol{y}-\boldsymbol{\Gamma}_k^*\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{\Gamma}_k^*\boldsymbol{x})^T\right]\right\|_\infty\\
\leq&\left\|\frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^n L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}^*)\boldsymbol{\Sigma}_k^*-\mathbb{E}[L_k(\boldsymbol{y};\boldsymbol{x},\boldsymbol{\Upsilon}^*)\boldsymbol{\Sigma}_k^*]\right)\right\|_\infty\\
&+\left\|\frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^n L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}^*)(\boldsymbol{y}_i-\boldsymbol{\Gamma}_k^*\boldsymbol{x}_i)(\boldsymbol{y}_i-\boldsymbol{\Gamma}_k^*\boldsymbol{x}_i)^T\right.\right.\\
&\left.\left.-\mathbb{E}\left[L_k(\boldsymbol{y};\boldsymbol{x},\boldsymbol{\Upsilon}^*)(\boldsymbol{y}-\boldsymbol{\Gamma}_k^*\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{\Gamma}_k^*\boldsymbol{x})^T\right]\right)\right\|_\infty.
\end{aligned}\tag{S3.34}$$

According to (S3.10), the first term is bounded by:

$$\left\|\frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n}L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}^*)\boldsymbol{\Sigma}_k^*-\mathbb{E}[L_k(\boldsymbol{y};\boldsymbol{x},\boldsymbol{\Upsilon}^*)\boldsymbol{\Sigma}_k^*]\right)\right\|_\infty \leq \|\boldsymbol{\Sigma}_k^*\|_\infty\sqrt{\frac{1}{2n}\log(2/\delta)}, \tag{S3.35}$$

with probability at least $1-\delta$. For the second term, denoted as II_2:

$$\begin{aligned} II_2 \leq & \frac{1}{2}\left\|\frac{1}{n}\sum_{i=1}^{n}L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}^*)\boldsymbol{y}_i\boldsymbol{y}_i^T-\mathbb{E}[L_k(\boldsymbol{y};\boldsymbol{x},\boldsymbol{\Upsilon}^*)\boldsymbol{y}\boldsymbol{y}^T]\right\|_\infty \\ & +\|\boldsymbol{\Gamma}_k^*\|_\infty\left\|\frac{1}{n}\sum_{i=1}^{n}L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}^*)\boldsymbol{y}_i\boldsymbol{x}_i^T-\mathbb{E}[L_k(\boldsymbol{y};\boldsymbol{x},\boldsymbol{\Upsilon}^*)\boldsymbol{y}\boldsymbol{x}^T]\right\|_\infty \\ & +\frac{1}{2}\|\boldsymbol{\Gamma}_k^*\|_\infty^2\left\|\frac{1}{n}\sum_{i=1}^{n}L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}^*)\boldsymbol{x}_i\boldsymbol{x}_i^T-\mathbb{E}[L_k(\boldsymbol{y};\boldsymbol{x},\boldsymbol{\Upsilon}^*)\boldsymbol{x}\boldsymbol{x}^T]\right\|_\infty. \end{aligned} \tag{S3.36}$$

Define $\xi_{jj'}=\frac{1}{n}\sum_{i=1}^{n}L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}^*)y_{ij}y_{ij'}-\mathbb{E}[L_k(\boldsymbol{y};\boldsymbol{x},\boldsymbol{\Upsilon}^*)y_jy_{j'}], j,j'=1,\ldots,p$. Note that:

$$y_{ij}y_{ij'}=\sum_{k'=1}^{K_0}I(c_i=k')\left[(\boldsymbol{\Gamma}_{k'}^*\boldsymbol{x}_i)_j(\boldsymbol{\Gamma}_{k'}^*\boldsymbol{x}_i)_{j'}+(\boldsymbol{\Gamma}_{k'}^*\boldsymbol{x}_i)_jV_{k'j'}+(\boldsymbol{\Gamma}_{k'}^*\boldsymbol{x}_i)_{j'}V_{k'j}+V_{k'j}V_{k'j'}\right].$$

Define $\xi_{jj',1}=\frac{1}{n}\sum_{i=1}^{n}L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}^*)I(c_i=k')(\boldsymbol{\Gamma}_{k'}^*\boldsymbol{x}_i)_j(\boldsymbol{\Gamma}_{k'}^*\boldsymbol{x}_i)_{j'}-\mathbb{E}[L_k(\boldsymbol{y};\boldsymbol{x},\boldsymbol{\Upsilon}^*)I(c_i=k')(\boldsymbol{\Gamma}_{k'}^*\boldsymbol{x})_j(\boldsymbol{\Gamma}_{k'}^*\boldsymbol{x})_{j'}]$. Since $L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}^*)I(c_i=k')(\boldsymbol{\Gamma}_{k'}^*\boldsymbol{x}_i)_j(\boldsymbol{\Gamma}_{k'}^*\boldsymbol{x}_i)_{j'}$ is sub-gaussian with a bounded norm $m\|\boldsymbol{\Gamma}_{k'}^*\|_\infty^2$, for some constant D_4,

$$\sup_{j,j'\in[p]}|\xi_{jj',1}|\leq\sqrt{\frac{4}{D_4}}m\|\boldsymbol{\Gamma}_{k'}^*\|_\infty^2\sqrt{\frac{2\log p+\log(e/\delta)}{n}},$$

with probability at least $1-\delta$.

Define $\xi_{jj',2}=\frac{1}{n}\sum_{i=1}^{n}L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}^*)I(c_i=k')(\boldsymbol{\Gamma}_{k'}^*\boldsymbol{x}_i)_jV_{k'j'}-\mathbb{E}[L_k(\boldsymbol{y};\boldsymbol{x},\boldsymbol{\Upsilon}^*)I(c_i=k')(\boldsymbol{\Gamma}_{k'}^*\boldsymbol{x})_jV_{k'j'}]$. Since $L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Upsilon}^*)I(c_i=k')(\boldsymbol{\Gamma}_{k'}^*\boldsymbol{x}_i)_jV_{k'j'}$ is sub-exponential with a

bounded norm $\|\mathbf{\Gamma}^*_{k'}\|_\infty(m\|\mathbf{\Sigma}^*_{k'}\|_\infty)^{1/2}$, for some constant D_5,

$$\sup_{j,j'\in[p]}|\xi_{jj',2}| \leq \sqrt{\frac{4}{D_5}}\|\mathbf{\Gamma}^*_{k'}\|_\infty(m\|\mathbf{\Sigma}^*_{k'}\|_\infty)^{1/2}\sqrt{\frac{2\log p+\log(2/\delta)}{n}},$$

with probability at least $1-\delta$.

Define $\xi_{jj',3} = \frac{1}{n}\sum_{i=1}^n L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\mathbf{\Upsilon}^*)I(c_i = k')V_{k'j}V_{k'j'} - \mathbb{E}[L_k(\boldsymbol{y};\boldsymbol{x},\mathbf{\Upsilon}^*)I(c_i = k')V_{k'j}V_{k'j'}]$. Since $L_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\mathbf{\Upsilon}^*)I(c_i = k')V_{k'j}V_{k'j'}$ is sub-exponential with parameter $\|\mathbf{\Sigma}^*_{k'}\|_\infty$, there exists some constant D_6,

$$\sup_{j,j'\in[p]}|\xi_{jj',3}| \leq \sqrt{\frac{4}{D_6}}\|\mathbf{\Sigma}^*_{k'}\|_\infty\sqrt{\frac{2\log p+\log(2/\delta)}{n}},$$

with probability at least $1-\delta$ and a sufficiently large n.

Similar to the upper bound for $\|I_1\|_\infty$ in (S3.29), II_2 is bounded by:

$$II_2 \leq \sqrt{\frac{1}{D_7}}\phi'_{K_0}\sqrt{\frac{\log p+\log(e/\delta)}{n}}, \tag{S3.37}$$

with probability at least $1-16K_0\delta$, where $\phi'_{K_0} = \sum_{k'=1}^{K_0}\left(m\|\mathbf{\Gamma}^*_{k'}\|_\infty + (\|\mathbf{\Sigma}^*_{k'}\|_\infty)^{1/2}\right)^2$. Combining (S3.28), (S3.29), (S3.32), (S3.33), (S3.34), (S3.35), (S3.37), and the bounded conditions (C2) and (C3), we have:

$$\|\nabla H_n(\mathbf{\Upsilon}^*|\mathbf{\Upsilon}^{(t-1)}) - \nabla H(\mathbf{\Upsilon}^*|\mathbf{\Upsilon}^{(t-1)})\|_\infty = O_p(\sqrt{(\log p+\log q)/n}).$$

This completes the proof. □

S4 Additional Simulation Results

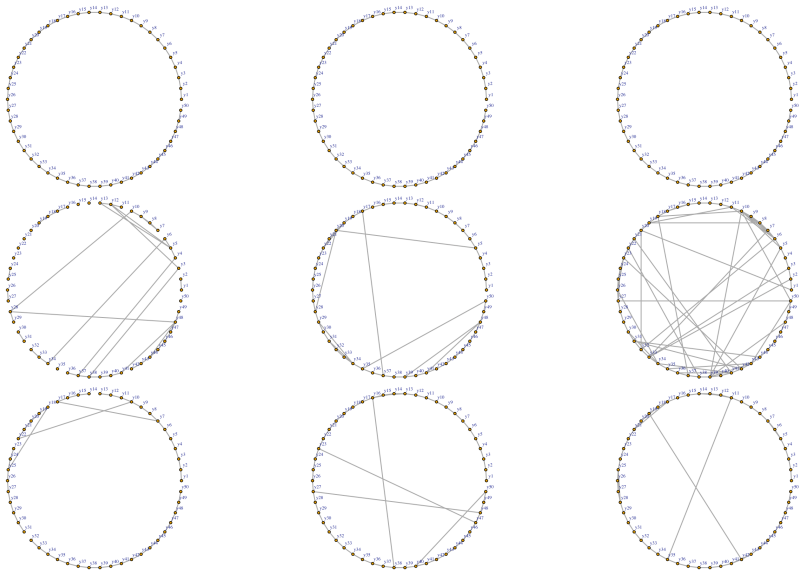

Figure S2: Simulation: true and estimated networks for three sample groups under S1. Top: true networks; Middle: estimation for one replicate with sizes (200, 200, 200); Bottom: estimation for one replicate with sizes (500, 500, 500).

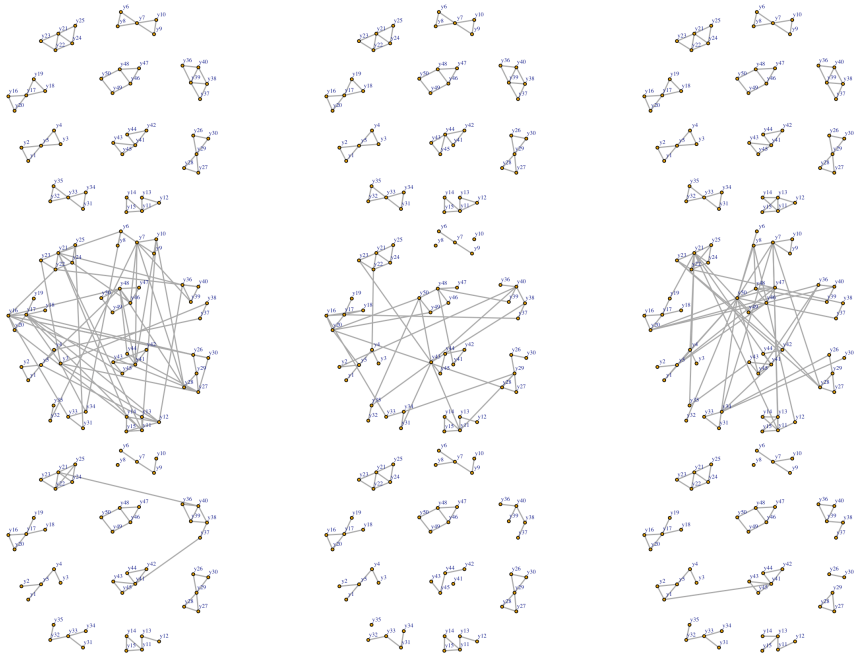

Figure S3: Simulation: true and estimated networks for three sample groups under S2. Top: true networks; Middle: estimation for one replicate with sizes (200, 200, 200); Bottom: estimation for one replicate with sizes (500, 500, 500).

Table S3: Simulation results under S1 with $p = q = 100$. In each cell, mean (sd).

n	Method		RMSE	TPR	FRP	RI	$\hat{K}_0$
(200,200,200)	Proposed	$\boldsymbol{\Theta}$	5.867(5.208)	0.952(0.054)	0.184(0.084)		
		$\boldsymbol{\Gamma}$	7.297(2.298)	0.792(0.091)	0.005(0.009)	0.644(0.291)	2.70(1.22)
	HeteroGGM	$\boldsymbol{\Theta}$	6.194(0.490)	0.980(0.009)	0.888(0.014)		
		$\boldsymbol{\Gamma}$	-	-	-	0.128(0.203)	4.75(1.77)
	CGLasso($K = 6$)	$\boldsymbol{\Theta}$	5.344(0.156)	0.828(0.012)	0.279(0.004)		
		$\boldsymbol{\Gamma}$	11.701(0.335)	0.770(0.061)	0.088(0.004)	0.022(0.012)	6(0)
	CGLasso($K = 4$)	$\boldsymbol{\Theta}$	5.683(0.245)	0.899(0.016)	0.249(0.017)		
		$\boldsymbol{\Gamma}$	10.469(1.419)	0.904(0.065)	0.069(0.008)	0.090(0.119)	4(0)
	CGLasso($K = 3$)	$\boldsymbol{\Theta}$	5.484(1.165)	0.937(0.012)	0.176(0.063)		
		$\boldsymbol{\Gamma}$	8.075(3.449)	0.969(0.034)	0.044(0.018)	0.350(0.350)	3(0)
	MRCE($K = 6$)	$\boldsymbol{\Theta}$	5.938(0.176)	0.846(0.013)	0.360(0.015)		
		$\boldsymbol{\Gamma}$	12.537(0.099)	0.284(0.048)	0.026(0.003)	0.022(0.012)	6(0)
	MRCE($K = 4$)	$\boldsymbol{\Theta}$	5.872(0.344)	0.910(0.021)	0.330(0.035)		
		$\boldsymbol{\Gamma}$	11.452(1.532)	0.431(0.192)	0.034(0.021)	0.090(0.119)	4(0)
	MRCE($K = 3$)	$\boldsymbol{\Theta}$	5.415(1.436)	0.955(0.026)	0.252(0.059)		
		$\boldsymbol{\Gamma}$	8.535(4.008)	0.782(0.242)	0.061(0.031)	0.350(0.350)	3(0)
	MCGGM($K = 3$)	$\boldsymbol{\Theta}$	6.671(1.435)	0.750(0.109)	0.086(0.085)		
		$\boldsymbol{\Gamma}$	-	-	-	0.666(0.173)	3(0)
(150,200,250)	Proposed	$\boldsymbol{\Theta}$	4.488(0.621)	0.963(0.026)	0.162(0.039)		
		$\boldsymbol{\Gamma}$	6.816(1.364)	0.781(0.111)	0.002(0.001)	0.705(0.159)	2.20(0.41)
	HeteroGGM	$\boldsymbol{\Theta}$	6.063(0.370)	0.983(0.009)	0.895(0.016)		
		$\boldsymbol{\Gamma}$	-	-	-	0.049(0.122)	5.30(1.49)
	CGLasso($K = 6$)	$\boldsymbol{\Theta}$	5.727(0.156)	0.859(0.012)	0.278(0.005)		
		$\boldsymbol{\Gamma}$	11.542(0.407)	0.806(0.056)	0.087(0.003)	0.030(0.014)	6(0)
	CGLasso($K = 4$)	$\boldsymbol{\Theta}$	6.111(0.348)	0.921(0.016)	0.253(0.018)		
		$\boldsymbol{\Gamma}$	10.980(1.108)	0.878(0.058)	0.071(0.007)	0.049(0.100)	4(0)
	CGLasso($K = 3$)	$\boldsymbol{\Theta}$	5.745(0.881)	0.956(0.014)	0.195(0.047)		
		$\boldsymbol{\Gamma}$	8.961(2.687)	0.951(0.055)	0.047(0.013)	0.239(0.265)	3(0)
	MRCE($K = 6$)	$\boldsymbol{\Theta}$	5.980(0.106)	0.874(0.013)	0.359(0.015)		
		$\boldsymbol{\Gamma}$	12.694(0.153)	0.235(0.070)	0.025(0.002)	0.030(0.014)	6(0)
	MRCE($K = 4$)	$\boldsymbol{\Theta}$	6.065(0.328)	0.928(0.016)	0.333(0.050)		
		$\boldsymbol{\Gamma}$	11.885(1.281)	0.357(0.208)	0.033(0.020)	0.049(0.100)	4(0)
	MRCE($K = 3$)	$\boldsymbol{\Theta}$	5.680(1.164)	0.961(0.018)	0.239(0.029)		
		$\boldsymbol{\Gamma}$	9.285(3.124)	0.785(0.216)	0.068(0.028)	0.239(0.265)	3(0)
	MCGGM($K = 3$)	$\boldsymbol{\Theta}$	6.544(0.569)	0.611(0.203)	0.035(0.033)		
		$\boldsymbol{\Gamma}$	-	-	-	0.700(0.142)	3(0)
(500,500,500)	Proposed	$\boldsymbol{\Theta}$	1.165(0.037)	0.998(0.002)	0.036(0.001)		
		$\boldsymbol{\Gamma}$	0.456(0.022)	1.000(0.000)	0.001(0.000)	1.000(0.000)	3.00(0.00)
	HeteroGGM	$\boldsymbol{\Theta}$	5.793(0.150)	0.993(0.004)	0.851(0.010)		
		$\boldsymbol{\Gamma}$	-	-	-	0.730(0.112)	5.35(0.99)
	CGLasso($K = 6$)	$\boldsymbol{\Theta}$	3.954(0.338)	0.957(0.024)	0.124(0.035)		
		$\boldsymbol{\Gamma}$	4.322(0.688)	0.999(0.002)	0.042(0.018)	0.613(0.036)	6(0)
	CGLasso($K = 4$)	$\boldsymbol{\Theta}$	3.134(0.300)	0.960(0.016)	0.050(0.021)		
		$\boldsymbol{\Gamma}$	2.560(0.483)	0.999(0.002)	0.022(0.007)	0.829(0.015)	4(0)
	CGLasso($K = 3$)	$\boldsymbol{\Theta}$	2.646(0.180)	0.978(0.020)	0.020(0.007)		
		$\boldsymbol{\Gamma}$	1.854(0.036)	1.000(0.000)	0.011(0.001)	0.964(0.007)	3(0)
	MRCE($K = 6$)	$\boldsymbol{\Theta}$	3.627(0.311)	0.987(0.007)	0.172(0.041)		
		$\boldsymbol{\Gamma}$	4.289(0.828)	0.979(0.047)	0.080(0.011)	0.613(0.036)	6(0)
	MRCE($K = 4$)	$\boldsymbol{\Theta}$	2.680(0.221)	0.995(0.007)	0.090(0.035)		
		$\boldsymbol{\Gamma}$	2.209(0.463)	0.998(0.008)	0.057(0.007)	0.829(0.015)	4(0)
	MRCE($K = 3$)	$\boldsymbol{\Theta}$	2.216(0.104)	0.999(0.002)	0.050(0.013)		
		$\boldsymbol{\Gamma}$	1.449(0.039)	1.000(0.000)	0.050(0.006)	0.964(0.007)	3(0)
	MCGGM($K = 3$)	$\boldsymbol{\Theta}$	6.234(0.680)	0.750(0.249)	0.036(0.038)		
		$\boldsymbol{\Gamma}$	-	-	-	0.729(0.185)	3(0)

Table S4: Simulation results under S2 with $p = q = 50$. In each cell, mean (sd).

n	Method		RMSE	TPR	FRP	RI	$\hat{K}_0$
(200,200,200)	Proposed	$\mathbf{\Theta}$	2.792(0.697)	0.874(0.046)	0.115(0.044)	0.785(0.220)	2.50(0.51)
		$\mathbf{\Gamma}$	4.066(1.536)	0.878(0.066)	0.005(0.004)		
	HeteroGGM	$\mathbf{\Theta}$	4.809(0.093)	0.975(0.008)	0.921(0.011)	0.107(0.172)	5.35(1.31)
		$\mathbf{\Gamma}$	-	-	-		
	CGLasso($K = 6$)	$\mathbf{\Theta}$	4.334(0.044)	0.701(0.016)	0.319(0.008)	0.023(0.014)	6(0)
		$\mathbf{\Gamma}$	8.715(0.290)	0.853(0.037)	0.141(0.003)		
	CGLasso($K = 4$)	$\mathbf{\Theta}$	4.583(0.037)	0.741(0.018)	0.290(0.007)	0.032(0.016)	4(0)
		$\mathbf{\Gamma}$	8.211(0.229)	0.927(0.020)	0.116(0.003)		
	CGLasso($K = 3$)	$\mathbf{\Theta}$	4.711(0.038)	0.762(0.019)	0.263(0.009)	0.029(0.023)	3(0)
		$\mathbf{\Gamma}$	8.166(0.428)	0.941(0.032)	0.097(0.005)		
	MRCE($K = 6$)	$\mathbf{\Theta}$	4.667(0.064)	0.736(0.020)	0.396(0.025)	0.023(0.014)	6(0)
		$\mathbf{\Gamma}$	9.489(0.196)	0.454(0.069)	0.058(0.010)		
	MRCE($K = 4$)	$\mathbf{\Theta}$	4.741(0.103)	0.777(0.025)	0.375(0.053)	0.032(0.016)	4(0)
		$\mathbf{\Gamma}$	8.921(0.346)	0.580(0.120)	0.071(0.024)		
	MRCE($K = 3$)	$\mathbf{\Theta}$	4.823(0.118)	0.784(0.024)	0.322(0.052)	0.029(0.023)	3(0)
		$\mathbf{\Gamma}$	8.804(0.619)	0.617(0.167)	0.070(0.028)		
	MCGGM($K = 3$)	$\mathbf{\Theta}$	4.459(0.578)	0.703(0.193)	0.143(0.112)	0.473(0.141)	3(0)
		$\mathbf{\Gamma}$	-	-	-		
(150,200,250)	Proposed	$\mathbf{\Theta}$	2.558(0.338)	0.879(0.028)	0.113(0.053)	0.934(0.136)	2.80(0.41)
		$\mathbf{\Gamma}$	3.482(0.751)	0.844(0.079)	0.003(0.003)		
	HeteroGGM	$\mathbf{\Theta}$	4.820(0.071)	0.971(0.007)	0.919(0.011)	0.144(0.243)	5.20(1.54)
		$\mathbf{\Gamma}$	-	-	-		
	CGLasso($K = 6$)	$\mathbf{\Theta}$	4.301(0.045)	0.702(0.022)	0.314(0.007)	0.029(0.012)	6(0)
		$\mathbf{\Gamma}$	8.350(0.235)	0.879(0.029)	0.137(0.004)		
	CGLasso($K = 4$)	$\mathbf{\Theta}$	4.543(0.069)	0.740(0.020)	0.286(0.010)	0.040(0.029)	4(0)
		$\mathbf{\Gamma}$	7.824(0.411)	0.937(0.025)	0.109(0.006)		
	CGLasso($K = 3$)	$\mathbf{\Theta}$	4.663(0.101)	0.765(0.028)	0.257(0.013)	0.047(0.063)	3(0)
		$\mathbf{\Gamma}$	7.974(0.796)	0.914(0.042)	0.092(0.007)		
	MRCE($K = 6$)	$\mathbf{\Theta}$	4.723(0.591)	0.708(0.142)	0.381(0.081)	0.029(0.012)	6(0)
		$\mathbf{\Gamma}$	9.107(0.228)	0.516(0.117)	0.067(0.019)		
	MRCE($K = 4$)	$\mathbf{\Theta}$	4.654(0.132)	0.771(0.038)	0.363(0.057)	0.040(0.029)	4(0)
		$\mathbf{\Gamma}$	8.436(0.515)	0.655(0.093)	0.082(0.023)		
	MRCE($K = 3$)	$\mathbf{\Theta}$	4.693(0.176)	0.792(0.030)	0.333(0.058)	0.047(0.063)	3(0)
		$\mathbf{\Gamma}$	8.352(0.898)	0.666(0.132)	0.088(0.023)		
	MCGGM($K = 3$)	$\mathbf{\Theta}$	4.612(0.555)	0.604(0.199)	0.098(0.109)	0.473(0.086)	3(0)
		$\mathbf{\Gamma}$	-	-	-		
(500,500,500)	Proposed	$\mathbf{\Theta}$	0.930(0.040)	0.980(0.006)	0.038(0.003)	1.000(0.000)	3.00(0.00)
		$\mathbf{\Gamma}$	0.369(0.062)	1.000(0.001)	0.001(0.001)		
	HeteroGGM	$\mathbf{\Theta}$	4.660(0.021)	0.980(0.004)	0.895(0.009)	0.642(0.047)	5.90(0.45)
		$\mathbf{\Gamma}$	-	-	-		
	CGLasso($K = 6$)	$\mathbf{\Theta}$	3.894(0.184)	0.800(0.018)	0.143(0.018)	0.298(0.049)	6(0)
		$\mathbf{\Gamma}$	4.972(0.603)	0.993(0.008)	0.057(0.019)		
	CGLasso($K = 4$)	$\mathbf{\Theta}$	3.679(0.303)	0.822(0.022)	0.090(0.022)	0.476(0.120)	4(0)
		$\mathbf{\Gamma}$	3.890(0.901)	0.999(0.002)	0.054(0.014)		
	CGLasso($K = 3$)	$\mathbf{\Theta}$	3.954(0.682)	0.827(0.033)	0.089(0.051)	0.400(0.303)	3(0)
		$\mathbf{\Gamma}$	4.610(2.453)	0.999(0.002)	0.052(0.026)		
	MRCE($K = 6$)	$\mathbf{\Theta}$	3.586(0.257)	0.872(0.036)	0.283(0.037)	0.298(0.049)	6(0)
		$\mathbf{\Gamma}$	4.854(0.698)	0.960(0.052)	0.126(0.013)		
	MRCE($K = 4$)	$\mathbf{\Theta}$	3.387(0.401)	0.907(0.032)	0.235(0.031)	0.476(0.120)	4(0)
		$\mathbf{\Gamma}$	3.910(1.060)	0.985(0.034)	0.081(0.006)		
	MRCE($K = 3$)	$\mathbf{\Theta}$	3.766(0.905)	0.899(0.067)	0.215(0.044)	0.400(0.303)	3(0)
		$\mathbf{\Gamma}$	4.825(2.778)	0.939(0.108)	0.057(0.004)		
	MCGGM($K = 3$)	$\mathbf{\Theta}$	4.493(0.544)	0.665(0.236)	0.064(0.052)	0.493(0.188)	3(0)
		$\mathbf{\Gamma}$	-	-	-		

Table S5: Simulation results under S2 with $p = q = 100$. In each cell, mean (sd).

n	Method		RMSE	TPR	FRP	RI	$\hat{K}_0$
(200,200,200)	Proposed	$\boldsymbol{\Theta}$	4.313(0.707)	0.869(0.021)	0.092(0.028)		
		$\boldsymbol{\Gamma}$	6.810(1.586)	0.846(0.059)	0.003(0.002)	0.678(0.191)	2.25(0.44)
	HeteroGGM	$\boldsymbol{\Theta}$	6.661(0.181)	0.966(0.007)	0.893(0.012)		
		$\boldsymbol{\Gamma}$	-	-	-	0.019(0.013)	5.15(1.53)
	CGLasso($K = 6$)	$\boldsymbol{\Theta}$	5.813(0.045)	0.723(0.014)	0.277(0.005)		
		$\boldsymbol{\Gamma}$	11.951(0.317)	0.823(0.044)	0.089(0.002)	0.025(0.012)	6(0)
	CGLasso($K = 4$)	$\boldsymbol{\Theta}$	6.210(0.038)	0.771(0.011)	0.254(0.004)		
		$\boldsymbol{\Gamma}$	11.419(0.416)	0.901(0.032)	0.073(0.002)	0.030(0.015)	4(0)
	CGLasso($K = 3$)	$\boldsymbol{\Theta}$	6.285(0.287)	0.799(0.011)	0.218(0.020)		
		$\boldsymbol{\Gamma}$	10.535(1.519)	0.941(0.037)	0.056(0.008)	0.095(0.127)	3(0)
	MRCE($K = 6$)	$\boldsymbol{\Theta}$	6.613(0.076)	0.756(0.018)	0.363(0.013)		
		$\boldsymbol{\Gamma}$	13.084(0.098)	0.333(0.046)	0.025(0.002)	0.025(0.012)	6(0)
	MRCE($K = 4$)	$\boldsymbol{\Theta}$	6.902(0.836)	0.747(0.157)	0.321(0.071)		
		$\boldsymbol{\Gamma}$	12.713(0.296)	0.357(0.115)	0.023(0.009)	0.030(0.015)	4(0)
	MRCE($K = 3$)	$\boldsymbol{\Theta}$	6.458(0.518)	0.815(0.024)	0.289(0.056)		
		$\boldsymbol{\Gamma}$	11.535(1.958)	0.583(0.254)	0.041(0.030)	0.095(0.127)	3(0)
	MCGGM($K = 3$)	$\boldsymbol{\Theta}$	6.746(0.510)	0.539(0.199)	0.023(0.027)		
		$\boldsymbol{\Gamma}$	-	-	-	0.729(0.362)	3(0)
(150,200,250)	Proposed	$\boldsymbol{\Theta}$	4.234(0.401)	0.865(0.020)	0.090(0.019)		
		$\boldsymbol{\Gamma}$	5.675(0.855)	0.842(0.091)	0.002(0.001)	0.723(0.168)	2.25(0.44)
	HeteroGGM	$\boldsymbol{\Theta}$	6.539(0.094)	0.965(0.006)	0.898(0.006)		
		$\boldsymbol{\Gamma}$	-	-	-	0.028(0.013)	5.70(0.66)
	CGLasso($K = 6$)	$\boldsymbol{\Theta}$	5.781(0.038)	0.722(0.012)	0.274(0.003)		
		$\boldsymbol{\Gamma}$	11.656(0.250)	0.833(0.026)	0.088(0.002)	0.023(0.011)	6(0)
	CGLasso($K = 4$)	$\boldsymbol{\Theta}$	6.154(0.122)	0.772(0.009)	0.250(0.008)		
		$\boldsymbol{\Gamma}$	10.815(0.722)	0.919(0.027)	0.071(0.004)	0.036(0.044)	4(0)
	CGLasso($K = 3$)	$\boldsymbol{\Theta}$	6.058(0.566)	0.805(0.021)	0.204(0.033)		
		$\boldsymbol{\Gamma}$	10.017(2.068)	0.903(0.056)	0.051(0.011)	0.161(0.210)	3(0)
	MRCE($K = 6$)	$\boldsymbol{\Theta}$	6.800(0.963)	0.717(0.169)	0.337(0.081)		
		$\boldsymbol{\Gamma}$	12.824(0.125)	0.366(0.095)	0.024(0.006)	0.023(0.011)	6(0)
	MRCE($K = 4$)	$\boldsymbol{\Theta}$	6.640(0.508)	0.772(0.085)	0.311(0.049)		
		$\boldsymbol{\Gamma}$	12.111(0.854)	0.495(0.120)	0.030(0.016)	0.036(0.044)	4(0)
	MRCE($K = 3$)	$\boldsymbol{\Theta}$	6.085(0.823)	0.825(0.036)	0.269(0.037)		
		$\boldsymbol{\Gamma}$	10.503(2.385)	0.676(0.204)	0.057(0.028)	0.161(0.210)	3(0)
	MCGGM($K = 3$)	$\boldsymbol{\Theta}$	6.528(0.628)	0.700(0.144)	0.088(0.075)		
		$\boldsymbol{\Gamma}$	-	-	-	0.633(0.132)	3(0)
(500,500,500)	Proposed	$\boldsymbol{\Theta}$	1.432(0.048)	0.958(0.011)	0.024(0.011)		
		$\boldsymbol{\Gamma}$	0.575(0.178)	0.999(0.002)	0.001(0.001)	1.000(0.000)	3.00(0.00)
	HeteroGGM	$\boldsymbol{\Theta}$	6.393(0.046)	0.981(0.006)	0.859(0.009)		
		$\boldsymbol{\Gamma}$	-	-	-	0.707(0.120)	5.25(1.02)
	CGLasso($K = 6$)	$\boldsymbol{\Theta}$	4.766(0.654)	0.836(0.012)	0.117(0.029)		
		$\boldsymbol{\Gamma}$	6.190(1.667)	0.992(0.012)	0.045(0.018)	0.440(0.152)	6(0)
	CGLasso($K = 4$)	$\boldsymbol{\Theta}$	3.541(0.279)	0.892(0.026)	0.055(0.021)		
		$\boldsymbol{\Gamma}$	3.156(0.556)	0.999(0.009)	0.023(0.008)	0.794(0.018)	4(0)
	CGLasso($K = 3$)	$\boldsymbol{\Theta}$	2.799(0.130)	0.940(0.013)	0.025(0.005)		
		$\boldsymbol{\Gamma}$	1.916(0.042)	1.000(0.000)	0.011(0.001)	0.956(0.007)	3(0)
	MRCE($K = 6$)	$\boldsymbol{\Theta}$	4.424(0.766)	0.884(0.025)	0.222(0.031)		
		$\boldsymbol{\Gamma}$	6.336(1.803)	0.932(0.082)	0.070(0.011)	0.440(0.152)	6(0)
	MRCE($K = 4$)	$\boldsymbol{\Theta}$	2.908(0.169)	0.952(0.014)	0.162(0.019)		
		$\boldsymbol{\Gamma}$	2.850(0.542)	0.999(0.004)	0.058(0.008)	0.794(0.018)	4(0)
	MRCE($K = 3$)	$\boldsymbol{\Theta}$	2.242(0.102)	0.971(0.008)	0.089(0.015)		
		$\boldsymbol{\Gamma}$	1.520(0.043)	1.000(0.000)	0.053(0.006)	0.956(0.007)	3(0)
	MCGGM($K = 3$)	$\boldsymbol{\Theta}$	6.303(0.641)	0.676(0.230)	0.042(0.039)		
		$\boldsymbol{\Gamma}$	-	-	-	0.787(0.125)	3(0)

Table S6: Simulation results under S3 with $p = q = 50$. In each cell, mean (sd).

n	Method		RMSE	TPR	FRP	RI	$\hat{K}_0$
(200,200,200)	Proposed	$\mathbf{\Theta}$	1.728(1.083)	0.938(0.034)	0.048(0.021)	0.994(0.029)	3.05(0.22)
		$\mathbf{\Gamma}$	1.241(1.336)	0.983(0.027)	0.009(0.017)		
	HeteroGGM	$\mathbf{\Theta}$	4.131(0.171)	0.983(0.010)	0.900(0.012)	0.473(0.244)	5.15(1.38)
		$\mathbf{\Gamma}$	-	-	-		
	CGLasso($K = 6$)	$\mathbf{\Theta}$	3.496(0.181)	0.869(0.018)	0.249(0.018)	0.387(0.089)	6(0)
		$\mathbf{\Gamma}$	5.589(0.541)	0.964(0.034)	0.102(0.006)		
	CGLasso($K = 4$)	$\mathbf{\Theta}$	3.284(0.292)	0.908(0.018)	0.184(0.025)	0.572(0.076)	4(0)
		$\mathbf{\Gamma}$	4.013(0.676)	0.992(0.007)	0.064(0.009)		
	CGLasso($K = 3$)	$\mathbf{\Theta}$	2.982(0.429)	0.919(0.021)	0.112(0.038)	0.738(0.145)	3(0)
		$\mathbf{\Gamma}$	2.853(1.138)	0.998(0.004)	0.037(0.012)		
	MRCE($K = 6$)	$\mathbf{\Theta}$	3.929(0.640)	0.867(0.149)	0.353(0.063)	0.387(0.089)	6(0)
		$\mathbf{\Gamma}$	6.463(0.697)	0.784(0.153)	0.116(0.030)		
	MRCE($K = 4$)	$\mathbf{\Theta}$	3.035(0.320)	0.959(0.013)	0.293(0.044)	0.572(0.076)	4(0)
		$\mathbf{\Gamma}$	4.066(0.717)	0.977(0.028)	0.148(0.022)		
	MRCE($K = 3$)	$\mathbf{\Theta}$	2.625(0.519)	0.979(0.013)	0.238(0.033)	0.738(0.145)	3(0)
		$\mathbf{\Gamma}$	2.640(1.222)	0.997(0.005)	0.120(0.013)		
	MCGGM($K = 3$)	$\mathbf{\Theta}$	4.654(1.275)	0.742(0.250)	0.141(0.115)	0.528(0.219)	3(0)
		$\mathbf{\Gamma}$	-	-	-		
(150,200,250)	Proposed	$\mathbf{\Theta}$	1.399(0.054)	0.949(0.018)	0.053(0.004)	1.000(0.000)	3.00(0.00)
		$\mathbf{\Gamma}$	0.636(0.173)	0.998(0.005)	0.005(0.002)		
	HeteroGGM	$\mathbf{\Theta}$	4.142(0.143)	0.988(0.007)	0.905(0.013)	0.467(0.270)	4.55(1.64)
		$\mathbf{\Gamma}$	-	-	-		
	CGLasso($K = 6$)	$\mathbf{\Theta}$	3.432(0.147)	0.899(0.020)	0.237(0.016)	0.337(0.067)	6(0)
		$\mathbf{\Gamma}$	5.204(0.500)	0.972(0.024)	0.094(0.007)		
	CGLasso($K = 4$)	$\mathbf{\Theta}$	3.196(0.233)	0.922(0.018)	0.173(0.026)	0.541(0.089)	4(0)
		$\mathbf{\Gamma}$	3.741(0.784)	0.995(0.008)	0.058(0.009)		
	CGLasso($K = 3$)	$\mathbf{\Theta}$	3.164(0.542)	0.927(0.028)	0.134(0.050)	0.638(0.226)	3(0)
		$\mathbf{\Gamma}$	3.440(1.657)	0.996(0.008)	0.044(0.016)		
	MRCE($K = 6$)	$\mathbf{\Theta}$	3.814(0.686)	0.927(0.019)	0.342(0.035)	0.337(0.067)	6(0)
		$\mathbf{\Gamma}$	5.767(0.563)	0.884(0.054)	0.146(0.029)		
	MRCE($K = 4$)	$\mathbf{\Theta}$	2.921(0.284)	0.969(0.011)	0.287(0.029)	0.541(0.089)	4(0)
		$\mathbf{\Gamma}$	3.737(0.860)	0.977(0.049)	0.150(0.015)		
	MRCE($K = 3$)	$\mathbf{\Theta}$	2.807(0.653)	0.970(0.013)	0.277(0.035)	0.638(0.226)	3(0)
		$\mathbf{\Gamma}$	3.278(1.770)	0.996(0.009)	0.133(0.017)		
	MCGGM($K = 3$)	$\mathbf{\Theta}$	4.515(0.381)	0.714(0.150)	0.064(0.061)	0.589(0.112)	3(0)
		$\mathbf{\Gamma}$	-	-	-		
(500,500,500)	Proposed	$\mathbf{\Theta}$	0.755(0.035)	0.997(0.005)	0.032(0.007)	1.000(0.000)	3.00(0.00)
		$\mathbf{\Gamma}$	0.324(0.021)	1.000(0.000)	0.001(0.000)		
	HeteroGGM	$\mathbf{\Theta}$	4.055(0.096)	0.990(0.005)	0.881(0.008)	0.669(0.085)	5.80(0.70)
		$\mathbf{\Gamma}$	-	-	-		
	CGLasso($K = 6$)	$\mathbf{\Theta}$	2.802(0.175)	0.922(0.019)	0.080(0.021)	0.517(0.041)	6(0)
		$\mathbf{\Gamma}$	2.951(0.404)	0.999(0.003)	0.043(0.016)		
	CGLasso($K = 4$)	$\mathbf{\Theta}$	2.581(0.173)	0.958(0.024)	0.065(0.026)	0.751(0.014)	4(0)
		$\mathbf{\Gamma}$	2.243(0.236)	1.000(0.000)	0.037(0.018)		
	CGLasso($K = 3$)	$\mathbf{\Theta}$	2.315(0.153)	0.970(0.026)	0.039(0.015)	0.877(0.015)	3(0)
		$\mathbf{\Gamma}$	1.555(0.052)	1.000(0.000)	0.022(0.001)		
	MRCE($K = 6$)	$\mathbf{\Theta}$	2.331(0.162)	0.982(0.009)	0.226(0.031)	0.517(0.041)	6(0)
		$\mathbf{\Gamma}$	2.726(0.382)	0.991(0.029)	0.113(0.012)		
	MRCE($K = 4$)	$\mathbf{\Theta}$	2.080(0.083)	0.997(0.004)	0.201(0.025)	0.751(0.014)	4(0)
		$\mathbf{\Gamma}$	1.950(0.117)	1.000(0.000)	0.093(0.012)		
	MRCE($K = 3$)	$\mathbf{\Theta}$	1.884(0.099)	0.996(0.007)	0.136(0.009)	0.877(0.015)	3(0)
		$\mathbf{\Gamma}$	1.299(0.067)	1.000(0.000)	0.060(0.003)		
	MCGGM($K = 3$)	$\mathbf{\Theta}$	4.220(0.432)	0.825(0.206)	0.070(0.046)	0.487(0.180)	3(0)
		$\mathbf{\Gamma}$	-	-	-		

Table S7: Simulation results under S3 with $p = q = 100$. In each cell, mean (sd).

n	Method		RMSE	TPR	FRP	RI	$\hat{K}_0$
(200,200,200)	Proposed	$\boldsymbol{\Theta}$	3.901(0.844)	0.952(0.024)	0.081(0.014)		
		$\boldsymbol{\Gamma}$	5.865(1.936)	0.885(0.070)	0.003(0.002)	0.721(0.210)	2.35(0.49)
	HeteroGGM	$\boldsymbol{\Theta}$	6.175(0.570)	0.979(0.007)	0.895(0.013)		
		$\boldsymbol{\Gamma}$	-	-	-	0.024(0.016)	5.30(1.34)
	CGLasso($K = 6$)	$\boldsymbol{\Theta}$	5.412(0.178)	0.827(0.016)	0.272(0.019)		
		$\boldsymbol{\Gamma}$	11.072(0.954)	0.826(0.090)	0.083(0.008)	0.126(0.159)	6(0)
	CGLasso($K = 4$)	$\boldsymbol{\Theta}$	4.636(0.984)	0.917(0.024)	0.176(0.066)		
		$\boldsymbol{\Gamma}$	6.655(3.210)	0.976(0.043)	0.049(0.019)	0.490(0.334)	4(0)
	CGLasso($K = 3$)	$\boldsymbol{\Theta}$	3.845(1.242)	0.926(0.018)	0.088(0.070)		
		$\boldsymbol{\Gamma}$	4.270(3.336)	0.983(0.044)	0.025(0.017)	0.775(0.355)	3(0)
	MRCE($K = 6$)	$\boldsymbol{\Theta}$	5.933(0.538)	0.838(0.099)	0.359(0.036)		
		$\boldsymbol{\Gamma}$	12.153(0.608)	0.267(0.068)	0.026(0.006)	0.126(0.159)	6(0)
	MRCE($K = 4$)	$\boldsymbol{\Theta}$	4.806(0.981)	0.947(0.034)	0.278(0.081)		
		$\boldsymbol{\Gamma}$	7.751(3.422)	0.716(0.285)	0.042(0.022)	0.490(0.334)	4(0)
	MRCE($K = 3$)	$\boldsymbol{\Theta}$	3.635(1.485)	0.962(0.026)	0.182(0.092)		
		$\boldsymbol{\Gamma}$	4.597(3.837)	0.903(0.196)	0.040(0.018)	0.775(0.355)	3(0)
	MCGGM($K = 3$)	$\boldsymbol{\Theta}$	6.464(0.863)	0.633(0.269)	0.067(0.094)		
		$\boldsymbol{\Gamma}$	-	-	-	0.869(0.128)	3(0)
(150,200,250)	Proposed	$\boldsymbol{\Theta}$	3.515(1.097)	0.948(0.017)	0.076(0.017)		
		$\boldsymbol{\Gamma}$	4.434(2.427)	0.905(0.115)	0.002(0.002)	0.809(0.203)	2.50(0.51)
	HeteroGGM	$\boldsymbol{\Theta}$	5.977(0.172)	0.985(0.007)	0.897(0.011)		
		$\boldsymbol{\Gamma}$	-	-	-	0.020(0.015)	5.40(1.27)
	CGLasso($K = 6$)	$\boldsymbol{\Theta}$	5.782(0.217)	0.856(0.012)	0.283(0.007)		
		$\boldsymbol{\Gamma}$	11.572(0.570)	0.822(0.064)	0.088(0.004)	0.027(0.024)	6(0)
	CGLasso($K = 4$)	$\boldsymbol{\Theta}$	4.759(1.327)	0.936(0.015)	0.173(0.072)		
		$\boldsymbol{\Gamma}$	6.592(3.503)	0.974(0.041)	0.047(0.021)	0.441(0.323)	4(0)
	CGLasso($K = 3$)	$\boldsymbol{\Theta}$	3.760(0.918)	0.935(0.024)	0.087(0.053)		
		$\boldsymbol{\Gamma}$	4.368(2.994)	0.959(0.100)	0.024(0.011)	0.790(0.276)	3(0)
	MRCE($K = 6$)	$\boldsymbol{\Theta}$	6.036(0.160)	0.875(0.010)	0.371(0.022)		
		$\boldsymbol{\Gamma}$	12.669(0.412)	0.217(0.108)	0.025(0.007)	0.027(0.024)	6(0)
	MRCE($K = 4$)	$\boldsymbol{\Theta}$	4.983(1.724)	0.910(0.215)	0.252(0.096)		
		$\boldsymbol{\Gamma}$	7.689(3.910)	0.675(0.364)	0.039(0.021)	0.441(0.323)	4(0)
	MRCE($K = 3$)	$\boldsymbol{\Theta}$	3.620(0.991)	0.972(0.011)	0.186(0.063)		
		$\boldsymbol{\Gamma}$	4.493(3.227)	0.926(0.152)	0.050(0.019)	0.790(0.276)	3(0)
	MCGGM($K = 3$)	$\boldsymbol{\Theta}$	6.876(0.609)	0.512(0.177)	0.023(0.033)		
		$\boldsymbol{\Gamma}$	-	-	-	0.787(0.087)	3(0)
(500,500,500)	Proposed	$\boldsymbol{\Theta}$	1.177(0.037)	0.995(0.003)	0.017(0.008)		
		$\boldsymbol{\Gamma}$	0.460(0.015)	1.000(0.000)	0.000(0.001)	1.000(0.000)	3.00(0.00)
	HeteroGGM	$\boldsymbol{\Theta}$	5.851(0.230)	0.991(0.006)	0.855(0.009)		
		$\boldsymbol{\Gamma}$	-	-	-	0.728(0.134)	5.35(1.18)
	CGLasso($K = 6$)	$\boldsymbol{\Theta}$	3.618(0.330)	0.944(0.029)	0.095(0.031)		
		$\boldsymbol{\Gamma}$	3.900(0.628)	0.999(0.003)	0.038(0.017)	0.651(0.038)	6(0)
	CGLasso($K = 4$)	$\boldsymbol{\Theta}$	2.962(0.390)	0.961(0.025)	0.049(0.019)		
		$\boldsymbol{\Gamma}$	2.546(0.623)	1.000(0.001)	0.021(0.005)	0.859(0.016)	4(0)
	CGLasso($K = 3$)	$\boldsymbol{\Theta}$	2.755(0.416)	0.964(0.026)	0.017(0.014)		
		$\boldsymbol{\Gamma}$	2.190(1.781)	0.984(0.074)	0.012(0.002)	0.956(0.125)	3(0)
	MRCE($K = 6$)	$\boldsymbol{\Theta}$	3.376(0.473)	0.974(0.039)	0.157(0.032)		
		$\boldsymbol{\Gamma}$	4.122(0.865)	0.957(0.070)	0.063(0.018)	0.651(0.038)	6(0)
	MRCE($K = 4$)	$\boldsymbol{\Theta}$	2.487(0.224)	0.993(0.017)	0.083(0.026)		
		$\boldsymbol{\Gamma}$	2.270(0.563)	0.993(0.024)	0.054(0.010)	0.859(0.016)	4(0)
	MRCE($K = 3$)	$\boldsymbol{\Theta}$	2.147(0.428)	0.998(0.002)	0.050(0.017)		
		$\boldsymbol{\Gamma}$	1.810(1.871)	0.984(0.070)	0.044(0.004)	0.956(0.125)	3(0)
	MCGGM($K = 3$)	$\boldsymbol{\Theta}$	5.771(0.457)	0.924(0.092)	0.068(0.034)		
		$\boldsymbol{\Gamma}$	-	-	-	0.716(0.132)	3(0)

Table S8: Simulation results under S1 with $p = q = 200$. In each cell, mean (sd).

n	Method		RMSE	TPR	FPR	RI	$\hat{K}_0$
$(200, 200, 200)$	Proposed	$\mathbf{\Theta}$	6.293(0.531)	0.960(0.032)	0.074(0.010)	0.567(0.007)	2.00(0.00)
		$\mathbf{\Gamma}$	9.672(1.588)	0.841(0.068)	0.002(0.001)		
	HeteroGGM	$\mathbf{\Theta}$	10.329(3.108)	0.971(0.009)	0.860(0.013)	0.027(0.011)	5.68(0.94)
		$\mathbf{\Gamma}$	-	-	-		
	CGLasso($K = 6$)	$\mathbf{\Theta}$	31.018(16.126)	0.515(0.201)	0.086(0.112)	0.013(0.006)	6(0)
		$\mathbf{\Gamma}$	20.934(2.608)	0.839(0.073)	0.274(0.118)		
	CGLasso($K = 4$)	$\mathbf{\Theta}$	6.819(0.304)	0.899(0.011)	0.262(0.002)	0.008(0.008)	4(0)
		$\mathbf{\Gamma}$	16.360(0.444)	0.883(0.036)	0.103(0.003)		
	CGLasso($K = 3$)	$\mathbf{\Theta}$	7.972(0.384)	0.945(0.007)	0.257(0.015)	0.052(0.135)	3(0)
		$\mathbf{\Gamma}$	15.384(2.028)	0.954(0.018)	0.085(0.009)		
	MRCE($K = 6$)	$\mathbf{\Theta}$	31.703(4.298)	0.591(0.029)	0.156(0.023)	0.013(0.006)	6(0)
		$\mathbf{\Gamma}$	20.647(0.348)	0.740(0.047)	0.235(0.019)		
	MRCE($K = 4$)	$\mathbf{\Theta}$	8.004(0.325)	0.934(0.003)	0.416(0.013)	0.008(0.008)	4(0)
		$\mathbf{\Gamma}$	16.809(0.509)	0.647(0.062)	0.054(0.004)		
	MRCE($K = 3$)	$\mathbf{\Theta}$	8.558(0.150)	0.969(0.005)	0.437(0.012)	0.053(0.135)	3(0)
		$\mathbf{\Gamma}$	16.688(0.552)	0.659(0.039)	0.033(0.003)		
	MCGGM($K = 3$)	$\mathbf{\Theta}$	16.254(8.581)	0.810(0.091)	0.105(0.023)	0.121(0.249)	3(0)
		$\mathbf{\Gamma}$	-	-	-		
$(500, 500, 500)$	Proposed	$\mathbf{\Theta}$	2.077(0.042)	0.980(0.002)	0.003(0.000)	1.000(0.000)	3.00(0.00)
		$\mathbf{\Gamma}$	1.011(0.144)	0.998(0.001)	0.001(0.000)		
	HeteroGGM	$\mathbf{\Theta}$	8.512(0.286)	0.993(0.003)	0.823(0.041)	0.269(0.234)	4.80(1.61)
		$\mathbf{\Gamma}$	-	-	-		
	CGLasso($K = 6$)	$\mathbf{\Theta}$	5.625(1.136)	0.979(0.008)	0.157(0.037)	0.536(0.206)	6(0)
		$\mathbf{\Gamma}$	7.479(3.227)	0.997(0.006)	0.043(0.012)		
	CGLasso($K = 4$)	$\mathbf{\Theta}$	4.017(0.892)	0.983(0.015)	0.068(0.023)	0.831(0.098)	4(0)
		$\mathbf{\Gamma}$	3.599(1.775)	1.000(0.000)	0.015(0.004)		
	CGLasso($K = 3$)	$\mathbf{\Theta}$	3.613(0.839)	0.984(0.004)	0.023(0.019)	0.963(0.124)	3(0)
		$\mathbf{\Gamma}$	2.927(1.848)	1.000(0.000)	0.007(0.003)		
	MRCE($K = 6$)	$\mathbf{\Theta}$	7.168(0.253)	0.991(0.002)	0.252(0.005)	0.536(0.206)	6(0)
		$\mathbf{\Gamma}$	5.678(0.148)	0.992(0.001)	0.073(0.002)		
	MRCE($K = 4$)	$\mathbf{\Theta}$	3.035(0.003)	0.999(0.001)	0.128(0.001)	0.831(0.098)	4(0)
		$\mathbf{\Gamma}$	2.535(0.048)	1.000(0.000)	0.049(0.001)		
	MRCE($K = 3$)	$\mathbf{\Theta}$	2.625(0.004)	0.999(0.001)	0.079(0.001)	0.963(0.124)	3(0)
		$\mathbf{\Gamma}$	1.989(0.058)	1.000(0.000)	0.018(0.001)		
	MCGGM($K = 3$)	$\mathbf{\Theta}$	7.769(3.645)	0.907(0.088)	0.049(0.014)	0.391(0.229)	3(0)
		$\mathbf{\Gamma}$	-	-	-		

Table S9: Simulation results under S2 with $p = q = 50, K_0 = 10$. In each cell, mean (sd).

n	Method		RMSE	TPR	FPR	RI	$\hat{K}_0$
$(200, 200, 200)$	Proposed	$\boldsymbol{\Theta}$	4.275(0.165)	0.736(0.017)	0.092(0.006)	0.743(0.085)	7.95(0.69)
		$\boldsymbol{\Gamma}$	7.917(0.425)	0.276(0.055)	0.001(0.001)		
	HeteroGGM	$\boldsymbol{\Theta}$	4.906(0.088)	0.962(0.006)	0.911(0.013)	0.199(0.196)	14.20(3.99)
		$\boldsymbol{\Gamma}$	-	-	-		
	CGLasso($K = 20$)	$\boldsymbol{\Theta}$	4.387(0.025)	0.661(0.008)	0.431(0.003)	0.031(0.006)	20(0)
		$\boldsymbol{\Gamma}$	9.938(0.093)	0.622(0.029)	0.242(0.003)		
	CGLasso($K = 10$)	$\boldsymbol{\Theta}$	4.977(0.025)	0.699(0.009)	0.396(0.004)	0.032(0.009)	10(0)
		$\boldsymbol{\Gamma}$	9.326(0.155)	0.686(0.050)	0.190(0.003)		
	CGLasso($K = 5$)	$\boldsymbol{\Theta}$	5.258(0.030)	0.734(0.012)	0.315(0.007)	0.038(0.015)	5(0)
		$\boldsymbol{\Gamma}$	9.092(0.261)	0.741(0.064)	0.123(0.003)		
	MRCE($K = 20$)	$\boldsymbol{\Theta}$	4.703(0.023)	0.740(0.011)	0.572(0.004)	0.031(0.006)	20(0)
		$\boldsymbol{\Gamma}$	9.693(0.038)	0.374(0.036)	0.191(0.003)		
	MRCE($K = 10$)	$\boldsymbol{\Theta}$	5.069(0.025)	0.787(0.010)	0.580(0.008)	0.032(0.009)	10(0)
		$\boldsymbol{\Gamma}$	9.789(0.047)	0.295(0.042)	0.056(0.002)		
	MRCE($K = 5$)	$\boldsymbol{\Theta}$	5.384(0.006)	0.806(0.021)	0.505(0.006)	0.038(0.015)	5(0)
		$\boldsymbol{\Gamma}$	8.936(0.011)	0.107(0.008)	0.029(0.001)		
	MCGGM($K = 10$)	$\boldsymbol{\Theta}$	10.428(5.145)	0.603(0.193)	0.069(0.014)	0.280(0.168)	10(0)
		$\boldsymbol{\Gamma}$	-	-	-		
$(500, 500, 500)$	Proposed	$\boldsymbol{\Theta}$	1.715(0.123)	0.948(0.004)	0.049(0.003)	0.999(0.001)	10.00(0.00)
		$\boldsymbol{\Gamma}$	2.962(0.282)	0.893(0.021)	0.076(0.004)		
	HeteroGGM	$\boldsymbol{\Theta}$	4.467(0.045)	0.978(0.005)	0.869(0.012)	0.770(0.073)	15.00(2.65)
		$\boldsymbol{\Gamma}$	-	-	-		
	CGLasso($K = 20$)	$\boldsymbol{\Theta}$	4.944(0.016)	0.690(0.011)	0.354(0.003)	0.040(0.003)	20(0)
		$\boldsymbol{\Gamma}$	9.404(0.121)	0.795(0.024)	0.153(0.002)		
	CGLasso($K = 10$)	$\boldsymbol{\Theta}$	5.259(0.014)	0.690(0.011)	0.279(0.004)	0.022(0.002)	10(0)
		$\boldsymbol{\Gamma}$	9.518(0.047)	0.740(0.024)	0.100(0.002)		
	CGLasso($K = 5$)	$\boldsymbol{\Theta}$	5.404(0.019)	0.702(0.018)	0.208(0.004)	0.009(0.002)	5(0)
		$\boldsymbol{\Gamma}$	9.578(0.126)	0.599(0.048)	0.073(0.002)		
	MRCE($K = 20$)	$\boldsymbol{\Theta}$	5.073(0.019)	0.722(0.012)	0.415(0.013)	0.040(0.002)	20(0)
		$\boldsymbol{\Gamma}$	9.873(0.097)	0.464(0.032)	0.045(0.003)		
	MRCE($K = 10$)	$\boldsymbol{\Theta}$	5.298(0.015)	0.754(0.023)	0.417(0.039)	0.022(0.002)	10(0)
		$\boldsymbol{\Gamma}$	9.935(0.059)	0.237(0.024)	0.025(0.001)		
	MRCE($K = 5$)	$\boldsymbol{\Theta}$	5.409(0.023)	0.778(0.039)	0.372(0.068)	0.009(0.002)	5(0)
		$\boldsymbol{\Gamma}$	9.822(0.113)	0.079(0.016)	0.022(0.001)		
	MCGGM($K = 10$)	$\boldsymbol{\Theta}$	5.141(0.035)	0.651(0.158)	0.020(0.008)	0.437(0.055)	10(0)
		$\boldsymbol{\Gamma}$	-	-	-		

Computational Cost We examine computer time of the different methods using setting S1 with $n = (200, 200, 200)$ and various (p, q) values. Computation is done on a regular laptop with a 2.3 GHz Quad-Core Intel Core i5 processor. The results based on 100 replicates are presented in Table S10. With a higher computational complexity, as expected, the proposed approach has a higher computational cost. However, it is still acceptable.

Additional Simulation for Theorem 1 We conduct simulation to "validate"

Table S10: Computer time (in seconds). In each cell, mean.

(p, q)	$(10, 10)$	$(20, 20)$	$(50, 50)$	$(100, 100)$
Proposed	4.420	7.429	38.502	532.505
HeteroGGM	1.008	9.253	76.959	102.286
CGLasso	0.222	0.411	1.339	6.337
MRCE	0.228	0.394	2.599	77.994
MCGGM	13.977	73.941	181.2049	226.804

Theorem 1. Under setting S1, we fix $p = q = 50$ and $K_0 = 3$ and vary the sample size from $n_k = 200$ to $n_k = 3000$ for $k = 1, \cdots, K_0$. Figure S4 shows the estimation error $\sum_{k=1}^{\hat{K}_0} \left(\|\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma}_k^*\|_F + \|\hat{\mathbf{\Theta}}_k - \mathbf{\Theta}_k^*\|_F \right)$ plotted on a log-log scale. It is observed that the estimation error decreases as the sample size increases. The red line, which is from a linear regression, has slope approximately -0.5, which suggests a convergence rate of $1/\sqrt{n}$. This can provide support to Theorem 1.

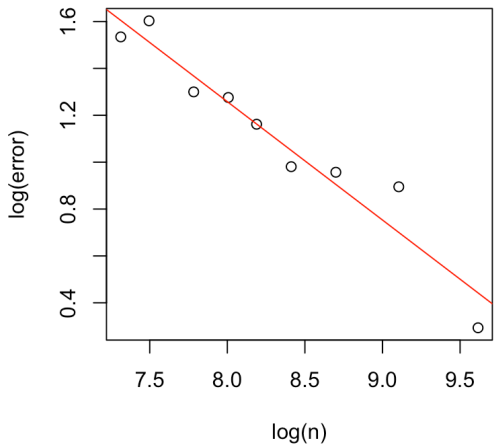

Figure S4: Estimation error as a function of sample size.

S5 Analysis of METABRIC data: additional results

Table S11: Analysis of METABRIC data: numbers of edges and overlapping edges for the six sample groups.

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
Group 1	684	134	118	140	68	110
Group 2		676	166	178	94	170
Group 3			432	136	72	148
Group 4				638	96	152
Group 5					380	94
Group 6						652

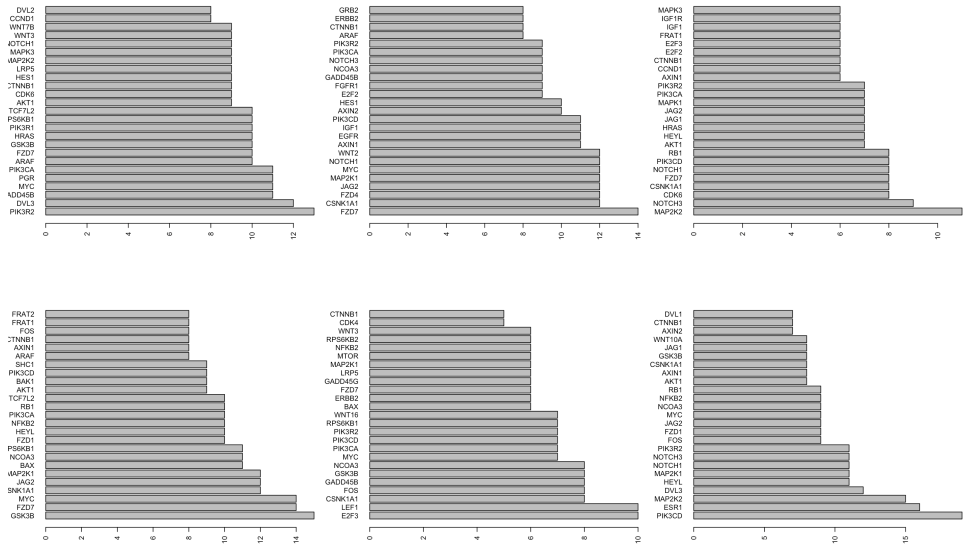

Figure S5: Analysis of METABRIC data: genes with the highest degrees for the six sample groups.

Table S12: Analysis of METABRIC data: analysis of variance for selected clinical variables.

Clinical variable	p-value
Tumor size	< 0.001
Mutation count	< 0.001
TMB nonsynonymous	< 0.001

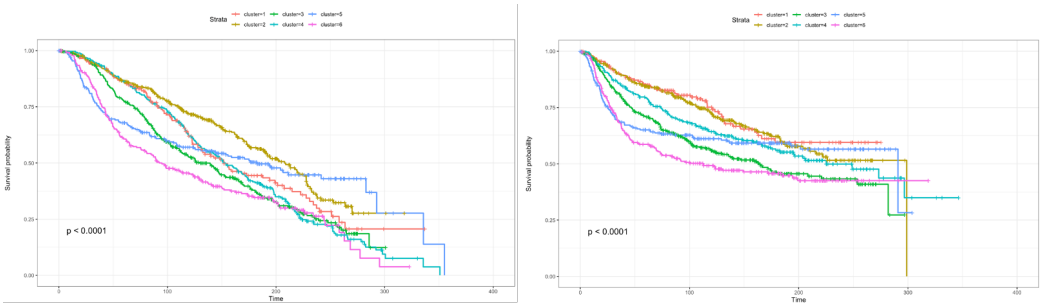

Figure S6: Analysis of METABRIC data: comparison of survival. Left: overall survival; Right: relapse free survival.

Table S13: Analysis of METABRIC data: comparison with Claudin subtypes.

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Sum
Basal	0	4	4	2	157	32	199
Claudin-low	2	80	4	0	79	34	199
Her2	2	11	32	20	9	146	220
LumA	114	250	80	219	2	14	679
LumB	83	42	236	62	1	37	461
Sum	201	387	356	303	248	263	1758

Table S14: Analysis of METABRIC data using different approaches. Diagonal: sample group sizes. Off-diagonal: Rand index.

	Proposed	CGLasso/MRCE	HeteroGGM	MCGGM
Proposed	201/387/356/303/248/263	0.777	0.834	0.423
CGLasso/MRCE		17/327/531/226/657	0.771	0.508
HeteroGGM			380/291/515/338/185/49	0.452
MCGGM				0/28/0/275/1455

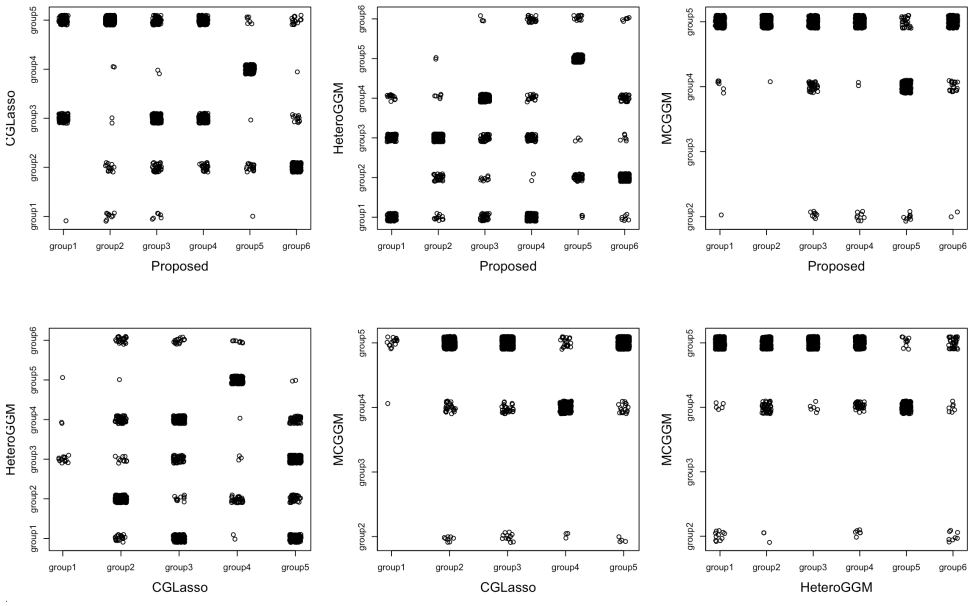

Figure S7: Analysis of METABRIC data using different approaches: comparison of grouping results.

S6 Analysis of TCGA data

To further demonstrate utility of the proposed approach, we conduct analysis of The Cancer Genome Atlas (TCGA) data. TCGA is a collective effort organized by the NIH and has comprehensively collected and published multiomics data on multiple cancer types. In particular, it is noted that gene expression-based heterogeneity analysis, including network-based, has been conducted using TCGA data (Cai and Li, 2017; Ni et al., 2018). In the TCGA breast cancer (BRCA) study, data is available for 1,048 patients. As in the main text, we consider the KEGG hsa05224 (breast cancer) pathway. Measurements are available for 147 gene expressions and their corresponding CNAs. We refer to published studies for information on data collection, management, and processing (Cancer Genome Atlas, 2012).

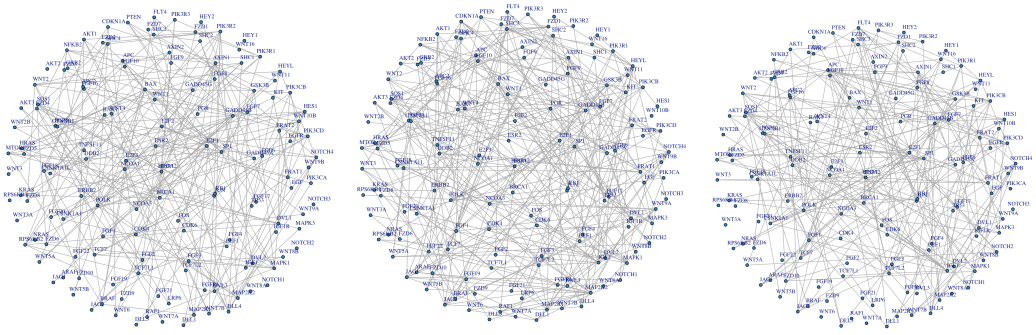

Figure S8: Analysis of TCGA BRCA data: network structures for the three sample groups.

Table S15: Analysis of TCGA BRCA data: numbers of edges and overlapping edges for the three sample groups.

	Group 1	Group 2	Group 3
Group 1	488	94	88
Group 2		596	158
Group 3			458

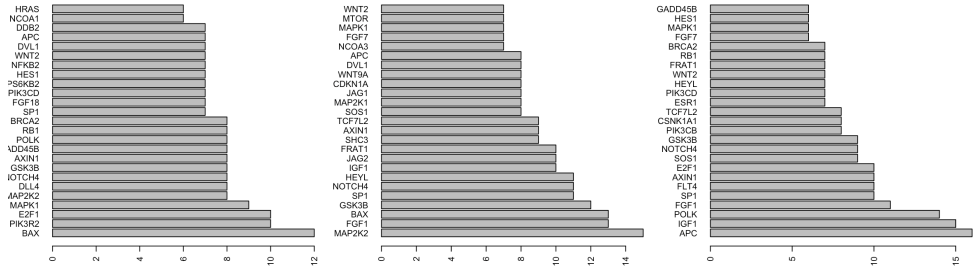

Figure S9: Analysis of TCGA BRCA data: genes with the highest degrees for the three sample groups.

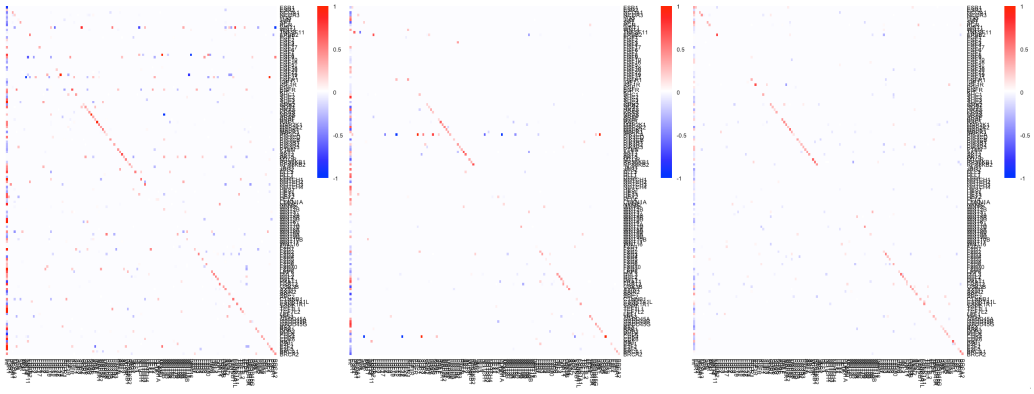

Figure S10: Analysis of TCGA BRCA data: heatmaps of the estimated coefficient matrices for the three sample groups.

The proposed method identifies three sample groups, with sizes 186, 251 and 611, respectively. Detailed information on sample grouping is available from the authors. The three gene expression network structures are shown in Figure S8. As shown in Table S15, they have 488, 596, and 458 edges, respectively. And they have limited overlapping edges, suggesting that the three networks are significantly different – this can also be seen from Figure S8. In Figure S9, we provide information on the genes with the highest degrees in the three networks. Significant differences are again observed. Many individual genes have significantly different connection properties in the networks. For example, gene ESR1 is an isolated node in the first network, has a few edges in the second network, and is a well-connected hub node in the third network. Genes IGF1, BAX and FGF1, which have been established as important biomarkers for breast cancer, also have different properties for the three sample groups. There are also some commonalities. For example, genes SP1, NOTCH4, WNT2, GSK3B, AXIN1, APC and POLK are key hubs for all the groups. In Figure S10, we show the heatmaps of the estimated coefficient matrices. Simply eyeballing the heatmaps suggests significant differences. In addition, we observe more and stronger cis regulations, which is as expected. However, a significant number of trans regulations are also observed.

Table S16: Analysis of TCGA BRCA data: comparison of clinical features across the three sample groups.

Clinical variable		Group 1	Group 2	Group 3	p-value
PR status	Negative	163	59	102	< 0.0001
	Positive	11	182	475	
ER status	Negative	153	28	40	< 0.0001
	Positive	23	213	538	
HER2 status	Negative	110	116	308	0.0005
	Positive	12	37	106	
ER/PR/HER2	ER+/PR+/HER2-	2	89	261	< 0.0001
	ER-/PR-/HER2+	10	6	16	
	ER-/PR-/HER2-	90	10	7	

Different from the METABRIC data, there is a lack of subtype information. As in the main text, to provide "indirect support" to the validity of heterogeneity analysis, we compare important clinical features across the three sample groups. The results are summarized in Table S16. It is noted that all these clinical features have been well established as highly critical for breast cancer risk, prognosis, and response to treatment, and the significant p-values suggest that the identified sample groups have clinically meaningful differences.

Table S17: Analysis of TCGA BRCA data using different approaches. Diagonal: sample group sizes. Off-diagonal: Rand index.

	Proposed	CGLasso/MRCE	HeteroGGM	MCGGM
Proposed	186/251/611	0.822	0.761	0.561
CGLasso/MRCE		175/277/596	0.710	0.547
HeteroGGM			567/316/165	0.539
MCGGM				69/47/932

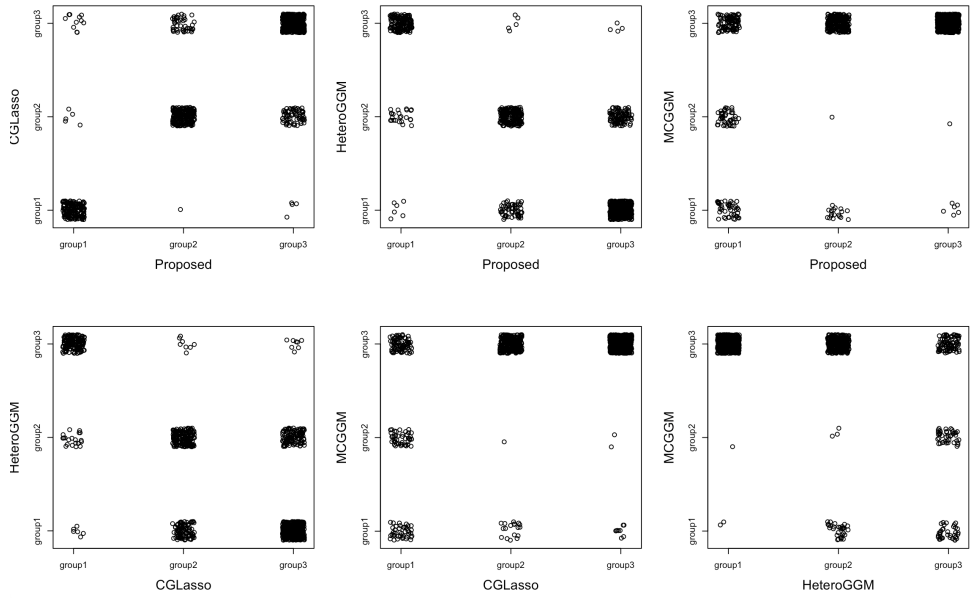

Figure S11: Analysis of TCGA BRCA data using different approaches: comparison of grouping results.

Data is also analyzed using the alternative approaches. For better comparability, we fix three sample groups for the alternatives. The heterogeneity comparison results are summarized in Table S17 and Figure S11. It is found that different approaches lead to different grouping structures. In particular, MCGGM generates highly imbalanced groups. The Rand index values in Table S17 further suggest moderate to strong overlappings. The three networks generated by CGLasso have 250, 230, and 188 edges. Those generated by HeteroGGM have 494, 736 and 596 edges. And those generated by MCGGM have 352, 120 and 168 edges. It is apparent that the network structures are also significantly different. More detailed results are available from

the authors.

Bibliography

Cai, M. and L. Li (2017). Subtype identification from heterogeneous tcga datasets on a genomic scale by multi-view clustering with enhanced consensus. *BMC Medical Genomics 10*, 65–79.

Cancer Genome Atlas, N. (2012). Comprehensive molecular portraits of human breast tumours. *Nature 490*(7418), 61–70.

Danaher, P., P. Wang, and D. M. Witten (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 76*(2), 373.

Fan, J. and J. Lv (2011). Nonconcave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory 57*(8), 5467–5484.

Hao, B., W. W. Sun, Y. Liu, and G. Cheng (2018). Simultaneous clustering and estimation of heterogeneous graphical models. *Journal of Machine Learning Research 18*, 1–58.

Ni, Y., P. Müller, Y. Zhu, and Y. Ji (2018). Heterogeneous reciprocal graphical models. *Biometrics 74*(2), 606–615.

Radchenko, P. and G. Mukherjee (2017). Convex clustering via l_1 fusion penalization. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 79*(5), 1527–1546.

Ren, M., S. Zhang, Q. Zhang, and S. Ma (2022). Gaussian graphical model-based heterogeneity analysis via penalized fusion. *Biometrics 78*(2), 524–535.

Wang, C. and B. Jiang (2020). An efficient admm algorithm for high dimensional precision matrix estimation via penalized quadratic loss. *Computational Statistics & Data Analysis 142*, 106812.

Wytock, M. and Z. Kolter (2013). Sparse gaussian conditional random fields: Algorithms, theory, and application to energy forecasting. In *International Conference on Machine Learning*, pp. 1265–1273. PMLR.