Joint Estimation and Inference for Multiple Multi-layered Gaussian Graphical Models

Subhabrata Majumdar

Abstract: The rapid development of high-throughput technologies has enabled generation of data from biological processes that span multiple layers, like genomic, proteomic or metabolomic data; and pertain to multiple sources, like disease subtypes or experimental conditions. In this work we propose a general statistical framework based on graphical models for horizontal (i.e. across conditions or subtypes) and vertical (i.e. across different layers containing data on molecular compartments) integration of information in such datasets. We start with decomposing the multi-layer problem into a series of two-layer problems. For each two-layer problem, we model the outcomes at a node in the lower layer as dependent on those of other nodes in that layer, as well as all nodes in the upper layer. Following the biconvexity of our objective function, this estimation problem decomposes into two parts, where we use neighborhood selection and subsequent refitting of the precision matrix to quantify the dependency of two nodes in a single layer, and use group-penalized least square estimation to quantify the directional dependency of two nodes in different layers. Finally, to test for differences in these directional dependencies across multiple sources, we devise a hypothesis testing procedure that utilizes already computed neighborhood selection coefficients for nodes in the upper layer. We establish theoretical results for the validity of this testing procedure and the consistency of our estimates, and also evaluate their performance through simulations and a real data application.

Keywords: Data integration; Gaussian Graphical Models; Neighborhood selection; Group lasso

1 Notations

We shall denote scalars by small letters, vectors by bold small letters and matrices by bold capital letters. For any matrix $\mathbf{A}$, $(\mathbf{A})_{ij}$ denote its element in the $(i,j)^{\text{th}}$ position. For $a, b \in \mathbb{N}$, we denote the set of all $a \times b$ real matrices by $\mathbb{M}(a,b)$. For any positive integer c, define $\mathcal{I}_c = \{1, \ldots, c\}$.

2 Model

Consider the two -layered setup:

$$\mathbb{X}^k = (X_1^k, \ldots, X_p^k)^T \sim \mathcal{N}(0, \Sigma_x^k) \tag{2.1}$$

$$\mathbb{Y}^k = \mathbb{X}^k \mathbf{B}^k + \mathbf{E}^k; \quad \mathbf{E}^k = (E_1^k, \ldots, E_p^k)^T \sim \mathcal{N}(0, \Sigma_y^k) \tag{2.2}$$

$$\mathbf{B}^k \in \mathbb{M}(p,q); \quad \Omega_x^k = (\Sigma_x^k)^{-1}; \quad \Omega_y^k = (\Sigma_y^k)^{-1} \tag{2.3}$$

Want to estimate $\{(\Omega_x^k, \Omega_y^k, \mathbf{B}^k); k \in \mathcal{I}_K$ from data $\mathcal{Z}^k = \{(\mathbf{Y}^k, \mathbf{X}^k); \mathbf{Y}^k \in \mathbb{M}(n,q), \mathbf{X}^k \in \mathbb{M}(n,p), k \in \mathcal{I}_K\}$. in presence of known grouping structures $\mathcal{G}_x, \mathcal{G}_y, \mathcal{H}$ respectively.

Estimation of $\{\Omega_x^k\}$ done using JSEM. For the other part, we use the following two-step procedure:

1. Run neighborhood selection on y-network incorporating effects of x-data and an additional block-wise group penalty:

$$\min_{\mathcal{B},\Theta} \left\{ \sum_{j=1}^{q} \frac{1}{n_k} \left[\sum_{k=1}^{K} \|\mathbf{Y}_j^k - (\mathbf{Y}_{-j}^k - \mathbf{X}^k \mathbf{B}_{-j}^k)\boldsymbol{\theta}_j^k - \mathbf{X}^k \mathbf{B}_j^k\|^2 + \sum_{j \neq i} \sum_{g \in \mathcal{G}_y^{ij}} \lambda_{ij}^g \|\boldsymbol{\theta}_{ij}^{[g]}\| \right] + \sum_{b \in \mathcal{G}_x \times \mathcal{G}_y \times \mathcal{H}} \eta^b \|\mathbf{B}^{[b]}\| \right\} \tag{2.4}$$

$$= \min \{ f(\mathcal{Y}, \mathcal{X}, \mathcal{B}, \Theta) + P(\Theta) + Q(\mathcal{B}) \} \tag{2.5}$$

   where $\Theta = \{\Theta_i\}, \mathcal{B} = \{\mathbf{B}^k\}, \mathcal{Y} = \{\mathbf{Y}^k\}, \mathcal{X} = \{\mathbf{X}^k\}, \mathcal{E} = \{\mathbf{E}^k\}$.

   This estimates $\mathcal{B}$ **(possibly refit and/or within-group threshold)** .

2. Step I part 2 and step II of JSEM (see 15-656 pg 6) follows to estimate $\{\Omega_y^k\}$.

The objective function is bi-convex, so we are going to do the following in step 1-

- Start with initial estimates of $\mathcal{B}$ and Θ, say $\mathcal{B}^{(0)}, \Theta^{(0)}$.
- Iterate:

$$\Theta^{(t+1)} = \arg\min \left\{ f(\mathcal{Y}, \mathcal{X}, \mathcal{B}^{(t)}, \Theta^{(t)}) + P(\Theta^{(t)}) \right\} \tag{2.6}$$

$$\mathcal{B}^{(t+1)} = \arg\min \left\{ f(\mathcal{Y}, \mathcal{X}, \mathcal{B}^{(t)}, \Theta^{(t+1)}) + Q(\mathcal{B}^{(t)}) \right\} \tag{2.7}$$

- Continue till convergence.

3 Two-sample testing

Suppose there are two disease subtypes: $k = 1, 2$, and we are interested in testing whether the downstream effect of a predictor is X-data is same across both subtypes, i.e. if $\mathbf{b}_i^1 = \mathbf{b}_i^2$ for some $i \in \{1, \ldots, p\}$. For this we consider the modified optimization problem:

$$\min_{\mathcal{B},\Theta} \frac{1}{n} \left\{ \sum_{j=1}^{q} \sum_{k=1}^{2} \|\mathbf{Y}_j^k - \mathbf{Y}_{-j}^k \boldsymbol{\theta}_j^k - \mathbf{X}^k \mathbf{B}_j^k\|^2 + \sum_{j \neq j'} \lambda_{jj'} \|\boldsymbol{\theta}_{jj'}^*\| + \sum_{i=1}^{p} \eta_i \|\mathbf{B}_{i*}^*\| \right\} \tag{3.1}$$

$$= \min \{ f(\mathcal{Y}, \mathcal{X}, \mathcal{B}, \Theta) + P(\Theta) + Q(\mathcal{B}) \} \tag{3.2}$$

with $n_1 = n_2 = n$ for simplicity; and $\mathbf{B}^k = (\mathbf{b}_1^k, \ldots, \mathbf{b}_q^k), (\mathbf{B}_{i*}^*) \in \mathbb{R}^{q \times K}$

4 Conditions

Conditions A1 from JSEM paper holds for $\mathcal{X}$ and $\mathcal{E}$. Also A2, A3 from JSEM paper.

5 Results

To prove the results in this section, we shall use a reparametrization of the neighborhood coefficients at the lower level. Specifically, notice that for $j \in \mathcal{I}_q, k \in \mathcal{I}_K$, the corresponding summand in $f(\mathcal{Y}, \mathcal{X}, \mathcal{B}, \Theta)$ can be rearranged as

$$\begin{aligned} \|\mathbf{Y}_j^k - \mathbf{X}^k \mathbf{B}_j^k - (\mathbf{Y}_{-j}^k - \mathbf{X}^k \mathbf{B}_{-j}^k) \boldsymbol{\theta}_j^k\|^2 &= \|\mathbf{Y}_j^k - \mathbf{Y}_{-j}^k \boldsymbol{\theta}_j^k - (\mathbf{X}^k \mathbf{B}_j^k - \mathbf{X}^k \mathbf{B}_{-j}^k \boldsymbol{\theta}_j^k)\|^2 \\ &= \|(\mathbf{Y} - \mathbf{X}\mathbf{B}) \mathbf{T}_j^k\|^2 \end{aligned}$$

where

$$T_{jj'}^k = \begin{cases} 1 \text{ if } j = j' \\ -\theta_{jj'}^k \text{ if } j \neq j' \end{cases}$$

Thus, with $\mathbf{T}^k := (\mathbf{T}_j^k)_{j \in \mathcal{I}_q}$, we have

$$f(\mathcal{Y}, \mathcal{X}, \mathcal{B}, \Theta) = \frac{1}{n} \sum_{j=1}^{p} \sum_{k=1}^{K} \|(\mathbf{Y}^k - \mathbf{X}^k \mathbf{B}^k) \mathbf{T}_j^k\|^2 = \frac{1}{n} \sum_{k=1}^{K} \|\mathbf{Y}^k - \mathbf{X}^k \mathbf{B}^k) \mathbf{T}^k\|_F^2 = \sum_{k=1}^{K} \mathrm{Tr}(\mathbf{S}^k (\mathbf{T}^k)^2)$$

where $\mathbf{S}^k = (1/n)(\mathbf{Y}^k - \mathbf{X}^k \mathbf{B}^k)(\mathbf{Y}^k - \mathbf{X}^k \mathbf{B}^k)^T$ is the sample covariance matrix.

Now suppose $\boldsymbol{\beta} = \mathrm{vec}(\mathbf{B})$, and any subscript or superscript on $\mathbf{B}$ will be passed on to $\boldsymbol{\beta}$. Denote by $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\Theta}}$ the generic estimators given by

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{pq}}{\arg\min} \left\{ -2\boldsymbol{\beta}^T \widehat{\boldsymbol{\gamma}} + \boldsymbol{\beta}^T \widehat{\boldsymbol{\Gamma}} \boldsymbol{\beta} + \lambda_n \sum_{g \in \mathcal{G}} \|\boldsymbol{\beta}^{[g]}\| \right\} \tag{5.1}$$

$$\widehat{\Theta}_j = \underset{\Theta_j \in \mathbb{M}(q-1,K)}{\arg\min} \left\{ \frac{1}{n} \sum_{k=1}^{K} \|\mathbf{Y}_j^k - \mathbf{X}^k \widehat{\mathbf{B}}_j^k - (\mathbf{Y}_{-j}^k - \mathbf{X}^k \widehat{\mathbf{B}}_{-j}^k) \boldsymbol{\theta}_j^k\|^2 + \sum_{j \neq j'} \sum_{g \in \mathcal{G}_y^{jj'}} \lambda_{jj'}^g \|\boldsymbol{\theta}_{jj'}^{[g]}\| \right\} \tag{5.2}$$

where

$$\widehat{\boldsymbol{\Gamma}} = \begin{bmatrix} (\widehat{\mathbf{T}}^1)^2 \otimes \frac{(\mathbf{X}^1)^T \mathbf{X}^1}{n} & & \\ & \ddots & \\ & & (\widehat{\mathbf{T}}^K)^2 \otimes \frac{(\mathbf{X}^K)^T \mathbf{X}^K}{n} \end{bmatrix}; \quad \widehat{\boldsymbol{\gamma}} = \begin{bmatrix} (\widehat{\mathbf{T}}^1)^2 \otimes \frac{(\mathbf{X}^1)^T}{n} \\ \vdots \\ (\widehat{\mathbf{T}}^K)^2 \otimes \frac{(\mathbf{X}^K)^T}{n} \end{bmatrix} \begin{bmatrix} \mathrm{vec}(\mathbf{Y}^1) \\ \vdots \\ \mathrm{vec}(\mathbf{Y}^K) \end{bmatrix}$$

with $\widehat{\mathbf{T}}^k$ defined the same way using $\widehat{\boldsymbol{\theta}}_j^k$ as we defined $\mathbf{T}^k$ using $\boldsymbol{\theta}_j^k$.

Theorem 5.1. *Assume fixed $\mathcal{X}, \mathcal{E}$ and deterministic $\widehat{\mathcal{B}} = \{\widehat{\mathbf{B}}^k\}$. Also for $k = 1, \ldots, K$,*
(T1) $\|\widehat{\mathbf{B}}^k - \mathbf{B}_0^k\|_F \leq v_\beta$, *where $v_\beta =$* **tbd**;
(T2) $\|\mathbf{X}^k(\widehat{\mathbf{B}}^k - \mathbf{B}_0^k)\|_\infty \leq c(v_\beta)$ *for some non-negative function $c(.)$;*
(T3) *Assumption (A1) holds for $\widehat{\mathbf{S}}^k := (\mathbf{Y}^k - \mathbf{X}^k\widehat{\mathbf{B}}^k)(\mathbf{Y}^k - \mathbf{X}^k\widehat{\mathbf{B}}^k)^T/n, k \in \mathcal{I}_K$.*
(T4) *Assumption (A2) holds for $\widehat{\mathbf{E}}^k$. Then*
(I) Estimation consistency.
Further if A1 holds with $s = s_0$, and A3 is satisfied then
(II) Direction consistency.

Proof of Theorem 5.1. Part I. Follows proof of thm 1 in 15-656. The proof has 3 parts: consistency of neighborhood regression, selection of edge sets, and finally the refitting step.

Define $\mathbf{T}_{0,j}^k$ the same way as $\mathbf{T}_j^k$. For any $g \in \mathcal{G}^{jj'}, k \in g$, and $j \neq j'$, let

$$\widehat{\boldsymbol{\epsilon}}_j^k = (\mathbf{Y}^k - \mathbf{X}^k\widehat{\mathbf{B}}^k)\mathbf{T}_{0,j}^k; \quad \widehat{\zeta}_{jj'}^k = \frac{(\widehat{\boldsymbol{\epsilon}}_j^k)^T\mathbf{Y}_{j'}^k}{n}; \quad \widehat{\boldsymbol{\zeta}}_{jj'}^{[g]} = (\widehat{\zeta}_{jj'}^k)_{k\in g}$$

Consider the random event $\mathcal{A} = \bigcap_{j,j'\neq j,g} \mathcal{A}_{jj'}^g$ with $\mathcal{A}_{jj'}^g = \{2\|\widehat{\boldsymbol{\zeta}}_{jj'}^{[g]}\| \leq \lambda_{jj'}^g\}$.

Proposition 5.2. *Given that $\lambda_{jj'}^g$ are chosen as*

$$\lambda_{jj'}^g \geq \max_{k\in g} \frac{2}{\sqrt{n\omega_{jj}^k}} \left(\sqrt{|g|(1 + 2c(v_\beta))} + \frac{\pi}{\sqrt{2}}\sqrt{r \log G_0}\right)$$

we shall have $\mathbb{P}(\mathcal{A}) \geq 1 - 2pG_0^{1-q}$ for some $r > 1$.

Proof of Proposition 5.2. We follow the proof of Lemma E.2 in 15-656, with $\mathbf{Y}_j^k, \widehat{\boldsymbol{\epsilon}}_j^k, \widehat{\zeta}_{jj'}^k, \widehat{\boldsymbol{\zeta}}_{jj'}^{[g]}$ in place of $\mathbf{X}_j^k, \boldsymbol{\epsilon}_i^k, \zeta_{ij}^k, \boldsymbol{\zeta}_{ij}^{[g]}$ respectively. Proceeding in a similar fashion we get

$$\|\widehat{\boldsymbol{\zeta}}_{jj'}^{[g]}\|^2 = \frac{1}{n}\left[\|\mathbf{Z}^{[g]}\|^2 + \sum_{k\in g}\left\{2Z^k(\mathbf{Q}_{j'}^k)^T\boldsymbol{\delta}_j^k + |(\mathbf{Q}_{j'}^k)^T\boldsymbol{\delta}_j^k|^2\right\}\right]$$

where $\mathbf{Z}^{[g]} = (Z^k)_{k\in g}; Z^k = (\mathbf{Q}_{j'}^k)^T\boldsymbol{\epsilon}_j^k$ with $\boldsymbol{\epsilon}_j^k := (\mathbf{Y}^k - \mathbf{X}^k\mathbf{B}_0^k)\mathbf{T}_{0,j}^k$, $\mathbf{Q}_j^k$ is the first eigenvector of $\mathbf{Y}_j^k(\mathbf{Y}_j^k)^T/n$, and $\boldsymbol{\delta}_j^k := \mathbf{X}^k(\mathbf{B}_0^k - \widehat{\mathbf{B}}^k)\mathbf{T}_{0,j}^k$.

By cauchy-schwarz inequality, $|(\mathbf{Q}_{j'}^k)^T\boldsymbol{\delta}_j^k| \leq \|\boldsymbol{\delta}_j^k\| \leq \|\mathbf{X}^k(\mathbf{B}_0^k - \widehat{\mathbf{B}}^k)\|_\infty\|\mathbf{T}_{0,j}\|_1$. Now since Ω_y^k is diagonally dominant,

$$\sum_{j\neq j'} |T_{0,jj'}^k| = \sum_{j\neq j'} |\theta_{0,jj'}^k| = \sum_{j\neq j'} \frac{|\sigma_{y,jj'}^k|}{\sigma_{y,jj}^k} \leq 1$$

Also $T_{0,jj}^k = 1$, so that $\|\mathbf{T}_{0,j}^k\|_1 \leq 2$. Thus $\|\boldsymbol{\delta}_j^k\| \leq 2\|\mathbf{X}^k(\mathbf{B}_0^k - \widehat{\mathbf{B}}^k)\|_\infty$. Hence by assumption (T2),

$$\|\widehat{\boldsymbol{\zeta}}_{ij}^{[g]}\| \leq \frac{1}{\sqrt{n}}(\|\mathbf{Z}^{[g]}\| + 2|g|c(v_\beta))$$

so that

$$\mathbb{P}(\{\mathcal{A}_{ij}^g\}^c) = \mathbb{P}\left(\|\widehat{\boldsymbol{\zeta}}_{ij}^{[g]}\| > \frac{\lambda_{ij}^g}{2}\right) \leq \mathbb{P}\left(\|\mathbf{Z}^{[g]}\| > \frac{\sqrt{n}\lambda_{ij}^g}{2} - 2|g|c(v_\beta)\right)$$

We now proceed through the proof of Lemma E.2 in 15-656 to end up with the choice of λ_{ij}^g. □

All subsequent derivations in the theorem go through with the new choice of λ_{ij}^g.

Part II. Proof of Thm 2 in 15-656 follows. We only need a new bound for $Var(\mathbf{Y}_i^k|\mathbf{Y}_{-i}^k, \mathbf{X}^k, \widehat{\mathbf{B}}_i^k)$. For this we have

$$Var(\mathbf{Y}_i^k|\mathbf{Y}_{-i}^k, \mathbf{X}^k, \widehat{\mathbf{B}}_i^k) = \mathbb{E}(\widehat{\boldsymbol{\epsilon}}_i^k)^2 = \mathbb{E}(\boldsymbol{\epsilon}_i^k + \boldsymbol{\delta}_i^k)^2 \leq \left(\frac{1}{d_0} + \frac{c(v_\beta)}{n}\right)^2$$

applying cauchy-schwarz inequality followed by assumption (A2). Now Replace $1/\sqrt{nd_0}$ in choice of λ, α_n in Thm 2 statement with $1/\sqrt{n}(\sqrt{1/d_0} + \sqrt{c(v_\beta)/n})$.

□

Proposition 5.3. *Consider deterministic $\widehat{\mathcal{B}}$ satisfying assumption (T1). Then for sample size $n \succsim \log(pq)$ and $k \in \mathcal{I}_K$,*

1. *We have $\|\mathbf{X}^k(\hat{\mathbf{B}}^k - \mathbf{B}_0^k)\|_\infty \leq c(v_\beta)$, where*

$$c(v_\beta) = \sqrt{n} v_\beta \left[\sqrt{\frac{\log 4 + \tau_1 \log p}{c_x^k n}} + \max_j \sigma_{x,jj}^k\right]^{1/2} \quad ; \quad c_x^k = \left[128(1 + 4\Lambda_{\max}(\Sigma_x^k))^2 \max_j (\sigma_{x,jj}^k)^2\right]^{-1}$$

   with probability $\geq 1 - 1/p^{\tau_1 - 2}, \tau_1 > 2$.

2. *$\widehat{\mathbf{E}}^k$ satisfies the RE condition: $\widehat{\mathbf{E}}^k \sim RE(\psi^*, \phi^*)$, where*

$$\psi^* = \frac{\Lambda_{\min}(\Sigma_x^k)}{2}; \quad \phi^* = \frac{\psi^* \log p}{n} + 2v_\beta c_2 [\Lambda_{\max}(\Sigma_x^k)\Lambda_{\max}(\Sigma_y^k)]^{1/2} \sqrt{\frac{\log(pq)}{n}}$$

   with probability $\geq 1 - 6c_1 \exp[-(c_2^2 - 1)\log(pq)], c_1 > 0, c_2 > 1$.

Proof of Proposition 5.3. For any sub-gaussian zero-mean design matrix $\mathbf{X} \in \mathbb{M}(n, p)$ with parameters (Σ_x, σ_x^2), and any $\hat{\mathbf{B}}, \mathbf{B}_0 \in \mathbb{M}(p, q)$ such that $\|\hat{\mathbf{B}} - \mathbf{B}_0\|_F \leq v_\beta$, we follow the proof of Proposition 3 in **?** to obtain that the following holds

$$\left\|(\widehat{\mathbf{B}} - \mathbf{B}_0)^T \left(\frac{\mathbf{X}^T\mathbf{X}}{n}\right)(\widehat{\mathbf{B}} - \mathbf{B}_0)\right\|_\infty \leq v_\beta^2 \left[\sqrt{\frac{\log 4 + \tau_1 \log p}{c_x n}} + \max_j \sigma_{x,jj}^k\right]$$

with probability $\geq 1 - 1/p^{\tau_1 - 2}$ for some $\tau_1 > 2$, where

$$c_x = \left[128(1 + 4\sigma_x^2)^2 \max_j (\sigma_{x,jj}^k)^2\right]^{-1}$$

Here we substitute $\mathbf{X}, \hat{\mathbf{B}}, \mathbf{B}_0$ with $\mathbf{X}^k, \hat{\mathbf{B}}^k, \hat{\mathbf{B}}_0^k$ respectively. Since rows of $\mathbf{X}^k$ come independently from $\mathcal{N}(\mathbf{0}, \Sigma_x^k)$, σ_x^2 in our case is the spectral norm of Σ_x^k (**?**), which is $\Lambda_{\max}(\Sigma_x^k)$. Finally

$$\|\mathbf{X}^k(\hat{\mathbf{B}}^k - \mathbf{B}_0^k)\|_\infty \leq \sqrt{\left\|(\widehat{\mathbf{B}}^k - \mathbf{B}_0^k)^T(\mathbf{X}^k)^T\mathbf{X}^k(\widehat{\mathbf{B}}^k - \mathbf{B}_0^k)\right\|_\infty}$$

The proof of part 1 is immediate now.

For part 2, we start with an auxiliary lemma:

Lemma 5.4. *For a sub-gaussian design matrix $\mathbf{X} \in \mathbb{M}(n, p)$ with columns having mean $\mathbf{0}_p$ and covariance matrix Σ_x, the sample covariance matrix $\widehat{\Sigma}_x = \mathbf{X}^T\mathbf{X}/n$ satisfies the RE condition*

$$\widehat{\Sigma}_x \sim RE\left(\frac{\Lambda_{\min}(\Sigma_x)}{2}, \frac{\Lambda_{\min}(\Sigma_x)\log p}{2n}\right)$$

This is the same as Lemma 2 in Appendix B of **?** and its proof can be found there. Now denote $\widehat{\mathbf{E}}^k = \mathbf{Y}^k - \mathbf{X}^k\widehat{\mathbf{B}}^k$. For $\mathbf{v} \in \mathbb{R}^q$, we have

$$\begin{aligned}
\mathbf{v}^T\widehat{\mathbf{S}}^k\mathbf{v} &= \frac{1}{n}\|\widehat{\mathbf{E}}^k\mathbf{v}\|^2 \\
&= \frac{1}{n}\|(\mathbf{E}^k + \mathbf{X}^k(\mathbf{B}_0^k - \widehat{\mathbf{B}}^k))\mathbf{v}\|^2 \\
&= \mathbf{v}^T\mathbf{S}^k\mathbf{v} + \frac{1}{n}\|\mathbf{X}^k(\mathbf{B}_0^k - \widehat{\mathbf{B}}^k)\mathbf{v}\|^2 + 2\mathbf{v}^T(\mathbf{B}_0^k - \widehat{\mathbf{B}}^k)^T\left(\frac{(\mathbf{X}^k)^T\mathbf{E}^k}{n}\right)\mathbf{v} \qquad (5.3)
\end{aligned}$$

For the first summand, $\mathbf{v}^T\mathbf{S}^k\mathbf{v} \geq \psi_y\|\mathbf{v}\|^2 - \phi_y\|\mathbf{v}\|_1^2$ with $\psi_y = \Lambda_{\min}(\Sigma_y^k)/2, \phi_y = \psi_y \log p/n$ by applying Lemma 5.4 on $\mathbf{S}^k$. The second summand is greater than or equal to 0. For the third summand,

$$2\mathbf{v}^T(\widehat{\mathbf{B}}^k - \mathbf{B}_0^k)^T\left(\frac{(\mathbf{X}^k)^T\mathbf{E}^k}{n}\right)\mathbf{v} \geq -2v_\beta\left\|\frac{(\mathbf{X}^k)^T\mathbf{E}^k}{n}\right\|_\infty \|\mathbf{v}\|_1^2$$

by assumption (T1). Now we use another lemma:

Lemma 5.5. *For zero-mean independent sub-gaussian matrices $\mathbf{X} \in \mathbb{M}(n,p), \mathbf{E} \in \mathbb{M}(n,q)$ with parameters (Σ_x, σ_x^2) and (Σ_e, σ_e^2) respectively, given that $n \gtrsim \log(pq)$ the following holds with probability $\geq 1 - 6c_1\exp[-(c_2^2-1)\log(pq)]$ for some $c_1 > 0, c_2 > 1$:*

$$\frac{1}{n}\|\mathbf{X}^T\mathbf{E}\|_\infty \leq c_2[\Lambda_{\max}(\Sigma_x)\Lambda_{\max}(\Sigma_e)]^{1/2}\sqrt{\frac{\log(pq)}{n}}$$

This is a part of Lemma 3 of Appendix B in **?**, and has been proved therein. Here we take $\mathbf{X} \equiv \mathbf{X}^k, \mathbf{E} \equiv \mathbf{E}^k$, and subsequently collecting all summands in (5.3) get

$$\mathbf{v}^T\widehat{\mathbf{S}}^k\mathbf{v} \geq \psi_y\|\mathbf{v}\|^2 - \left(\phi_y + 2v_\beta c_2[\Lambda_{\max}(\Sigma_x^k)\Lambda_{\max}(\Sigma_y^k)]^{1/2}\sqrt{\frac{\log(pq)}{n}}\right)\|\mathbf{v}\|_1^2$$

with probability $\geq 1 - 6c_1\exp[-(c_2^2-1)\log(pq)]$. This concludes the proof. □

Now concentrate on the k-population estimation problem. We want to obtain

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{pqK}}{\arg\min}\{-2\boldsymbol{\beta}\widehat{\boldsymbol{\gamma}} + \boldsymbol{\beta}^T\boldsymbol{\Gamma}\boldsymbol{\beta} + \|\boldsymbol{\beta}\|_{2,g}\}$$

with

$$\boldsymbol{\beta} = \begin{bmatrix}\mathrm{vec}(\mathbf{B}^1)\\ \vdots \\ \mathrm{vec}(\mathbf{B}^K)\end{bmatrix}; \quad \boldsymbol{\Gamma} = \begin{bmatrix} I_q \otimes (\mathbf{X}^1)TX^1/n) & & \\ & \ddots & \\ & & I_q \otimes (\mathbf{X}^K)^TX^K/n)\end{bmatrix}$$

Theorem 5.6.