A Consistent Method for Graph Based Anomaly Localization – Supplemental Material –

Satoshi Hara **Tetsuro Morimura** **Toshihiro Takahashi** **Hiroki Yanagisawa**
IBM Research – Tokyo, Japan, {satohara, tetsuro, e30137, yanagis}@jp.ibm.com

Taiji Suzuki
Tokyo Institute of Technology & PRESTO, JST, Japan, s-taiji@is.titech.ac.jp

6 Settings for the Synthetic Data

6.1 Generation of Synthetic Matrices

Let Λ_{A} and Λ_{B} be a pair of matrices. For our simulation in Section 4.1, we generated matrices that satisfy Assumption 1. Specifically, we generated matrices that simultaneously satisfy three conditions:

- positive-definite: $\Lambda_u \succeq \nu I_d$ for $u \in \{\mathrm{A}, \mathrm{B}\}$,
- sparse: $\Lambda_{u,ij} = 0$ for $(i,j) \in \Omega_u$ with $u \in \{\mathrm{A}, \mathrm{B}\}$,
- homogeneous: $\Lambda_{\mathrm{A},ij} = \Lambda_{\mathrm{B},ij}$ for $(i,j) \in \Pi$,

where ν is some positive constant, and Ω_{A}, Ω_{B}, and Π are properly chosen subsets of $\{1, 2, \ldots, d\}^2$. In particular, all pairs of healthy variable indices have to be involved in Π so that the homogeneity is satisfied. In addition, based on the sparseness and homogeneity conditions, $\Omega_{\mathrm{A}} \cap \Omega_{\mathrm{B}} \subseteq \Pi$ must be satisfied. We note that, although the first condition is not necessary for general adjacency matrices, we used it since both the covariance and precision matrices we used in the simulation are positive-definite matrices.

The matrix generation procedure starts by specifying ν, Ω_{A}, Ω_{B}, Π, and two reference matrices $C_{\mathrm{A}}, C_{\mathrm{B}} \in \mathbb{R}^{d \times d}$. Here we chose C_{A} and C_{B} to be symmetric but not necessarily to be positive-definite. We then derive Λ_{A} and Λ_{B} as the solution to the problem:

$$\begin{aligned}\min_{\{\Lambda_u\}_{u\in\{\mathrm{A},\mathrm{B}\}}} \frac{1}{2} \sum_{u\in\{\mathrm{A},\mathrm{B}\}} \|\Lambda_u - C_u\|_{\mathrm{F}}^2,\\ \text{s.t. } \Lambda_u \succeq \nu I_d \ (u \in \{\mathrm{A},\mathrm{B}\}),\\ \Lambda_{u,ij} = 0 \text{ for } (i,j) \in \Omega_u \ (u \in \{\mathrm{A},\mathrm{B}\}),\\ \Lambda_{\mathrm{A},ij} = \Lambda_{\mathrm{B},ij} \text{ for } (i,j) \in \Pi,\end{aligned}$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius-norm of the matrix. This problem corresponds to searching for the matrices Λ_{A} and Λ_{B} that are closest to the reference matrices C_{A} and C_{B} under the specified conditions. The problem is convex and we can solve it by using ADMM [24]. We first rewrite the problem into an equivalent form:

$$\begin{aligned}\min_{\{X_u\}_{u\in\{\mathrm{A},\mathrm{B}\}},\{Y_u\}_{u\in\{\mathrm{A},\mathrm{B}\}}} \frac{1}{2} \sum_{u\in\{\mathrm{A},\mathrm{B}\}} (\|X_u - C_u\|_{\mathrm{F}}^2 \\ + \delta_{\Omega_u}(X_u) + \tilde{\delta}(Y_u)),\\ \text{s.t. } X_u - Y_u + \nu I_d = 0 \ (u \in \{\mathrm{A},\mathrm{B}\}),\\ X_{\mathrm{A},ij} - X_{\mathrm{B},ij} = 0 \text{ for } (i,j) \in \Pi,\end{aligned}$$

where $\delta_{\Omega_u}(X_u)$ and $\tilde{\delta}(Y_u)$ are the indicator functions defined as

$$\delta_{\Omega_u}(X_u) := \begin{cases} 0 & \text{if } X_{u,ij} = 0 \text{ for all } (i,j) \in \Omega_u, \\ \infty & \text{otherwise,} \end{cases}$$

$$\tilde{\delta}(Y_u) := \begin{cases} 0 & \text{if } Y_u \succeq 0, \\ \infty & \text{otherwise,} \end{cases}$$

and we have $\Lambda_u = Y_u$ as the solution for $u \in \{\mathrm{A}, \mathrm{B}\}$. Let $Z_u \in \mathbb{R}^{d \times d}$ ($u \in \{\mathrm{A}, \mathrm{B}, \mathrm{O}\}$) be the matrix of Lagrange multipliers. We then define the Augmented Lagrangian (AL) function as

$$\begin{aligned}\mathcal{L}_\beta(X,Y,Z) = \frac{1}{2} \sum_{u\in\{\mathrm{A},\mathrm{B}\}} (\|X_u - C_u\|_{\mathrm{F}}^2 + \delta_{\Omega_u}(X_u) + \tilde{\delta}(Y_u)) \\ + \frac{\beta}{2} (\sum_{u\in\{\mathrm{A},\mathrm{B}\}} \|X_u - Y_u + \nu I_d + \frac{1}{\beta} Z_u\|_{\mathrm{F}}^2 \\ + \|E_\Pi \odot (X_{\mathrm{A}} - X_{\mathrm{B}}) + \frac{1}{\beta} Z_{\mathrm{O}}\|_{\mathrm{F}}^2),\end{aligned}$$

where we set $X := \{X_u\}_{u\in\{\mathrm{A},\mathrm{B}\}}$, $Y := \{Y_u\}_{u\in\{\mathrm{A},\mathrm{B}\}}$, and $Z := \{Z_u\}_{u\in\{\mathrm{A},\mathrm{B},\mathrm{O}\}}$ to simplify the notation, $\odot$ denotes the Hadamard product of matrices, and E_Π is an indicator matrix of the set Π defined as

$$E_{\Pi,ij} := \begin{cases} 1 & \text{if } (i,j) \in \Pi, \\ 0 & \text{otherwise.} \end{cases}$$

The optimization procedure of ADMM is defined using this AL function. Specifically, we repeat these steps until one of the termination criteria is fulfilled:

$$\begin{cases} X^{(k+1)} \in \mathrm{argmin}_X \, \mathcal{L}_\beta(X, Y^{(k)}, Z^{(k)}), \\ Y^{(k+1)} \in \mathrm{argmin}_Y \, \mathcal{L}_\beta(X^{(k+1)}, Y, Z^{(k)}), \\ Z_u^{(k+1)} = Z_u^{(k)} + \beta(X_u^{(k+1)} - Y_u^{(k+1)} + \nu I_d) \;\; (u \in \{\mathrm{A}, \mathrm{B}\}), \\ Z_{\mathrm{O}}^{(k+1)} = Z_{\mathrm{O}}^{(k)} + \beta E_\Pi \odot (X_{\mathrm{A}}^{(k+1)} - X_{\mathrm{B}}^{(k+1)}). \end{cases}$$

The first step, the update of X, can be decomposed into individual problems on each (i,j)th entry given as

$$\min_{X_{\mathrm{A},ij}, X_{\mathrm{B},ij}} \frac{1}{2} \sum_{u \in \{\mathrm{A},\mathrm{B}\}} \{(X_{u,ij} - C_{u,ij})^2 + \beta(X_{u,ij} - P_{u,ij}^{(k)})^2\} + \frac{\beta}{2}\{E_{\Pi,ij}(X_{\mathrm{A},ij} - X_{\mathrm{B},ij}) + \frac{1}{\beta} Z_{\mathrm{O},ij}\}^2,$$

$$\text{s.t. } X_{u,ij} = 0 \text{ for } (i,j) \in \Omega_u \; (u \in \{\mathrm{A},\mathrm{B}\}),$$

where $P_u^{(k)}$ is the matrix defined as

$$P_u^{(k)} = Y_u^{(k)} - \frac{1}{\beta} Z_u^{(k)} - \nu I_d \; (u \in \{\mathrm{A},\mathrm{B}\}).$$

When $E_{\Pi,ij} = 0$ or $(i,j) \notin \Pi$, the problem can be further reduced into the individual problems on $X_{\mathrm{A},ij}$ and $X_{\mathrm{B},ij}$. Hence, we have the solution

$$X_{u,ij}^{(k+1)} = \begin{cases} \frac{1}{1+\beta}(C_{u,ij} + \beta P_{u,ij}^{(k)}) & \text{if } (i,j) \notin \Omega_u, \\ 0 & \text{otherwise.} \end{cases}$$

For the case of $E_{\Pi,ij} = 1$ and $(i,j) \notin \Omega_u$ $(u \in \{\mathrm{A},\mathrm{B}\})$, we have the solution

$$\begin{bmatrix} X_{\mathrm{A},ij} \\ X_{\mathrm{B},ij} \end{bmatrix} = \frac{1}{(1+\beta)(1+3\beta)} \times \begin{bmatrix} 1+2\beta & \beta \\ \beta & 1+2\beta \end{bmatrix} \begin{bmatrix} C_{\mathrm{A},ij} + \beta P_{\mathrm{A},ij}^{(k)} - Z_{\mathrm{O},ij}^{(k)} \\ C_{\mathrm{B},ij} + \beta P_{\mathrm{B},ij}^{(k)} + Z_{\mathrm{O},ij}^{(k)} \end{bmatrix}.$$

The update problem for Y can be decomposed into individual problems on Y_{A} and Y_{B}, which are given as

$$\min_{Y_u} \frac{1}{2}\|Y_u - Q_u^{(k)}\|_{\mathrm{F}}^2, \text{ s.t. } Y_u \succeq 0 \; (u \in \{\mathrm{A},\mathrm{B}\}),$$

$$Q_u^{(k)} = X_u^{(k+1)} + \frac{1}{\beta} Z_u^{(k)} + \nu I_d.$$

This problem is equivalent to the Euclidean projection of the matrix $Q_u^{(k)}$ onto the positive semidefinite cone. Hence, this can be computed analytically using the eigenvalue decomposition. Here we assume all of the matrices are symmetric, which can be assured by initializing all of the matrices to be symmetric. Let $Q_u^{(k)} = UDU^\top$ be the eigenvalue decomposition with $D = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_d)$. We then have $Y_u^{(k+1)} = U\tilde{D}U^\top$ with $D = \mathrm{diag}(\tilde{\sigma}_1, \tilde{\sigma}_2, \ldots, \tilde{\sigma}_d)$ where $\tilde{\sigma}_i = \max(\sigma_i, 0)$.

6.2 Parameter Settings

This subsection explains how the sparsity patterns Ω_{A} and Ω_{B} and the shared pattern Π are chosen. For the sparsity pattern, we randomly pick index pairs so that the average size of Ω_u is $3,000$ when $d = 100$. This corresponds to choosing 70% of the matrix entries to be zero. For $d = 200$, they are set to be $8,000$ (80%). The set Π, the shared pattern, is chosen as $\Pi = (\mathcal{I} \times \mathcal{I}) \cup (\Omega_{\mathrm{A}} \cap \Omega_{\mathrm{B}}) \cup \Pi_+$. The first set is all pairs of healthy variables and the second set corresponds to the shared zero entries between the two matrices. The set Π_+ specifies the common non-zero entries, so that $\Pi_+ \subseteq \Omega_{\mathrm{A}}^c \cap \Omega_{\mathrm{B}}^c$ holds where c denotes the compliment of the set. For the construction of Π_+, we consider the subset of $\Omega_{\mathrm{A}}^c \cap \Omega_{\mathrm{B}}^c$ defined as

$$\Pi_0 := (\mathcal{I} \times \mathcal{I})^c \cap (\Omega_{\mathrm{A}}^c \cap \Omega_{\mathrm{B}}^c),$$

which means that Π_0 specifies the set of index pairs whose corresponding edges are non-zeros, and hence connected to anomalous variables. We then randomly pick index pairs $(i,j) \in \Pi_0$ and add to Π_+. We set the size $|\Pi|$ to be 70% of $|\Pi_0|$ for both $d = 100$ and 200.

Across all of the settings, the value of ν is set as 10^{-3}. Here is how the reference matrices C_{A} and C_{B} are generated. We first generate random matrices $L_{\mathrm{A}}, L_{\mathrm{B}} \in \mathbb{R}^{d \times \lceil\sqrt{d}\rceil}$ where each entry of each of the matrices is generated from a standard Gaussian distribution $\mathcal{N}(0,1)$. We then set $C_{\mathrm{A}} = L_{\mathrm{A}} L_{\mathrm{A}}^\top$ and $C_{\mathrm{B}} = L_{\mathrm{B}} L_{\mathrm{B}}^\top$, and rescale them so that their diagonals are one. Note that the generated matrices have at most rank $\lceil\sqrt{d}\rceil$ which implies that they are rank deficient. Therefore, the resulting covariance and precision matrices encourage variables to have higher dependencies with other variables.

6.3 Generation of Matrices with a Concentrated Anomaly Pattern

In Section 4.1, Figure 6, we used a concentrated anomaly pattern for the simulation. The synthetic matrices were generated by modifying the generated matrices in Section 6.1. Here, we denote by $M_{\mathcal{I}\times\mathcal{J}}$ the sub-matrix consisting of the components of a matrix M with indices in $\mathcal{I} \times \mathcal{J}$. We choose a subset $\mathcal{H} \subseteq \mathcal{I}$ and update the matrix Λ_{A} by $\Lambda_{\mathrm{A},\mathcal{H}\times\mathcal{I}} \leftarrow a\Lambda_{\mathrm{A},\mathcal{H}\times\mathcal{I}}$ and $\Lambda_{\mathrm{A},\mathcal{I}\times\mathcal{H}} \leftarrow a\Lambda_{\mathrm{A},\mathcal{I}\times\mathcal{H}}$ where $a > 1$ is some large positive value. The matrix Λ_{B} is updated in the same manner. Since the modified matrices Λ_{A} and Λ_{B} are no longer positive definite, we add an identity matrix $\Lambda_{\mathrm{A}} \leftarrow \Lambda_{\mathrm{A}} + cI_d$ and $\Lambda_{\mathrm{B}} \leftarrow \Lambda_{\mathrm{B}} + cI_d$ so that the matrices become positive definite. In our simulation in Section 4.1, we set the size of $\mathcal{H}$ to be 5 and $a = 10$.

7 Proofs of Theorems

Here, we provide the proofs of the theorems in the manuscript. We first enumerate key lemmas relevant to a GGM estimator that are needed in the proofs, and then present the proofs of theorems.

7.1 Key Lemmas on GGM estimators

Suppose that we have two groups of i.i.d. samples $\{x_i^{(\mathrm{A})}\}_{i=1}^{n_\mathrm{A}} \sim p_\mathrm{A}$ and $\{x_i^{(\mathrm{B})}\}_{i=1}^{n_\mathrm{B}} \sim p_\mathrm{B}$ ($x_i^{(\mathrm{A})} \in \mathbb{R}^d$ and $x_i^{(\mathrm{B})} \in \mathbb{R}^d$). To estimate the precision matrices Λ_A and Λ_B corresponding to p_A and p_B, respectively, we use the following graphical–Lasso type estimators:

$$\hat{\Lambda}_\mathrm{A} := \mathrm{argmin}_{\Lambda \succ 0} -\log\det(\Lambda) + \mathrm{tr}[\hat{\Sigma}_\mathrm{A}\Lambda] + \lambda_n^{(\mathrm{A})}\|\Lambda\|_{1,\mathrm{off}},$$
$$\hat{\Lambda}_\mathrm{B} := \mathrm{argmin}_{\Lambda \succ 0} -\log\det(\Lambda) + \mathrm{tr}[\hat{\Sigma}_\mathrm{B}\Lambda] + \lambda_n^{(\mathrm{B})}\|\Lambda\|_{1,\mathrm{off}},$$

where $\hat{\Sigma}_\mathrm{A} := \sum_{i=1}^{n_\mathrm{A}} x_i^{(\mathrm{A})} x_i^{(\mathrm{A})\top}/n_\mathrm{A}$, $\hat{\Sigma}_\mathrm{B} := \sum_{i=1}^{n_\mathrm{B}} x_i^{(\mathrm{B})} x_i^{(\mathrm{B})\top}/n_\mathrm{B}$, and $\|\Lambda\|_{1,\mathrm{off}} := \sum_{i\neq j}|\Lambda_{ij}|$.

Here we write $\Sigma_\mathrm{A} = \Lambda_\mathrm{A}^{-1}$, and $\Sigma_\mathrm{B} = \Lambda_\mathrm{B}^{-1}$. We define a linear operator $\Phi_\mathrm{A} : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ as $\Phi_\mathrm{A}\Lambda = \Sigma_\mathrm{A}\mathrm{tr}[\Sigma_\mathrm{A}\Lambda]$ (in other words, $\Phi_\mathrm{A} = \Sigma_\mathrm{A} \otimes \Sigma_\mathrm{A}$), and $\Phi_\mathrm{B} : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ in the same manner.

Now let $S_\mathrm{A} := \{(i,j) \mid \Lambda_{\mathrm{A},ij} \neq 0\} \cup \{(1,1),\ldots,(d,d)\}$ and similarly $S_\mathrm{B} := \{(i,j) \mid \Lambda_{\mathrm{B},ij} \neq 0\} \cup \{(1,1),\ldots,(d,d)\}$. Let $s_\mathrm{A} := \max_{i=1,\ldots,p}|\{j \mid \Lambda_{\mathrm{A},ij} \neq 0\}|$ and $s_\mathrm{B} := \max_{i=1,\ldots,p}|\{j \mid \Lambda_{\mathrm{B},ij} \neq 0\}|$. For matrix $M \in \mathbb{R}^{p\times p}$ ($p \in \mathbb{N}$), we denote by $M_{\mathcal{I}\times\mathcal{J}}$ the sub-matrix consisting of the components of M with indices in $\mathcal{I}\times\mathcal{J}$, and let $\|M\|_{\infty,\infty}$ be an operator norm of M as an operator from ℓ_∞ to ℓ_∞, that is, $\|M\|_{\infty,\infty} := \max_{i=1,\ldots,p}\sum_{j=1}^p |M_{ij}|$.

Using these notations, we define

$$K_{\Sigma_\mathrm{A}} := \|\Sigma_\mathrm{A}\|_{\infty,\infty},$$
$$K_{\Psi_\mathrm{A}} := \|(\Psi_{\mathrm{A},S\times S})^{-1}\|_{\infty,\infty} = \|((\Sigma_\mathrm{A}\otimes\Sigma_\mathrm{A})_{S\times S})^{-1}\|_{\infty,\infty},$$

and define K_{Σ_B} and K_{Ψ_B} in the same manner.

Finally, we assume the following conditions on the distributions p_A and p_B.

Assumption 2

1. *(Sub-Gaussianity) Both p_A and p_B have zero means and sub-Gaussian tails: $\exists\sigma > 0$, $\forall t \in \mathbb{R}$,*
$$\mathrm{E}_{p_\mathrm{A}}[\exp(tX)] \le \exp(\sigma^2 t^2),$$
$$\mathrm{E}_{p_\mathrm{B}}[\exp(tX)] \le \exp(\sigma^2 t^2).$$
2. *(Incoherence) There exists a real number $\alpha \in (0,1]$ such that*
$$\|\Psi_{\mathrm{A},S^c\times S}(\Psi_{\mathrm{A},S\times S})^{-1}\|_{\infty,\infty} \le 1-\alpha,$$
$$\|\Psi_{\mathrm{B},S^c\times S}(\Psi_{\mathrm{B},S\times S})^{-1}\|_{\infty,\infty} \le 1-\alpha.$$

Now let $n := \min\{n_\mathrm{A}, n_\mathrm{B}\}$ and we have the following lemma.

Lemma 1 *Under Assumption 2, if the sample size n satisfies*

$$n > C_3 s_\mathrm{A}^2(1+3/\alpha)^2(2\log(2d/\eta)+\log 4),$$
$$C_3 := \{48\sqrt{2}(1+4\sigma^2)\max_i(\Sigma_{\mathrm{A},ii})\max\{K_{\Sigma_\mathrm{A}}K_{\Psi_\mathrm{A}}, K_{\Sigma_\mathrm{A}}^3 K_{\Psi_\mathrm{A}}^2\}\}^2,$$

and the regularization parameter $\lambda_n^{(\mathrm{A})}$ satisfies

$$\lambda_n^{(\mathrm{A})} = (8/\alpha)\sqrt{128(1+4\sigma^2)^2\max_i(\Sigma_{\mathrm{A},ii})^2}\sqrt{\frac{2\log(2d/\eta)}{n}},$$

then with probability greater than $1-\eta/2$ ($\forall\eta > 0$), the estimated $\hat{\Lambda}_\mathrm{A}$ satisfies the bound,

$$|||\hat{\Lambda}_\mathrm{A} - \Lambda_\mathrm{A}|||_\infty \le 16\sqrt{2}(1+4\sigma^2)\max_i(\Sigma_{\mathrm{A},ii})(1+8\alpha^{-1}) \times K_{\Psi_\mathrm{A}}\sqrt{\frac{2\log(2d/\eta)+\log 4}{n}},$$

and the non-zero elements of $\hat{\Lambda}_A$ are included in Λ_A,

$$\{(i,j) \mid \hat{\Lambda}_{\mathrm{A},ij} \neq 0\} \subseteq S_\mathrm{A}.$$

(Proof of Lemma 1) The statement is a modification of Corollary 1 of Ravikumar et al. [10]. Assumption 2 yields the conditions assumed in Theorem 1 and Corollary 1 (sub-Gaussian condition) [10], thus we can apply these theorems. The assertion is immediately proven by replacing $\bar{\delta}_f(n,p^\tau)$ and $\bar{n}_f(n,p^\tau)$ with $\bar{\delta}_f(n,\eta/(2d^2))$ and $\bar{n}_f(n,\eta/(2d^2))$ in the proof of Corollary 1 in the paper. If we replace d^τ with $\eta/(2d^2)$, then the tail probability of the noise can be evaluated as

$$P[|||\hat{\Sigma}-\Sigma^*|||_\infty \ge \bar{\delta}_f(n,\eta/(2d^2))] \le \frac{\eta}{2},$$

by Lemma 8 and its proof [10]. Finally, we notice $\log(2d^2/\eta) \le 2\log(2d/\eta)$, and obtain the assertion. □

We have the same statement for the estimation of Λ_B.

Lemma 2 *Under the conditions assumed in Lemma 1, $\hat{\Gamma}-\Gamma$ satisfies the following inequality with probability greater than $1-\eta$ ($\eta \in (0,1)$):*

$$|||\hat{\Gamma}-\Gamma|||_\infty \le C_1\sqrt{\frac{2\log(2d/\eta)+\log 4}{n}},$$
$$C_1 := 16\sqrt{2}(1+4\sigma^2)(1+8\alpha^{-1}) \times (\max_i(\Sigma^*_{\mathrm{A},ii})K_{\Gamma^*_\mathrm{A}} + \max_i(\Sigma^*_{\mathrm{B},ii})K_{\Gamma^*_\mathrm{B}}).$$

In particular, we have

$$\|\hat{\Gamma}-\Gamma\|_{S_\infty} \le \|\hat{\Gamma}-\Gamma\|_F \le C_1\sqrt{\frac{d\min\{s_{\rm A}+s_{\rm B},d\}\{2\log(2d/\eta)+\log 4\}}{n}},$$

with probability greater than $1-\eta$, where $\|\cdot\|_{S_\infty}$ is the spectrum norm and $\|\cdot\|_F$ is the Frobenius norm.

(Proof of Lemma 2) Since

$$\begin{aligned}&-|\hat{\Lambda}_{{\rm A},ij}-\Lambda_{{\rm A},ij}-\hat{\Lambda}_{{\rm B},ij}+\Lambda_{{\rm B},ij}|\\ &\le |\hat{\Lambda}_{{\rm A},ij}-\hat{\Lambda}_{{\rm B},ij}|-|\Lambda_{{\rm A},ij}-\Lambda_{{\rm B},ij}|\\ &\le |\hat{\Lambda}_{{\rm A},ij}-\Lambda_{{\rm A},ij}-\hat{\Lambda}_{{\rm B},ij}+\Lambda_{{\rm B},ij}|,\end{aligned}$$

we obtain

$$\begin{aligned}\|\!|\hat{\Gamma}-\Gamma|\!\|_\infty &\le \|\!|\hat{\Lambda}_{\rm A}-\Lambda_{\rm A}-\hat{\Lambda}_{\rm B}+\Lambda_{\rm B}|\!\|_\infty\\ &\le \|\!|\hat{\Lambda}_{\rm A}-\Lambda_{\rm A}|\!\|_\infty+\|\!|\hat{\Lambda}_{\rm B}-\Lambda_{\rm B}|\!\|_\infty.\end{aligned}$$

Thus, applying the bound derived in Lemma 1 to both $\|\!|\hat{\Lambda}_{\rm A}-\Lambda_{\rm A}|\!\|_\infty$ and $\|\!|\hat{\Lambda}_{\rm B}-\Lambda_{\rm B}|\!\|_\infty$, we obtain the first assertion.

As for the second assertion, since $\hat{\Gamma}-\Gamma$ has at most $\min\{d(s_{\rm A}+s_{\rm B}),d^2\}$ non-zero components by the second assertion of Lemma 1, the Frobenius norm between $\hat{\Gamma}-\Gamma$ has the bound,

$$\begin{aligned}\|\hat{\Gamma}-\Gamma\|_F &\le \sqrt{\min\{d(s_{\rm A}+s_{\rm B}),d^2\}\|\!|\hat{\Gamma}-\Gamma|\!\|_\infty^2}\\ &= \sqrt{\min\{d(s_{\rm A}+s_{\rm B}),d^2\}}\|\!|\hat{\Gamma}-\Gamma|\!\|_\infty,\end{aligned}$$

which gives the assertion. □

This lemma gives a bound on the discrepancy between $\lambda_{\min}(\hat{\Gamma})$ and $\lambda_{\min}(\Gamma)$ as follows.

Lemma 3 *Under the conditions assumed in Lemma 1, we have the following bound with probability greater than $1-\eta$,*

$$|\lambda_{\min}(\hat{\Gamma})-\lambda_{\min}(\Gamma)| \le C_1\sqrt{\frac{d\min\{s_{\rm A}+s_{\rm B},d\}\{2\log(2d/\eta)+\log 4\}}{n}}.$$

(Proof of Lemma 3) For a real symmetric matrix $Q\in\mathbb{R}^{d\times d}$, let $\lambda_i(Q)$ be the i-th smallest eigenvalue ($\lambda_1(Q)\le\lambda_2(Q)\le\cdots\le\lambda_d(Q)$). For all real symmetric matrices $Q,R\in\mathbb{R}^{d\times d}$, the well-known Hoffman and Wielandt inequality (see Theorem 6.3.5 and its corollaries in Horn et al. [25]) yields

$$\sqrt{\sum_{i=1}^d(\lambda_i(Q)-\lambda_i(R))^2}\le\|Q-R\|_F.$$

Then applying Lemma 2 to the right-hand side of this inequality and noticing the relation $|\lambda_{\min}(Q)-\lambda_{\min}(R)|\le\sqrt{\sum_{i=1}^d(\lambda_i(Q)-\lambda_i(R))^2}$, we obtain the assertion. □

7.2 Proof of Theorem 1

The theorem is true for $k=d$ and $k=0$ since $\mathcal{I}=\hat{\mathcal{I}}=\{1,2,\ldots,d\}$ and $\mathcal{I}=\hat{\mathcal{I}}=\emptyset$ hold, respectively. Therefore, we only need to consider the case when $1\le k\le d-1$.

Let $\epsilon:=\|\!|\hat{\Gamma}-\Gamma|\!\|_\infty$ and $f(\mathcal{K},\mathcal{K}';M):=\sum_{i\in\mathcal{K},j\in\mathcal{K}'}M_{ij}$ for a matrix M. We also set the index sets $\mathcal{P}$, $\mathcal{Q}$, and $\mathcal{R}$ as $\mathcal{P}:=\mathcal{I}\backslash\hat{\mathcal{I}}$, $\mathcal{Q}:=\hat{\mathcal{I}}\backslash\mathcal{I}$, and $\mathcal{R}:=\mathcal{I}\cap\hat{\mathcal{I}}$. We now have

$$\begin{aligned}&f(\mathcal{I},\mathcal{I};\hat{\Gamma})-f(\hat{\mathcal{I}},\hat{\mathcal{I}};\hat{\Gamma})\\ &\quad= f(\mathcal{P},\mathcal{P};\hat{\Gamma})+2f(\mathcal{P},\mathcal{R};\hat{\Gamma})-f(\mathcal{Q},\mathcal{Q};\hat{\Gamma})-2f(\mathcal{Q},\mathcal{R};\hat{\Gamma})\\ &\quad\le f(\mathcal{P},\mathcal{P};\Gamma)+2f(\mathcal{P},\mathcal{R};\Gamma)-f(\mathcal{Q},\mathcal{Q};\Gamma)-2f(\mathcal{Q},\mathcal{R};\Gamma)\\ &\qquad+\epsilon(|\mathcal{P}|^2+|\mathcal{Q}|^2+2|\mathcal{P}||\mathcal{R}|+2|\mathcal{Q}||\mathcal{R}|)\\ &\quad\le f(\mathcal{I},\mathcal{I};\Gamma)-f(\hat{\mathcal{I}},\hat{\mathcal{I}};\Gamma)+\epsilon(k^2+d^2),\end{aligned}$$

where, in the first inequality, we used the fact that $|f(\mathcal{K},\mathcal{K}';\hat{\Gamma})-f(\mathcal{K},\mathcal{K}';\Gamma)|\le\epsilon|\mathcal{K}||\mathcal{K}'|$, and in the second inequality, we used $|\mathcal{P}|^2+|\mathcal{Q}|^2+2|\mathcal{P}||\mathcal{R}|+2|\mathcal{Q}||\mathcal{R}|\le(|\mathcal{P}|+|\mathcal{R}|)^2+(|\mathcal{Q}|+|\mathcal{R}|)^2=|\mathcal{I}|^2+|\hat{\mathcal{I}}|^2\le k^2+d^2$. Since this inequality is valid for all $\hat{\mathcal{I}}\neq\mathcal{I}$, we have

$$f(\mathcal{I},\mathcal{I};\hat{\Gamma})-f(\hat{\mathcal{I}},\hat{\mathcal{I}};\hat{\Gamma})\le\epsilon(k^2+d^2)-h.$$

From the assumption that $\mathcal{I}$ is unique, we have $h>0$. If $\epsilon<h/(k^2+d^2)$, then the right hand side becomes negative implying $\mathcal{I}$ is the minimizer of (1), which proves the claim. □

7.3 Proof of Theorem 2

The proof of the theorem immediately follows from the next two lemmas:

Lemma 4 *Let $\hat{A}_{\mu,\mathcal{I}\times\mathcal{I}}\in\mathbb{R}^{k\times k}$ and $\hat{A}_{\mu,\mathcal{J}\times\mathcal{I}}\in\mathbb{R}^{(d-k)\times k}$ be sub-matrices of $\hat{A}_\mu$ indexed by $\mathcal{I}$ and $\mathcal{J}$. Suppose the following conditions hold for some $\tau,\tau'>0$:*

$$\hat{A}^{-1}_{\mu,\mathcal{I}\times\mathcal{I}}\mathbf{1}_k\ge\tau\mathbf{1}_k, \tag{5}$$

$$\hat{A}_{\mu,\mathcal{J}\times\mathcal{I}}\mathbf{1}_k\ge\frac{1+\tau'}{\tau}\mathbf{1}_{d-k}. \tag{6}$$

Then $\tilde{\mathcal{I}}_0=\mathcal{I}$.

Lemma 5 *Suppose the conditions in Theorem 2 hold true. Then there exists $\tau,\tau'>0$ that satisfy the conditions (5) and (6).*

(Proof of Lemma 4) Since the problem (3) is a convex quadratic programming with a positive-definite matrix $\hat{A}_\mu$, the KKT conditions given here are both necessary

and sufficient for the optimality of the solution:

$$\hat{A}_\mu \tilde{\boldsymbol{s}} - \tilde{\boldsymbol{\zeta}} - \tilde{\nu}\mathbf{1}_d = \mathbf{0}_d, \tag{7}$$

$$\tilde{\boldsymbol{s}}, \tilde{\boldsymbol{\zeta}} \geq \mathbf{0}_d, \tag{8}$$

$$\tilde{\boldsymbol{s}} \odot \tilde{\boldsymbol{\zeta}} = \mathbf{0}_d, \tag{9}$$

$$\mathbf{1}_d^\top \tilde{\boldsymbol{s}} - 1 = 0, \tag{10}$$

where $\tilde{\boldsymbol{\zeta}}$ and $\tilde{\nu}$ are the dual parameters, and $\odot$ denotes the Hadamard product.

Here, let $\boldsymbol{t}_\mathcal{I}$, $\boldsymbol{t}_\mathcal{J}$, $\boldsymbol{\xi}_\mathcal{I}$, and $\boldsymbol{\xi}_\mathcal{J}$ be sub-vectors of $\boldsymbol{t}, \boldsymbol{\xi} \in \mathbb{R}^d$ indexed by $\mathcal{I}$ and $\mathcal{J}$. We define vectors $\boldsymbol{t}$ and $\boldsymbol{\xi}$ by $\boldsymbol{t}_\mathcal{I} = \hat{A}^{-1}_{\mu,\mathcal{I}\times\mathcal{I}}\mathbf{1}_k$, $\boldsymbol{t}_\mathcal{J} = \mathbf{0}_{d-k}$, $\boldsymbol{\xi}_\mathcal{I} = \mathbf{0}_k$, $\boldsymbol{\xi}_\mathcal{J} = \hat{A}_{\mu,\mathcal{J}\times\mathcal{I}}\hat{A}^{-1}_{\mu,\mathcal{I}\times\mathcal{I}}\mathbf{1}_k - \mathbf{1}_{d-k}$. We also define a scalar γ by $\gamma = 1/\mathbf{1}_k^\top \hat{A}^{-1}_{\mu,\mathcal{I}\times\mathcal{I}}\mathbf{1}_k$. We show that setting $\tilde{\boldsymbol{s}} \leftarrow \gamma\boldsymbol{t}$, $\tilde{\boldsymbol{\zeta}} \leftarrow \gamma\boldsymbol{\xi}$, and $\tilde{\nu} \leftarrow \gamma$ satisfies the conditions (7)–(10). From the definitions of $\boldsymbol{t}, \boldsymbol{\xi}$, and γ, the conditions (7), (9), and (10) are obvious, and there remains only the condition (8) to be verified. The condition $\tilde{\boldsymbol{s}} \geq \mathbf{0}_d$ is guaranteed from the condition (5). By combining (5) and (6), we also have

$$\boldsymbol{\xi}_\mathcal{J} \geq \tau\hat{A}_{\mu,\mathcal{J}\times\mathcal{I}}\mathbf{1}_k - \mathbf{1}_{d-k} \geq \tau'\mathbf{1}_{d-k} > \mathbf{0}_{d-k},$$

which guarantees $\tilde{\boldsymbol{\zeta}} \geq \mathbf{0}_d$. This result indicates the vector $\tilde{\boldsymbol{s}}$ defined here is the optimal solution to the problem (3) with $\tilde{\mathcal{I}}_0 = \mathcal{I}$, which completes the proof. □

(Proof of Lemma 5) Let $\epsilon := |||\hat{A}_\mu - A_\mu|||_\infty$. Since $B_\delta \leq 1/2$ for any $\delta > 0$, we have $\epsilon < \mu/2d$.

We first consider the condition (5). We note that, from Assumption 1, $\Gamma_{\mathcal{I}\times\mathcal{I}} = 0_{|\mathcal{I}|\times|\mathcal{I}|}$ holds and thus $A_{\mu,\mathcal{I}\times\mathcal{I}} = \mu I_k$. Hence, $\hat{A}_{\mu,\mathcal{I}\times\mathcal{I}}$ can be expressed as $\hat{A}_{\mu,\mathcal{I}\times\mathcal{I}} = \mu I_k + E$ with $|||E|||_\infty \leq \epsilon$. From Ravikumar et al. [10, pp. 972], we have

$$(\mu I_k + E)^{-1}\mathbf{1}_k = \frac{1}{\mu}\mathbf{1}_k - \frac{1}{\mu^2}E\mathbf{1}_k + \frac{1}{\mu^3}E^2J\mathbf{1}_k,$$

where $J = \sum_{m=0}^\infty (-1)^m(\mu^{-1}E)^m$. From $|||E|||_\infty \leq \epsilon$, we have $\|E^2\|_{\infty,\infty} \leq d^2\epsilon^2$ and $\|J\|_{\infty,\infty} \leq \sum_{m=0}^\infty \|\mu^{-1}E\|^m_{\infty,\infty} \leq \sum_{m=0}^\infty(\mu^{-1}d\epsilon)^m \leq 2$ where we used $\epsilon < \mu/2d$ for the last inequality. Hence, we have

$$|||E^2J\mathbf{1}_k|||_\infty \leq \|E^2J\|_{\infty,\infty} \leq \|E^2\|_{\infty,\infty}\|J\|_{\infty,\infty} \leq 2d^2\epsilon^2.$$

Thus, we can conclude

$$(\hat{A}^{-1}_{\mu,\mathcal{I}\times\mathcal{I}}\mathbf{1}_k)_i \geq \frac{1}{\mu} - \frac{d}{\mu^2}\epsilon - \frac{2d^2}{\mu^3}\epsilon^2,$$

where $(\cdot)_i$ denotes the ith entry of the vector in the parenthesis. In addition, the right hand side of this inequality becomes positive when $\epsilon < \mu/2d$, which is assured by the assumption. This indicates that we can choose $\tau = 1/\mu - d\epsilon/\mu^2 - 2d^2\epsilon^2/\mu^3$ and the condition (5) holds.

We now turn to the condition (6). We have

$$(\hat{A}_{\mu,\mathcal{J}\times\mathcal{I}}\mathbf{1}_k)_i \geq (A_{\mu,\mathcal{J}\times\mathcal{I}}\mathbf{1}_k)_i - k\epsilon \geq (1+\delta)\mu - k\epsilon,$$

from the assumption in Theorem 2, and we can choose τ' as $\tau' = \tau\{(1+\delta)\mu - k\epsilon\} - 1$. The condition $\tau' > 0$ is assured when

$$\begin{aligned}\tau' &= \delta - \frac{k}{\mu}\epsilon - \left(\frac{d}{\mu}\epsilon + 2\frac{d^2}{\mu^2}\epsilon^2\right)\left(1 + \delta - \frac{k}{\mu}\epsilon\right) \\ &\geq \delta - (2+\delta)\frac{d}{\mu}\epsilon - 2(1+\delta)\frac{d^2}{\mu^2}\epsilon^2 > 0,\end{aligned}$$

which is guaranteed by the assumption $\epsilon < B_\delta\mu/d$, and therefore the condition (6) holds. □

7.4 Proof of Theorem 3

From Lemma 2, we have $|||\hat{\Gamma} - \Gamma|||_\infty < h/(k^2+d^2)$ when $\eta > 4d\exp\{-h^2n/2C_1^2(k^2+d^2)^2\}$, which proves the claim. □

7.5 Proof of Theorem 4

From Lemmas 2 and 3, with probability greater than $1-\eta$, we have

$$\begin{aligned}|||\hat{A}_\mu - A_\mu|||_\infty &\leq |||\hat{\Gamma} - \Gamma|||_\infty + |\lambda_{\min}(\hat{\Gamma}) - \lambda_{\min}(\Gamma)| \\ &\leq C_1\left(1 + \sqrt{d\min\{s_\mathrm{A} + s_\mathrm{B}, d\}}\right) \\ &\quad \times \sqrt{\frac{2\log(2d/\eta) + \log 4}{n}} \\ &= C_2\sqrt{\frac{2\log(2d/\eta) + \log 4}{n}},\end{aligned}$$

where $C_2 := C_1\left(1 + \sqrt{d\min\{s_\mathrm{A} + s_\mathrm{B}, d\}}\right)$.

From Lemma 2, we have $|||\hat{A}_\mu - A_\mu|||_\infty < B_\delta\mu/d$ when $\eta > 4d\exp\{-B_\delta^2\mu^2n/2C_2^2d^2\}$, which proves the claim. □

References

[24] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning* 3(1):1122, 2011.

[25] R. A. Horn, and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, New York, 1985.