Trans-Glasso: A Transfer Learning Approach to Precision Matrix Estimation

Boxin Zhao[1], Cong Ma[2], and Mladen Kolar[3]

[1]Booth School of Business, University of Chicago
[2]Department of Statistics, University of Chicago
[3]Department of Data Sciences and Operations, Marshall School of Business, University of Southern California

Abstract

Precision matrix estimation is essential in various fields, yet it is challenging when samples for the target study are limited. Transfer learning can enhance estimation accuracy by leveraging data from related source studies. We propose Trans-Glasso, a two-step transfer learning method for precision matrix estimation. First, we obtain initial estimators using a multi-task learning objective that captures shared and unique features across studies. Then, we refine these estimators through differential network estimation to adjust for structural differences between the target and source precision matrices. Under the assumption that most entries of the target precision matrix are shared with source matrices, we derive non-asymptotic error bounds and show that Trans-Glasso achieves minimax optimality under certain conditions. Extensive simulations demonstrate Trans-Glasso 's superior performance compared to baseline methods, particularly in small-sample settings. We further validate Trans-Glasso in applications to gene networks across brain tissues and protein networks for various cancer subtypes, showcasing its effectiveness in biological contexts. Additionally, we derive the minimax optimal rate for differential network estimation, representing the first such guarantee in this area.

1 Introduction

Estimating the precision matrix, i.e., the inverse covariance matrix, is a fundamental task in statistical analysis and has broad applications, including in portfolio optimization, speech recognition, and genomics [Best and Grauer, 1992, Lauritzen, 1996, Yuan and Lin, 2007, Saon and Chien, 2011]. The precision matrix is closely tied to Gaussian graphical models: estimating the support of the precision matrix corresponds to uncovering the network structure of conditional dependencies between multivariate normal variables [Lauritzen, 1996]. However, estimating a precision matrix accurately is often challenging when the sample size is small compared to the dimension—a typical scenario in high-dimensional settings.

In many applications, sample sizes are constrained for the target study of interest, yet data from related studies may be available. Transfer learning [Pan and Yang, 2009] provides a promising approach in these scenarios by leveraging information from related source studies to improve estimation accuracy in the target study. For example, in gene expression studies across different tissue types, sample sizes may be small for specific tissues, but data from related tissues can help improve estimates [Li et al., 2023a]. Similarly, protein network studies for different cancer subtypes

can benefit from transfer learning, as leveraging data from related subtypes can enhance estimation for a particular subtype with limited data [Peterson et al., 2015].

A critical aspect of transfer learning is establishing similarity between the target and source tasks. Here, we assume that most entries of the target precision matrix are shared with those of the source matrices, with only a few differences. Based on this assumption, we propose Trans-Glasso, a novel two-step transfer learning method for precision matrix estimation. First, we obtain initial estimators through a multi-task learning objective that captures shared and unique dependencies across datasets. Second, we refine these estimators using differential network estimation to adjust for differences between the target and source matrices [Zhao et al., 2014, Yuan et al., 2017].

We provide a theoretical analysis of Trans-Glasso, deriving non-asymptotic error bounds and establishing that the method achieves minimax optimality in a wide range of parameter regimes. Through extensive simulations, we demonstrate that Trans-Glasso outperforms several baseline methods, particularly in scenarios where the target sample size is small. We also apply Trans-Glasso to gene networks across brain tissues and protein networks for various cancer subtypes, showing its practical effectiveness in biological applications. Additionally, as a byproduct of our analysis, we derive the minimax optimal rate for differential network estimation, to our knowledge, the first of its kind.

1.1 Related Work

Precision matrix estimation. Estimation of sparse precision matrices in a single study is well studied. Common methods include penalized M-estimator [Yuan and Lin, 2007, Friedman et al., 2008, Rothman et al., 2008, Lam and Fan, 2009, Ravikumar et al., 2011] and constrained L_1 minimization [Cai et al., 2011, Ren et al., 2015, Cai et al., 2016b]. There is also extensive literature on multi-task precision matrix estimation, which estimates multiple related but nonidentical precision matrices from multiple studies [Guo et al., 2011, Danaher et al., 2014, Zhu et al., 2014, Mohan et al., 2014, Lee and Liu, 2015, Cai et al., 2016a, Ma and Michailidis, 2016, Saegusa and Shojaie, 2016]. See Tsai et al. [2022] for a survey. While related to transfer learning, multi-task learning aims to estimate parameters of all studies, whereas transfer learning only focuses on the target study.

Transfer learning. Transfer learning has a long history [Pan and Yang, 2009] and has been applied in various contexts [Turki et al., 2017, Hajiramezanali et al., 2018, Bastani, 2021]. Recently, interest in transfer learning for statistical problems has grown. Li et al. [2022], He et al. [2024] studied high-dimensional linear regression. The fused regularizer in He et al. [2024] is similar to our multi-task objective; however, He et al. [2024] focuses on linear regression, while we focus on precision matrix estimation, making the techniques different. Li et al. [2023b], Tian and Feng [2023] studied high-dimensional generalized linear regression. Pathak et al. [2022], Ma et al. [2023], Wang [2023], Ge et al. [2024] addressed covariate shift. Cai and Wei [2021] studied nonparametric classification and Cai et al. [2024a] investigated multi-armed bandit problems. Liu [2023] proposed a unified transfer learning model for high-dimensional linear regression problems. Hanneke and Kpotufe [2019, 2022] studied the transfer learning problem from a learning theory perspective. Lin and Reimherr [2022], Cai et al. [2024b] studied transfer learning for functional data analysis.

The most relevant work to this paper is Li et al. [2023a], which also studied transfer learning for precision matrix estimation. The key difference is the similarity assumption. Li et al. [2023a] assumes that the divergence matrices between the target and source precision matrices are sparse. We assume that most entries of the target precision matrix are shared across source precision matrices, with few different entries. Although the assumption in Li et al. [2023a] is motivated by the KL divergence between Gaussian distributions, ours is a structural assumption, making it applicable beyond Gaussian data and easier to interpret. Consequently, our method differs significantly from that of Li et al. [2023a].

Differential Network Estimation. Our approach leverages differential network estimation techniques, which aim to directly estimate the difference between two precision matrices without the need to estimate the individual ones [Zhao et al., 2014, Yuan et al., 2017, Liu et al., 2014, Ma et al., 2021]. Fazayeli and Banerjee [2016] explored this concept in the context of Ising models. Additionally, Zhao et al. [2019, 2022] extended differential network estimation methods to functional data, while Tugnait [2023] broadened its application to multi-attribute data.

1.2 Organization and Notation

The rest of the paper is organized as following. In Section 2, we introduce the problem setup. In Section 3, we introduce the methodology of the paper. We then describe how to implement our method in practice in Section 4. The theoretical results are developed in Section 5. Besides, we implement extensive simulation experiments in Section 6. Furthermore, in Section 7, we apply our method on two real-world datasets. Finally, we conclude our paper with Section 8. The technical proofs and details about optimization algorithms are provided in the appendix.

Notation. For a vector $v \in \mathbb{R}^d$, we use $\|v\|_p$ to denote its L_p-norm. More specifically, we have $\|v\|_p = (\sum_{i=1}^d |v_i|^p)^{\frac{1}{p}}$ for $1 \leqslant p \leqslant \infty$, where $\|v\|_\infty = \max_i |v_i|$. For a matrix $A \in \mathbb{R}^{d\times d}$, we use $|\cdot|$ to denote its elementwise norm and $\|\cdot\|$ to denote its operator norm. For example, $|A|_1 = \sum_{i=1}^d \sum_{j=1}^d |A_{ij}|$, $|A|_0 = \sum_{i=1}^d \sum_{j=1}^d \mathbb{1}\{A_{ij} \neq 0\}$, $|A|_\infty = \max_{1\leqslant i,j\leqslant d} |A_{ij}|$; $\|A\|_1 = \max_{1\leqslant j\leqslant d} \sum_{i=1}^d |A_{ij}|$, $\|A\|_\infty = \max_{1\leqslant i\leqslant d} \sum_{j=1}^d |A_{ij}|$, and $\|A\|_2$ denote the largest singular value of A. We use $\|A\|_{\mathrm{F}} = (\sum_{i=1}^d \sum_{j=1}^d |A_{ij}|^2)^{1/2}$ to denote the Frobenius norm of A. In addition, we use $\langle A, B\rangle = \mathrm{tr}(A^\top B) = \sum_{i,j} A_{ij}B_{ij}$ for $A, B \in \mathbb{R}^{d\times d}$ to define the inner product between two matrices. We use $\mathrm{vec}(A)$ to denote the d^2-vector obtained by stacking the columns of A. When A is symmetric, we let $\gamma_{\min}(A)$ and $\gamma_{\max}(A)$ denote its smallest and largest eigenvalue. For $A \in \mathbb{R}^{n_1\times n_2}$ and $B \in \mathbb{R}^{m_1\times m_2}$, we let $A \otimes B = [A_{ij}B_{lm}]_{i,j,l,m} \in \mathbb{R}^{n_1 m_1 \times n_2 m_2}$ denote the Kronecker product of two matrices. We define $\mathbb{S}^{d\times d}$ as the set of symmetric matrices with dimension d. The universal constants may vary from one line to another without further clarification. Finally, following Ravikumar et al. [2011], we say a random vector $X \in \mathbb{R}^d$ with $\mathbb{E}[X] = 0$ is sub-Gaussian if there exists a constant $\sigma > 0$ such that

$$\mathbb{E}\left[\exp\left(\lambda X_j/\sqrt{\Sigma_{jj}}\right)\right] \leqslant \exp\left(\sigma^2\lambda^2/2\right) \quad \text{for all } \lambda \in \mathbb{R} \text{ and } 1 \leqslant j \leqslant d,$$

where $\Sigma = \mathrm{Cov}(X)$.

In addition, we use the following standard notation in the paper. For two positive sequences $\{f(n)\}_{n\geqslant 1}$ and $\{g(n)\}_{n\geqslant 1}$, $f(n) = O(g(n))$ or $f(n) \lesssim g(n)$ means that there exists a universal constant $c > 0$ such that $f(n) \leqslant cg(n)$ holds for sufficiently large n; $f(n) = \Omega(g(n))$ or $f(n) \gtrsim g(n)$ means that there exists a universal constant $c > 0$ such that $f(n) \geqslant cg(n)$ holds for sufficiently large n; $f(n) = \Theta(g(n))$ or $f(n) \asymp g(n)$ means that there exist universal constants $c_1, c_2 > 0$ such that $c_1 g(n) \leqslant f(n) \leqslant c_2 g(n)$ holds for sufficiently large n; $f(n) = o(g(n))$ indicates that $f(n)/g(n) \to 0$ as $n \to \infty$.

2 Problem Setup

Imagine that we observe n_0 i.i.d. samples $\{\mathbf{x}_i^{(0)}\}_{i=1}^{n_0} := \mathcal{D}_0$ from a sub-Gaussian target distribution $\mathcal{P}_0$. Each sample $\mathbf{x}_i^{(0)} \in \mathbb{R}^d$ is assumed to have zero mean and covariance matrix $\Sigma^{(0)}$. Our goal is to estimate the target precision matrix $\Omega^{(0)} = \left\{\Sigma^{(0)}\right\}^{-1}$. Additionally, we have access to K sub-Gaussian source distributions $\{\mathcal{P}_k\}_{k=1}^K$, each with n_k i.i.d. samples $\{\mathbf{x}_i^{(k)}\}_{i=1}^{n_k} := \mathcal{D}_k$. For $1 \leqslant k \leqslant K$, each $\mathbf{x}_i^{(k)} \in \mathbb{R}^d$ also has zero mean, covariance matrix $\Sigma^{(k)}$, and corresponding precision matrix

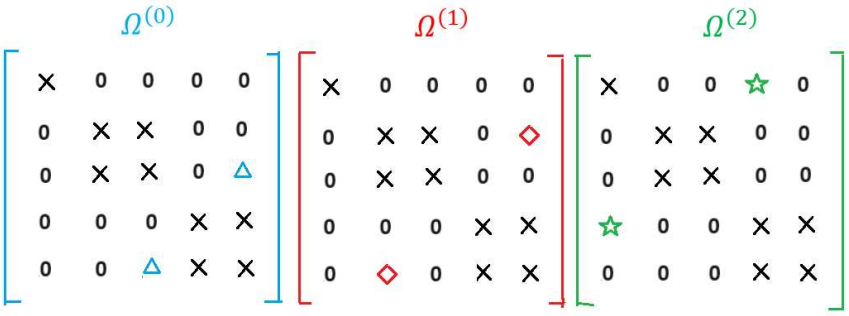

Figure 1: Illustration of Assumption 1. The target precision matrix, $\Omega^{(0)}$, is shown alongside two source precision matrices, $\Omega^{(1)}$ and $\Omega^{(2)}$. Black crosses represent the shared entries across the matrices, while colored shapes indicate individual, unique entries.

$\Omega^{(k)} = \left\{\Sigma^{(k)}\right\}^{-1}$. The goal is to leverage samples from both the target and source distributions to accurately estimate $\Omega^{(0)}$.

To facilitate transfer learning, we assume structural similarity between the target and source precision matrices, whereby most entries in $\Omega^{(0)}$ are shared with those in $\Omega^{(k)}$, with relatively few differences. This assumption enables us to efficiently utilize source samples to enhance estimation of the target precision matrix.

Formally, we characterize the relationship between the target and source precision matrices using the following assumption.

Assumption 1. *For each $0 \leqslant k \leqslant K$, there exists a shared component $\Omega^\star$ and a unique component $\Gamma^{(k)\star}$ with disjoint supports such that*

$$\Omega^{(k)} = \Omega^\star + \Gamma^{(k)\star}, \tag{1}$$

where $|\Omega^\star|_0 \leqslant s$, $|\Gamma^{(k)\star}|_0 \leqslant h$, and $|\Gamma^{(k)\star}|_1 \leqslant M_\Gamma$. The sparsity parameters s and h satisfy $s \gg h$, indicating that the majority of the structure is shared, while unique components are minimal.

See Figure 1 for a visual illustration of Assumption 1.

Assumption 1 is inspired from the assumptions widely used in the differential network estimation literature [Zhao et al., 2014, Yuan et al., 2017, Zhao et al., 2022]. To see this connection, for each $1 \leqslant k \leqslant K$, define $\Psi^{(k)} = \Omega^{(k)} - \Omega^{(0)}$ to be the differential network between the target $\Omega^{(0)}$ and the source $\Omega^{(k)}$. Two immediate implications of Assumption 1 are

$$\left|\Psi^{(k)}\right|_0 \leqslant 2h, \qquad \text{and} \qquad \left|\Psi^{(k)}\right|_1 \leqslant 2M_\Gamma, \quad \text{for all } 1 \leqslant k \leqslant K.$$

This matches exactly Condition 1 in the paper Zhao et al. [2014].

In addition, Assumption 1 is naturally interpretable within Gaussian graphical models. Suppose we have an undirected graphical model $G = (V, E)$ where nodes represent variables and edges represent conditional dependencies. In this model, an edge exists between nodes i and j if and only if $\Omega_{ij}^{(k)} \neq 0$ [Lauritzen, 1996]. Under Assumption 1, the target precision matrix $\Omega^{(0)}$ and the source matrices $\left\{\Omega^{(k)}\right\}$ share a large subset of edges, with only a small number of unique edges in each source, corresponding to sparse deviations $\Psi^{(k)}$.

Last but not least, we contrast Assumption 1 with the similarity assumptions based on divergence measures, used in Li et al. [2023a]. Li et al. [2023a] assumes that the divergence matrix

$$\Upsilon^{(k)} = \Omega^{(0)}\Sigma^{(k)} - I_d \tag{2}$$

is sparse. However, our structural approach offers broader applicability beyond Gaussian data, as it does not rely on assumptions specific to Gaussian graphical models and provides a straightforward interpretation of shared structure. See a more detailed comparison in Appendix A.

Next, we discuss estimating $\Omega^{(0)}$ using samples from both target and source distributions under Assumption 1.

3 Trans-Glasso Algorithm

In this section, we introduce Trans-Glasso, our transfer learning method for precision matrix estimation, which consists of two main steps. First, we initialize estimators of the precision matrices by solving a multi-task learning problem. Second, we refine these estimators using differential network estimation to adjust for structural differences between the target and source matrices.

3.1 Initialization via Multi-Task Learning

To leverage shared structure across the target and source matrices, we begin by jointly estimating precision matrices for both the target and source distributions. Based on Assumption 1, we employ a multi-task variant of the graphical lasso estimator [Friedman et al., 2008], which we refer to as Trans-MT-Glasso (**Trans**fer **M**ulti-**T**ask **G**raphical **lasso**).

Let $\widehat{\Sigma}^{(k)} = \frac{1}{n_k}\sum_{i=1}^{n_k} \mathbf{x}_i^{(k)}\mathbf{x}_i^{(k)\top}$ denote the sample covariance matrix for $0 \leqslant k \leqslant K$. We define $\Theta := \left(\Omega, \left\{\Gamma^{(k)}\right\}_{k=0}^{K}\right)$ for the shared component and the sparse unique components. The Trans-MT-Glasso objective is then given by

$$\widehat{\Theta} = \left(\widehat{\Omega}, \{\widehat{\Gamma}^{(k)}\}_{k=0}^{K}\right) \in \arg\min_{\Theta \in \mathcal{C}(M_{\text{op}})} \left\{\mathcal{L}(\Theta) + \lambda_{\text{M}}\Phi(\Theta)\right\}, \tag{3}$$

where

$$\mathcal{L}(\Theta) := \sum_{k=0}^{K} \alpha_k \left\{\left\langle \Omega + \Gamma^{(k)}, \widehat{\Sigma}^{(k)} \right\rangle - \log\det\left(\Omega + \Gamma^{(k)}\right)\right\}, \tag{4}$$

$$\Phi(\Theta) := |\Omega|_1 + \sum_{k=0}^{K} \sqrt{\alpha_k}\left|\Gamma^{(k)}\right|_1, \tag{5}$$

$$\mathcal{C}(M_{\text{op}}) := \left\{\Theta = \left(\Omega, \left\{\Gamma^{(k)}\right\}_{k=0}^{K}\right) : \Omega + \Gamma^{(k)} \succ 0 \text{ and } \left\|\Omega + \Gamma^{(k)}\right\|_2 \leqslant M_{\text{op}} \text{ for all } 0 \leqslant k \leqslant K\right\}. \tag{6}$$

Here, $\alpha_k = n_k/N$ with $N = \sum_{k=0}^{K} n_k$ controls the contribution of each source, $\lambda_{\text{M}} > 0$ is a regularization parameter, and $M_{\text{op}} > 0$ is a predefined constant. The constraint $\left\|\Omega + \Gamma^{(k)}\right\|_2 \leqslant M_{\text{op}}$ is primarily included to facilitate theoretical analysis. While we theoretically need to set M_{op} to be a sufficiently large constant, as detailed in a later section, in practice, we simply set $M_{\text{op}} = \infty$ to effectively remove this constraint.

The first term of (3) measures parameter fitness with observed data, while the second term $\Phi(\Theta)$promotes sparsity in both the shared and individual components. The sparsity penalization level for $\Gamma^{(k)}$ is proportional to $\sqrt{n_k/N}$, a factor that is crucial to balance contributions from the target and sources. See Section 5.1 for details.

In the end, we construct the initial estimators of $\Omega^{(k)}$ as $\breve{\Omega}^{(k)} = \widehat{\Omega} + \widehat{\Gamma}^{(k)}$ for $0 \leqslant k \leqslant K$.

3.2 Refinement via Differential Network Estimation

To further enhance accuracy, we refine these initial estimators by estimating the differential networks $\Psi^{(k)} = \Omega^{(k)} - \Omega^{(0)}$, which capture the structural differences between each source and the target. This refinement step corrects for potential biases in the initial estimates.

Estimating $\Psi^{(k)}$ has been extensively studied in the differential network estimation literature [Zhao et al., 2014, Yuan et al., 2017], with a variety of good estimators. For instance, Yuan et al. [2017] proposed estimation by solving

$$\widehat{\Psi}^{(k)} \in \arg\min_{\Psi} L_D\left(\Psi; \widehat{\Sigma}^{(0)}, \widehat{\Sigma}^{(k)}\right) + \lambda_{\Psi}^{(k)} |\Psi|_1 , \tag{7}$$

where

$$L_D\left(\Psi; \widehat{\Sigma}^{(0)}, \widehat{\Sigma}^{(k)}\right) = \frac{1}{4}\left(\left\langle \widehat{\Sigma}^{(0)}\Psi, \Psi\widehat{\Sigma}^{(k)} \right\rangle + \left\langle \widehat{\Sigma}^{(k)}\Psi, \Psi\widehat{\Sigma}^{(0)} \right\rangle\right) - \left\langle \Psi, \widehat{\Sigma}^{(0)} - \widehat{\Sigma}^{(k)} \right\rangle,$$

and $\lambda_{\Psi}^{(k)}$ is a tuning parameter.

For our purpose, any reasonable differential network estimator can correct for the bias. Thus, we treat differential network estimation as a black-box algorithm, and we obtain the estimates $\widehat{\Psi}^{(k)}$.

With the initial estimator $\check{\Omega}^{(k)}$ and refined differential network estimators $\widehat{\Psi}^{(k)}$, we construct the final transfer learning estimator for $\Omega^{(0)}$ as

$$\widehat{\Omega}^{(0)} = \sum_{k=0}^{K} \alpha_k \left(\check{\Omega}^{(k)} - \widehat{\Psi}^{(k)}\right), \tag{8}$$

where $\widehat{\Psi}^{(0)} = 0$ by definition. The final estimator Trans-Glasso (**Trans**fer learning **G**raphical **lasso**) integrates both shared information and source-specific refinements, yielding a transfer learning approach that leverages structural similarities across datasets for improved precision matrix estimation.

Compared to Li et al. [2023a], our method is more sample efficient as it does not require sample splitting between the steps.

4 Implementation in Practice

In this section, we provide practical guidelines for implementing Trans-Glasso, including optimization techniques, hyperparameter selection, and a method for identifying the informative set when not all sources are useful.

4.1 Optimization Algorithms

To implement Trans-Glasso, we first solve the Trans-MT-Glasso objective from Equation (3) to obtain initial estimates for $\Omega^{\star}$ and $\Gamma^{(k)\star}$. This is a constrained optimization problem that can be efficiently solved using the Alternating Direction Method of Multipliers (ADMM) [Boyd et al., 2011]. In this section, we slightly abuse notation by using superscripts to denote the iteration round and subscripts to represent the population.

Define

$$X = \begin{bmatrix} \Omega_0 \\ \Omega_1 \\ \vdots \\ \Omega_K \end{bmatrix} \in \mathbb{R}^{(K+1)d \times d}, \quad Y = \begin{bmatrix} \Omega \\ \Gamma_0 \\ \Gamma_1 \\ \vdots \\ \Gamma_K \end{bmatrix} \in \mathbb{R}^{(K+2)d \times d},$$

and

$$B = \begin{bmatrix} I_d & I_d & 0 & \cdots & 0 \\ I_d & 0 & I_d & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ I_d & 0 & 0 & \cdots & I_d \end{bmatrix} \in \mathbb{R}^{(K+1)d \times (K+2)d}.$$

With this notation, the objective in Equation (3) can be reformulated as

$$\text{minimize } f(X) + g(Y), \qquad \text{subject to } X = BY,$$

where

$$f(X) = \sum_{k=0}^{K} f_k(\Omega_k), \quad f_k(\Omega_k) = \alpha_k \left\{ -\log\det(\Omega_k) + \left\langle \widehat{\Sigma}^{(k)}, \Omega_k \right\rangle \right\} + \mathbb{I}(\Omega_k \succ 0), \quad 0 \leqslant k \leqslant K,$$

$$g(Y) = \lambda_{\mathrm{M}} |\Omega|_1 + \lambda_{\mathrm{M}} \sum_{k=0}^{K} \sqrt{\alpha_k}\, |\Gamma_k|_1,$$

where $\mathbb{I}(\Omega \succ 0) = 0$ if $\Omega \succ 0$ and $\mathbb{I}(\Omega \succ 0) = \infty$ otherwise.

The augmented Lagrangian for this problem is:

$$L_\rho(X, Y, Z) = f(X) + g(Y) + \rho \langle Z, X - BY \rangle + \frac{\rho}{2} \|X - BY\|_{\mathrm{F}}^2,$$

where

$$Z = \begin{bmatrix} Z_0 \\ Z_1 \\ \vdots \\ Z_K \end{bmatrix} \in \mathbb{R}^{(K+1)d \times d},$$

is a dual variable and ρ isa penalty parameter.

After initializing $Y^{(0)}$ and $Z^{(0)}$ such that $\Omega^{(0)}$, $\Gamma_k^{(0)}$, and $Z_k^{(0)}$ are symmetric, ADMM iteratively updates:

$$X^{(t)} = \arg\min_X L_\rho \left(X, Y^{(t-1)}, Z^{(t-1)} \right), \tag{9}$$

$$Y^{(t)} \in \arg\min_Y L_\rho \left(X^{(t)}, Y, Z^{(t-1)} \right), \tag{10}$$

$$Z^{(t)} = Z^{(t-1)} + \rho \left(X^{(t)} - BY^{(t)} \right). \tag{11}$$

Note that (11) is equivalent to $Z_k^{(t)} = Z_k^{(t-1)} + \rho \left(\Omega_k^{(t)} - \Omega^{(t)} - \Gamma_k^{(t)} \right)$ for $0 \leqslant k \leqslant K$. See Appendix H for detailed steps and stopping criteria.

To refine the initial estimators, we solve the D-Trace loss objective (7) for differential network estimation. Here, we use a different proximal gradient descent algorithm [Parikh and Boyd, 2014] following Zhao et al. [2022, 2019]. For simplicity, let $L_D(\Psi) \coloneqq L_D \left(\Psi ; \widehat{\Sigma}^{(0)}, \widehat{\Sigma}^{(k)} \right)$ for the chosen k. In iteration t, we update $\Psi^{(t-1)}$ by solving

$$\Psi^{(t)} = \arg\min_{\Psi} \left\{ \frac{1}{2} \left\| \Psi - \left(\Psi^{(t-1)} - \eta \nabla L_D \left(\Psi^{(t-1)} \right) \right) \right\|_F^2 + \eta \cdot \lambda_\Psi^{(k)} |\Psi|_1 \right\}, \tag{12}$$

where η is a user-specified step size. Note that $\nabla L_D(\cdot)$ is Lipschitz continuous with constant $\|\widehat{\Sigma}^{(0)} \otimes \widehat{\Sigma}^{(k)}\|_2 = \|\widehat{\Sigma}^{(0)}\|_2 \|\widehat{\Sigma}^{(k)}\|_2$. Thus, for $0 < \eta \leqslant \|\widehat{\Sigma}^{(0)}\|_2^{-1} \|\widehat{\Sigma}^{(k)}\|_2^{-1}$, the proximal gradient method converges [Beck and Teboulle, 2009]. The update in (12) has a closed-form solution:

$$\Psi_{jl}^{(t)} = \left[\left| A_{jl}^{(t-1)} \right| - \lambda_\Psi^{(k)} \eta \right]_+ \cdot A_{jl}^{(t-1)} \Big/ \left| A_{jl}^{(t-1)} \right|, \qquad 1 \leqslant j, l \leqslant d, \tag{13}$$

where $A^{(t-1)} = \Psi^{(t-1)} - \eta \nabla L_D(\Psi^{(t-1)})$ and $x_+ = \max\{0, x\}$, $x \in \mathbb{R}$.

Details on the optimization algorithms, including stopping criteria and descriptions, are in Section H.

4.2 Hyperparameter Selection

This section covers the selection of hyperparameters, specifically λ_{M} in Trans-MT-Glasso (3) and $\lambda_{\Psi}^{(k)}$ in D-Trace loss (7).

We choose $\lambda_{\Psi}^{(k)}$ for $k \in [K]$ to minimize the Bayesian information criterion (BIC) of D-Trace loss:

$$\mathrm{BIC}_{\Psi}^{(k)} = (n_0 + n_k) \left\| \frac{1}{2} \left(\widehat{\Sigma}^{(0)} \widehat{\Psi}^{(k)} \widehat{\Sigma}^{(k)} + \widehat{\Sigma}^{(k)} \widehat{\Psi}^{(k)} \widehat{\Sigma}^{(0)} \right) - \widehat{\Sigma}^{(0)} + \widehat{\Sigma}^{(k)} \right\|_{\mathrm{F}} + \log(n_0 + n_k) \cdot \left| \widehat{\Psi}^{(k)} \right|_0, \quad (14)$$

following Yuan et al. [2017]. After selecting $\lambda_{\Psi}^{(k)}$ and obtaining $\widehat{\Psi}^{(k)}$ for all $k \in [K]$, $\widehat{\Omega}^{(0)}$ depends on λ_{M}. Recall that $N = \sum_{k=0}^{K} n_k$. We choose λ_{M} to minimize the BIC of Trans-Glasso, defined as

$$\mathrm{BIC}_{\mathrm{Trans}} = N \cdot \left[\left\langle \widehat{\Sigma}^{(0)}, \widehat{\Omega}^{(0)} \right\rangle - \log\det\left(\widehat{\Omega}^{(0)} \right) \right] + \log N \cdot \left| \widehat{\Omega}^{(0)} \right|_0. \quad (15)$$

4.3 Identifying the Informative Set

In practice, it is not necessarily true that all source distributions are structurally similar to the target. We propose a data-driven method to estimate the informative set $\mathcal{A} \subseteq [K]$.

We obtain differential network estimations $\widehat{\Psi}^{(k)}$ for all $k \in [K]$, with the hyperparameter $\lambda_{\Psi}^{(k)}$ chosen by minimizing the BIC criterion in (14). We then rank sources according to the sparsity level of $\widehat{\Psi}^{(k)}$. Let R_k be the rank of the source k. For any $1 \leqslant k_1, k_2 \leqslant K$, $|\widehat{\Psi}^{(k_1)}|_0 \leqslant |\widehat{\Psi}^{(k_2)}|_0$ implies $R_{k_1} \leqslant R_{k_2}$. After ranking sources, we input samples into Trans-Glasso and determine the number of sources based on the cross-validation (CV) error. For $K_{\mathrm{chosen}} = 0, 1, \ldots, K$, we select sources k with $R_k \leqslant K_{\mathrm{chosen}}$. When $K_{\mathrm{chosen}} = 0$, we obtain $\widehat{\Omega}^{(0)}$ from graphical lasso [Friedman et al., 2008] using the target data alone. We then compute the CV error of K_{chosen} by the following procedure:

(i) We randomly split the *target samples* into M-fold.

(ii) For $m = 1, \ldots, M$, we select the m-th fold as the validation set, $\mathcal{D}_{\mathrm{val}}$, and the rest as the training set. We input the training set and chosen source samples into Trans-Glasso to obtain $\widehat{\Omega}^{(0)}$. We compute the CV error for the m-th fold as

$$\mathrm{CV}_m = \frac{1}{2d} \left\{ \frac{1}{|\mathcal{D}_{\mathrm{val}}|} \sum_{i \in \mathcal{D}_{\mathrm{val}}} \mathrm{tr}\left(\mathbf{x}_i^{(0)} \mathbf{x}_i^{(0)\top} \widehat{\Omega}^{(0)} \right) - \log\det\left(\widehat{\Omega}^{(0)} \right) \right\} + \frac{1}{2} \log \pi, \quad (16)$$

and define $\mathrm{CV}(K_{\mathrm{chosen}}) = \frac{1}{M} \sum_{m=1}^{M} \mathrm{CV}_m$.

We set the estimated informative set as $\widehat{\mathcal{A}} = \{k \in [K] : R_k \leqslant K_{\mathrm{chosen}}^{\star}\}$, where $K_{\mathrm{chosen}}^{\star} = \arg\min_{k=0,1,\ldots,K} \mathrm{CV}(k)$. Source samples from $\widehat{\mathcal{A}}$ are used to estimate $\Omega^{(0)}$. If $\widehat{\mathcal{A}} = \varnothing$, we obtain $\widehat{\Omega}^{(0)}$ via graphical lasso on target samples. We note that sample splitting is not necessary between estimating $\mathcal{A}$ and $\Omega^{(0)}$. This procedure is called Trans-Glasso-CV. We demonstrate its empirical performance in Section 6.

5 Theoretical Analysis

In this section, we establish theoretical guarantees for the Trans-Glasso algorithm. We begin by analyzing the initial estimation step using Trans-MT-Glasso, followed by an error bound for the

complete Trans-Glasso estimator. Finally, we derive a minimax lower bound, demonstrating that Trans-Glasso is minimax optimal in a wide range parameter regimes.

To simplify the theoretical statements, we assume the following condition throughout this section.

Assumption 2. *Assume that*

$$M_\Sigma := \max_{0\leqslant k\leqslant K} \left|\Sigma^{(k)}\right|_\infty = O(1) \qquad \text{and} \qquad M_\Omega := \max_{0\leqslant k\leqslant K} \left\|\Omega^{(k)}\right\|_2 = O(1). \tag{17}$$

5.1 Analysis of Trans-MT-Glasso

We first provide error bounds for the initial multi-task estimation step, Trans-MT-Glasso. This method estimates both the shared precision matrix component $\Omega^\star$ and the deviation matrices $\Gamma^{(k)\star}$ based on the structural similarity outlined in Assumption 1.

The following theorem provides a high probability upper bound on the Frobenius norm error for the Trans-MT-Glasso estimator. Recall that $N = \sum_{k=0}^K n_k$.

Theorem 1. *Suppose Assumptions 1 and 2 hold. Fix a failure probability $\delta \in (0, 1]$. Suppose that the local sample size is large enough so that*

$$\min_{0\leqslant k\leqslant K} n_k \geqslant 2\log(2\,(K+2)\,d^2/\delta). \tag{18}$$

Set $M_{\text{op}} \geqslant M_\Omega$ and the penalty parameter λ_M such that

$$\lambda_M \geqslant 160 M_\Sigma \sqrt{\frac{\log(2\,(K+2)\,d^2/\delta)}{2N}}. \tag{19}$$

Then with probability at least $1-\delta$, the estimator satisfies

$$\sum_{k=0}^K \alpha_k \|\check{\Omega}^{(k)} - \Omega^{(k)}\|_{\text{F}}^2 \leqslant \frac{18\,(s+(K+1)h)\,\lambda_M^2}{\kappa^2}, \tag{20}$$

where $\kappa = (2M_\Omega + M_{\text{op}})^{-2}$.

Note that the loss function defined in (4) is not strong convex with respect to the Euclidean norm. This prevents us from obtaining error guarantee for individual prevision matrix. Nevertheless, we make a key observation that the loss function exhibits strong convexity with respect to the weighted norm used in (20). See Appendix C for a detailed proof.

When we choose $\lambda_M \asymp \sqrt{(\log d)/N}$, the rate shown in Theorem 1 consists of two parts. The first part, which is of the order $(s\log d)/N$, refers to the estimation error of the shared component. In words, Trans-Glasso uses all the samples to estimate the shared component. The second part, of the order $(Kh\log d)/N$, relates to the estimation error of the individual components, i.e., on average, there are N/K samples to estimate each individual ocmponent.

5.2 Analysis of Trans-Glasso

After the initial estimates are obtained via Trans-MT-Glasso, the differential network estimation step refines these estimates by isolating the deviations $\Psi^{(k)}$. This yields the final Trans-Glasso estimator $\widehat{\Omega}^{(0)}$.

As discussed in Section 3, any differential network estimator can be used in Step 2 for refinement. The differential network estimates $\widehat{\Psi}^{(k)}$ are treated as the result of a black-box algorithm, obeying

$$\left\|\widehat{\Psi}^{(k)} - \Psi^{(k)}\right\|_{\text{F}} \lesssim g_{\text{F}}^{(k)}(n_0, n_k, d, h, M_\Gamma, \delta) := g_{\text{F}}^{(k)} \tag{21}$$

holds simultaneously for all $k = 1, \ldots, K$ with probability at least $1 - \delta$. We now establish a non-asymptotic error bound for this estimator, which combines the initial estimation error with the error from differential network estimation.

Theorem 2. *Let $\widehat{\Omega}^{(0)}$ be the Trans-Glasso estimator obtained in Equation (8). Under the same conditions as in Theorem 1, and assuming Equation (21) holds for the differential network estimators. Then with probability at least $1 - 2\delta$, one has*

$$\left\|\widehat{\Omega}^{(0)} - \Omega^{(0)}\right\|_{\mathrm{F}}^2 \lesssim \left(\frac{s}{N} + \frac{h}{\bar{n}}\right) \log(2\,(K+2)\,d^2/\delta) + \sum_{k=0}^{K} \alpha_k g_{\mathrm{F}}^{(k)}.$$

The error rates depend on differential network estimators' performance. Next, we provide specific error rates using the D-Trace loss estimator.

5.2.1 A differential network estimator: D-Trace loss minimization

We characterize $g_{\mathrm{F}}^{(k)}$ in the case when the D-Trace loss estimator in (7) is used. We use the tighter D-Trace loss estimator analysis from Zhao et al. [2019, 2022], Tugnait [2023].

We then have the following theorem providing a high-probability error upper bound for D-Trace loss estimator. The theorem is derived directly from Theorem 7 in Appendix D.

Theorem 3. *Suppose that Assumptions 1-2 hold. Assume that* $\min_{0\leqslant k\leqslant K} n_k \gg h^2 \log\left(2(K+1)d^2/\delta\right)$. *Set*

$$\lambda_{\Psi}^{(k)} = C\sqrt{\frac{\log\left(2(K+1)d^2/\delta\right)}{\min\{n_k, n_0\}}} \quad \text{for all } k \in [k]$$

for some large constant $C > 0$. Then

$$\left\|\widehat{\Psi}^{(k)} - \Psi^{(k)}\right\|_{\mathrm{F}} \lesssim \sqrt{h} M_\Gamma \sqrt{\frac{\log\left(2(K+1)d^2/\delta\right)}{\min\{n_k, n_0\}}}$$

holds simultaneously for all $k \in [K]$ with probability at least $1 - \delta$.

By Theorem 3, we have

$$g_{\mathrm{F}}^{(k)} = \sqrt{h} M_\Gamma \sqrt{\frac{\log\left(2(K+1)d^2/\delta\right)}{\min\{n_k, n_0\}}}$$

for D-Trace loss estimator. Plug the above results into Theorem 2, we have the following corollary.

Corollary 1. *Let $\widehat{\Omega}^{(0)}$ be obtained by Trans-Glasso (8) with the D-Trace loss estimator used in Step 1. Instate the assumptions in Theorems 1 and 3. For a given $\delta \in (0, 1]$, letting*

$$\lambda_M \asymp \sqrt{\frac{\log(2\,(K+2)\,d^2/\delta)}{N}}, \quad \lambda_{\Psi}^{(k)} \asymp M_\Gamma \sqrt{\frac{\log\left(2(K+1)d^2/\delta\right)}{\min\{n_k, n_0\}}} \quad \text{for all } k \in [K],$$

we have that

$$\left\|\widehat{\Omega}^{(0)} - \Omega^{(0)}\right\|_{\mathrm{F}} \lesssim \left(\sqrt{\frac{s}{N}} + (1 + M_\Gamma)\sqrt{\frac{h}{\bar{n}}} + M_\Gamma\sqrt{\frac{h}{n_0}}\right) \sqrt{\log(2\,(K+2)\,d^2/\delta)}$$

holds with probability at least $1 - 2\delta$.

The estimation error consists of three parts: shared component estimation, individual component estimation, and differential network estimation. If $\bar{n} \geqslant n_0$ and M_Γ is bounded by a universal constant, the error scales as $\sqrt{\frac{s \log d}{N}} + \sqrt{\frac{h \log d}{n_0}}$. When $N \gg n_0$, the error rate can be significantly reduced compared to the rate obtained by only using target samples, which is in the order of $\sqrt{\frac{s \log d}{n_0}} + \sqrt{\frac{h \log d}{n_0}}$.

5.3 Minimax Lower Bounds and Optimality

To evaluate the theoretical performance of Trans-Glasso, we derive the minimax lower bound for estimating the target precision matrix $\Omega^{(0)}$ over the parameter space defined by Assumptions 1-2. More precisely, we define the relevant parameter space:

$$\mathcal{G}(s,h) := \left\{ \left\{\Omega^{(k)}\right\}_{k=0}^{K} : \Omega^{(k)} \succ 0,\ \Omega^{(k)} = \Omega^{\star} + \Gamma^{(k)\star},\ \mathrm{supp}\left(\Omega^{\star}\right) \cap \mathrm{supp}\left(\Gamma^{(k)\star}\right) = \varnothing\ \forall 0 \leqslant k \leqslant K, \right.$$
$$\left. |\Omega^{\star}|_0 \leqslant s,\ \max_{0\leqslant k\leqslant K} \left|\Gamma^{(k)\star}\right|_0 \leqslant h,\ \max_{0\leqslant k\leqslant K} \left|\Gamma^{(k)\star}\right|_1 \leqslant M_{\Gamma},\ \max_{0\leqslant k\leqslant K} \left\|\Omega^{(k)}\right\|_2 \leqslant M_{\Omega},\ \max_{0\leqslant k\leqslant K} \left|\Sigma^{(k)}\right|_{\infty} \leqslant M_{\Sigma} \right\}, \tag{22}$$

where $M_{\Gamma} > 0$, $M_{\Omega}, M_{\Sigma} > 1$ are universal constants.

Intuitively, the performance limit of any transfer learning estimator is dictated by the information-theoretic lower bounds of estimating two parts, namely the shared component and the individual component. Hence, to derive the minimax lower bound for the transfer learning estimator, we need to provide lower bounds for estimating these two parts.

Lower bound for estimating shared component. The following theorem provides the minimax lower bound for estimating the shared component when all the distributions are the same.

Theorem 4. *Assume that we have n i.i.d. samples $X_1, \ldots, X_n$ from $N(0, \Omega^{-1})$, where*

$$\Omega \in \mathcal{G}_1 = \left\{\Omega \in \mathbb{S}^{d\times d} : \Omega \succ 0,\ |\Omega|_0 \leqslant s,\ 0 < c_1 \leqslant \gamma_{\min}(\Omega) \leqslant \gamma_{\max}(\Omega) \leqslant c_2 < \infty\right\},$$

and c_1, c_2 are universal constants. In addition, assume that $s \geqslant d \geqslant c' n^{\beta}$ for some universal constants $\beta > 1$ and $c' > 0$, and

$$\lceil s/d \rceil = o\left(\frac{n}{(\log d)^{\frac{3}{2}}}\right).$$

We then have

$$\inf_{\widehat{\Omega}} \sup_{\Omega\in\mathcal{G}_1} \mathbb{E}\left[\left\|\widehat{\Omega} - \Omega\right\|_{\mathrm{F}}^2\right] \gtrsim \frac{s\log d}{n}.$$

The proof is based on [Cai et al., 2016b, Theorem 6.1]. See Appendix E for more details. By Theorem 4, it is easy to see that the squared Frobenius error of estimating the shared component is lower bounded by $s\log d/N$ where N is the total number of samples.

Lower bound for estimating the individual components. We also provide a lower bound for estimating the individual component. When $\Omega^{(k)} = I_d$ for all $1 \leqslant k \leqslant K$ and $\Omega^{(0)} = I_d + \Delta$ with $\mathrm{diag}(\Delta) = 0$, the source samples are not helpful to estimate $\Omega^{(0)}$ at all. Thus, the minimax lower bound for estimating Δ provides a valid minimax lower bound for the transfer learning problem.

Theorem 5. *Assume that we have n i.i.d. samples $X_1, \ldots, X_n$ from $N(0, \Omega^{-1})$, where*

$$\Omega \in \mathcal{G}_2 = \left\{\Omega \in \mathbb{S}^{d\times d} : \Omega \succ 0,\ \Omega = I_d + \Delta,\ \Delta_{jj} = 0 \text{ for all } 1 \leqslant j \leqslant d,\right.$$
$$\left. |\Delta|_0 \leqslant h,\ |\Delta|_1 \leqslant C_{\Gamma},\ 0 < c_1 \leqslant \gamma_{\min}(\Omega) \leqslant \gamma_{\max}(\Omega) \leqslant c_2 < \infty\right\}, \tag{23}$$

where $C_{\Gamma} > 0$, $c_1 < 1$ and $c_2 > 1$ are constants. In addition, assume that

$$d \geqslant 4h,\ h\log d \geqslant 8\log 3,\ \frac{h\log d}{n} \leqslant \min\left\{2, 8(1-c_1)^2, 8(1-c_2)^2\right\},\ h\sqrt{\frac{\log d}{n}} \leqslant 4C_{\Gamma}. \tag{24}$$

We then have

$$\inf_{\widehat{\Omega}} \sup_{\Omega\in\mathcal{G}_2} \mathbb{E}\left[\left\|\widehat{\Omega} - \Omega\right\|_{\mathrm{F}}^2\right] \gtrsim \frac{h\log d}{n}.$$

See Appendix F for the proof, which relies on a novel construction of the packing set of the parameter space and on the celebrated Fano's method [Wainwright, 2019, Section 15.3].

It is worth noting that Theorem 5 also provides a minimax lower bound for estimating the differential network $\Omega_X - \Omega_Y$ for two precision matrices Ω_X and Ω_Y when the L_1-norm of the differential network is bounded. To our knowledge, this is also the first lower bound for differential network estimation [Zhao et al., 2014, Yuan et al., 2017]. As a result, we can derive the first minimax optimal rate for differential network estimation. See Appendix I for a more detailed discussion.

Combining pieces together. Combining Theorem 4 and Theorem 5, we have the following lower bound for any transfer learning estimator.

Theorem 6. *Suppose that we have n_k i.i.d. samples from a sub-Gaussian distribution $\mathcal{P}_k$ with zero mean and precision matrix $\Omega^{(k)}$ for all $0 \leqslant k \leqslant K$. Besides, assume that $s \geqslant d \geqslant c' N^{\beta}$ for some universal constants $\beta > 1$, $c' > 0$ and*

$$\lceil s/d \rceil = o\left(\frac{N}{(\log d)^{\frac{3}{2}}} \right).$$

In addition, assume that

$$d \geqslant 4h,\ h \log d \geqslant 8 \log 3,\ h\sqrt{\frac{\log d}{n_0}} \leqslant 4M_\Gamma,\ \frac{h \log d}{n_0} \leqslant \min\left\{ 2,\ 8\left(1 - M_\Omega\right)^2,\ 8\left(1 - \frac{1}{M_\Sigma}\right)^2 \right\},$$

where $M_\Gamma > 0, M_\Omega, M_\Sigma > 1$ are universal constants defined in (22). We then have

$$\inf_{\widehat{\Omega}} \sup_{\{\Omega^{(k)}\}_{k=0}^K \in \mathcal{G}(s,h)} \mathbb{E}\left[\left\| \widehat{\Omega} - \Omega^{(0)} \right\|_{\mathrm{F}}^2 \right] \gtrsim \frac{s \log d}{N} + \frac{h \log d}{n_0}.$$

Theorem 6 demonstrates that Trans-Glasso achieves minimax optimality for the parameter space specified in (22) when $\bar{n} \geqslant n_0$. The obtained minimax optimal rate is reasonable, considering we have N samples for estimating the shared component with s non-zero entries and n_0 samples for estimating the individual component with h non-zero entries. Furthermore, from a practical viewpoint, the rate suggests that the target sample size only needs to be sufficiently large in relation to the sparsity level h of the individual component. In contrast, if we only have target samples, the target sample size needs to be large enough to match the sparsity level $s + h$ of the entire precision matrix, which can be significantly larger.

6 Simulations

In this section, we demonstrate the empirical performance of Trans-Glasso through a series of simulations. We evaluate its accuracy in comparison with several baseline methods under different settings, including varying sample sizes and sparsity levels.

We set the dimensionality $d = 100$ and the number of source distributions $K = 5$ for all experiments. In each experiment, we vary parameters such as the target sample size n_0, the source sample size $n_k = n_{\text{source}}$, and the sparsity level h to assess the robustness of Trans-Glasso across diverse conditions. The data are simulated under three different models, each reflecting a specific structure for the shared and individual components of the precision matrices.

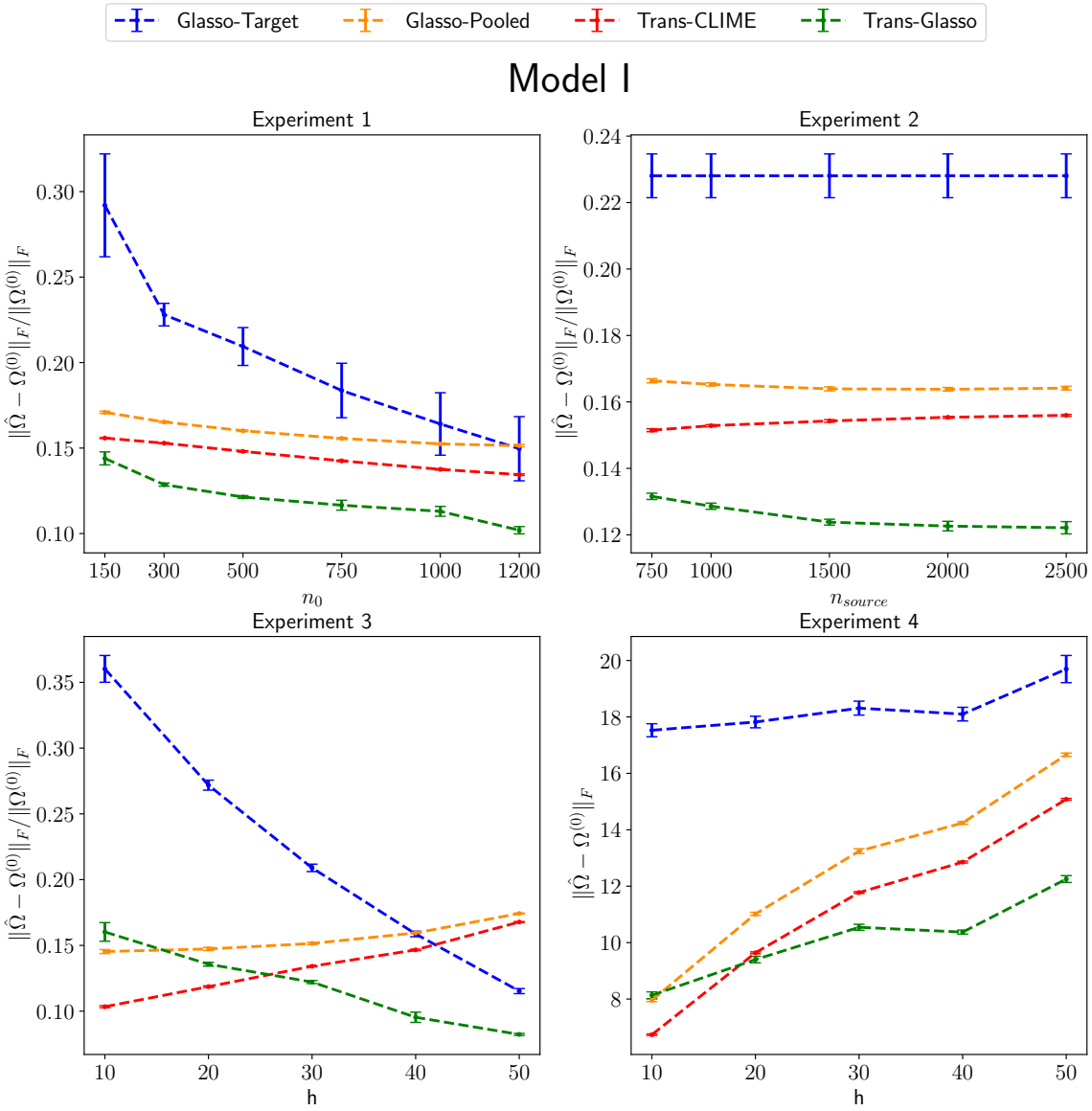

Figure 2: Simulation results for Model I. The default setting is $n_0 = 300$, $n_{\text{source}} = 1000$ and $h = 40$. In the first experiment, we increase n_0 while fixing n_{source} and h. In the second experiment, we increase n_{source} while fixing n_0 and h. In the third experiment, we increase both n_0 and n_{source} while increasing h. More specifically, we let $n_{\text{source}} = 3n_0$, and $n_0 = 70$ when $h = 10$, $n_0 = 150$ when $h = 20$, $n_0 = 300$ when $h = 30$, $n_0 = 600$ when $h = 40$ and $n_0 = 1200$ when $h = 50$. In the fourth experiment, we fix both n_0 and n_{source} while increasing h. Each dot represents the empirical mean across 30 repetitions and the vertical bar represents Mean $\pm \frac{2}{\sqrt{30}} \times$ Standard Error.

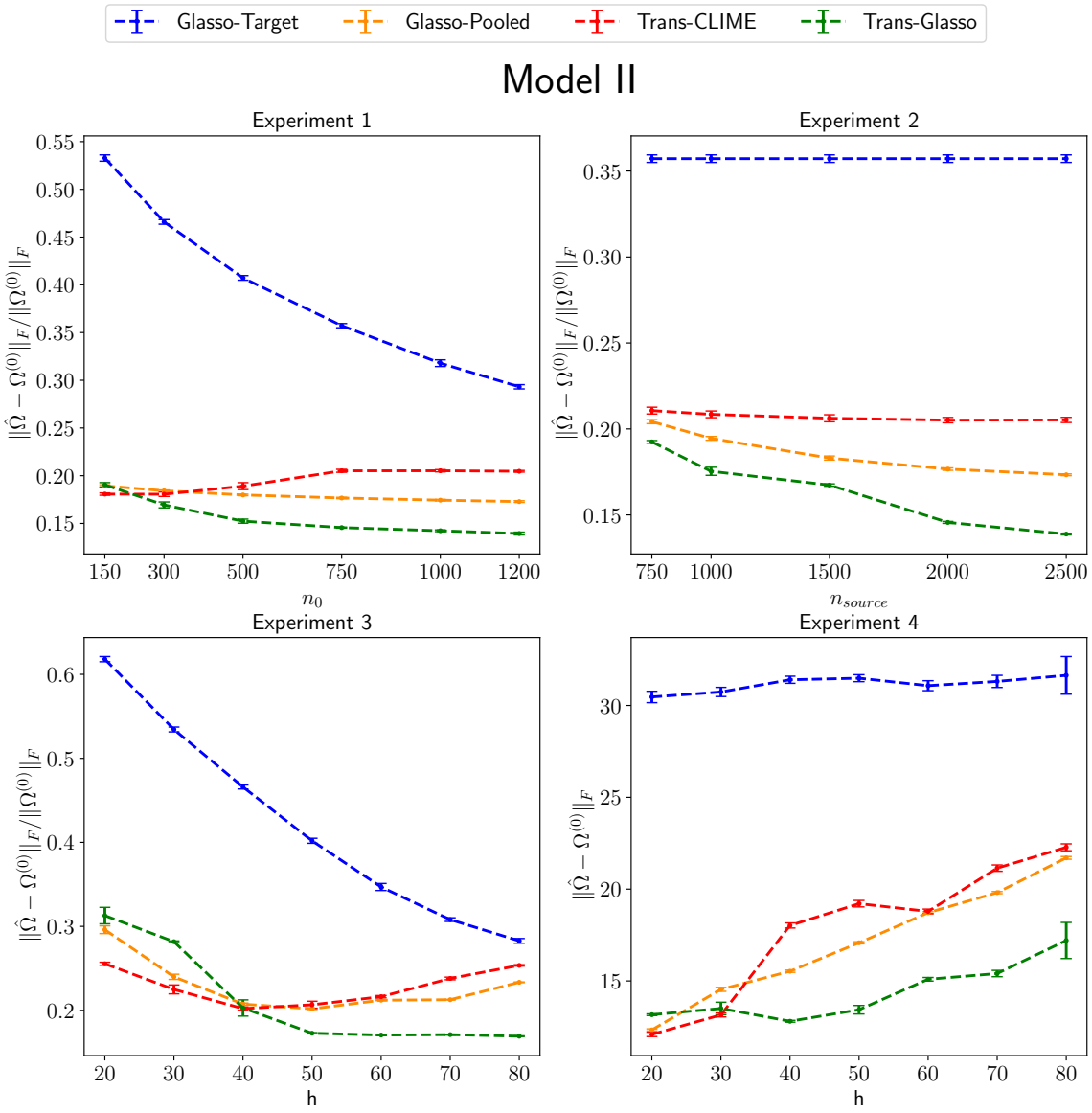

Figure 3: Simulation results for Model II. The default setting is $n_0 = 750$, $n_{\text{source}} = 2000$ and $h = 40$. In the first experiment, we increase n_0 while fixing n_{source} and h. In the second experiment, we increase n_{source} while fixing n_0 and h. In the third experiment, we increase both n_0 and n_{source} while increasing h. More specifically, we let $n_{\text{source}} = 3n_0$, and $n_0 = 100$ when $h = 20$, $n_0 = 200$ when $h = 30$, $n_0 = 300$ when $h = 40$, $n_0 = 500$ when $h = 50$, $n_0 = 800$ when $h = 60$, $n_0 = 1000$ when $h = 70$ and $n_0 = 1200$ when $h = 80$. In the fourth experiment, we fix both n_0 and n_{source} while increasing h. Each dot represents the empirical mean across 30 repetitions and the vertical bar represents Mean $\pm \frac{2}{\sqrt{30}} \times$ Standard Error.

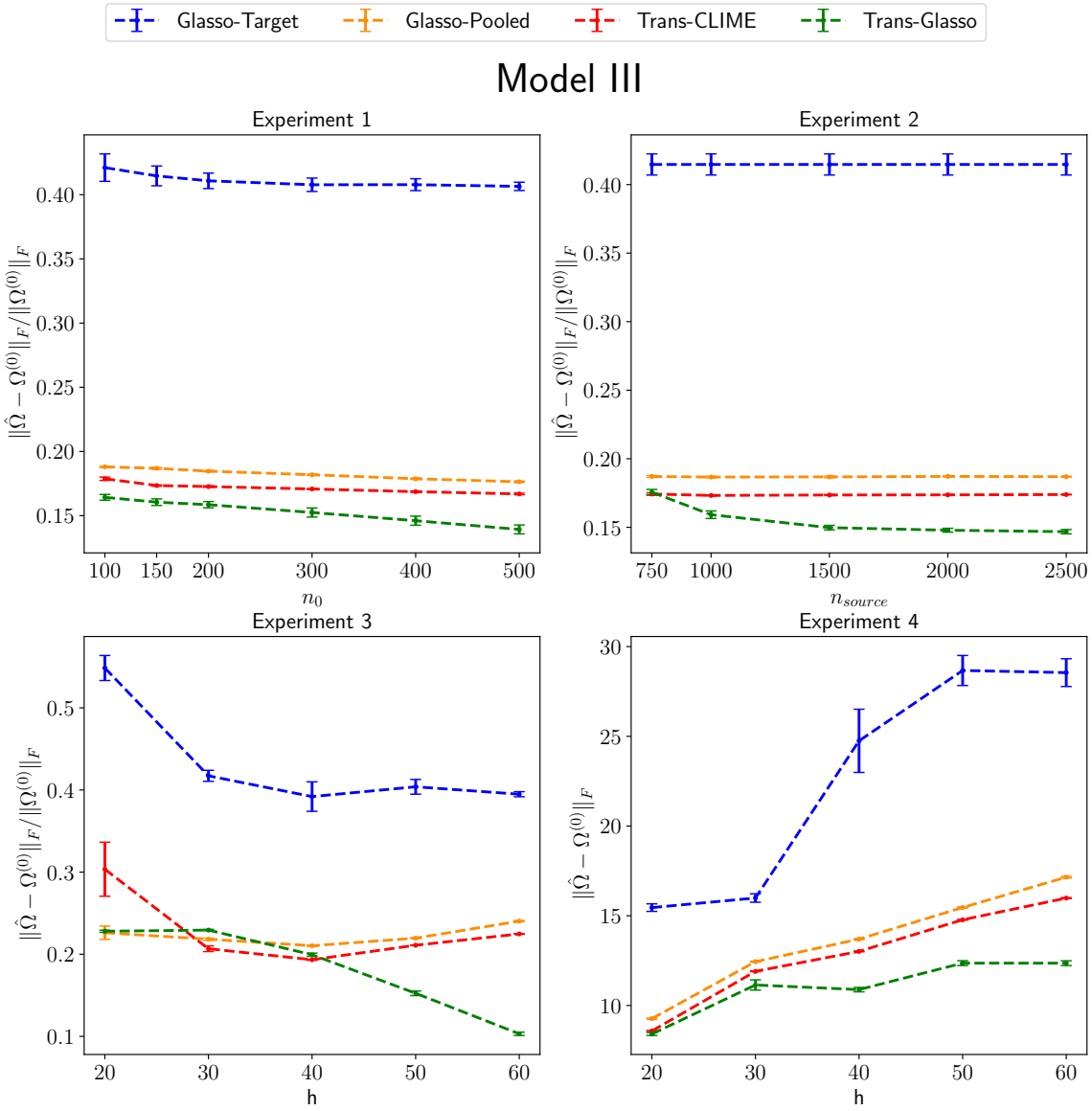

Figure 4: Simulation results for Model III. The default setting is $n_0 = 150$, $n_{\text{source}} = 1000$ and $h = 40$. In the first experiment, we increase n_0 while fixing n_{source} and h. In the second experiment, we increase n_{source} while fixing n_0 and h. In the third experiment, we increase both n_0 and n_{source} while increasing h. More specifically, we let $n_{\text{source}} = 4n_0$, and $n_0 = 15$ when $h = 20$, $n_0 = 30$ when $h = 30$, $n_0 = 80$ when $h = 40$, $n_0 = 300$ when $h = 50$, and $n_0 = 1000$ when $h = 60$. In the fourth experiment, we fix both n_0 and n_{source} while increasing h. Each dot represents the empirical mean across 30 repetitions and the vertical bar represents Mean $\pm \frac{2}{\sqrt{30}} \times$ Standard Error.

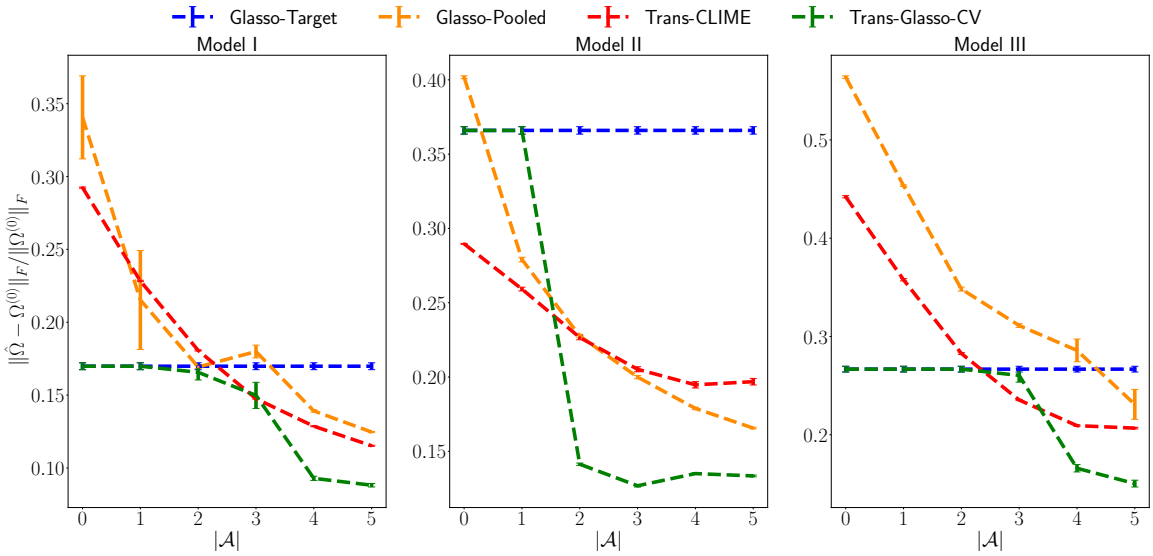

Figure 5: Simulation results when the informative set $\mathcal{A}$ is unknown. We set $n_0 = 300$ and $n_{\text{source}} = 1000$ for Model I; $n_0 = 750$ and $n_{\text{source}} = 2000$ for Model II; and $n_0 = 300$ and $n_{\text{source}} = 1000$ for Model III. Each dot represents the empirical mean across 30 repetitions and the vertical bar represents Mean $\pm \frac{2}{\sqrt{30}} \times$ Standard Error.

6.1 Data Generation Models

We generate data from three distinct models to assess the flexibility of Trans-Glasso. Each model starts with a shared component, followed by individual components. The final precision matrices are made positive definite by adding a diagonal matrix. Specifically, each model is set as follows.

- **Model I:** The shared component is a banded matrix with bandwidth 1, where each entry $\tilde{\Omega}_{ij} = 5 \times 0.6^{|i-j|} \mathbb{1}(|i-j| \leqslant 1)$ for $1 \leqslant i, j \leqslant d$. For a given h and each $k = 0, 1, \ldots, K$, we uniformly choose $\lceil h/2 \rceil$ entries (i, j) such that $1 \leqslant i \leqslant \lfloor \frac{d}{2} \rfloor$ and $\lfloor \frac{d}{2} \rfloor + 1 \leqslant j \leqslant d$, denoted as $\mathcal{S}_{\Gamma^{(k)},\text{up}}$. We let $\tilde{\Gamma}^{(k)}_{ij} = u_{ij} \mathbb{1}\{(i, j) \in \mathcal{S}_{\Gamma^{(k)},\text{up}}\}$, $1 \leqslant i, j \leqslant d$, where u_{ij}'s are from Unif$[-3, 3]$. Then $\Gamma^{(k)\star} = \tilde{\Gamma}^{(k)} + (\tilde{\Gamma}^{(k)})^\top$. Finally, we let $\Omega^{(k)} = \tilde{\Omega} + \Gamma^{(k)\star} + \sigma I_d$, where σ ensures $\gamma_{\min}(\Omega^{(k)}) \geqslant 0.1$ for $0 \leqslant k \leqslant K$.
- **Model II:** Model II is similar to Model I but with a wider bandwidth of 5, introducing a more connected structure in the shared component.
- **Model III:** We generate the shared component from an Erdos–Renyi graph. Specifically, let $\tilde{\Omega}_{ii} = 5$, $1 \leqslant i \leqslant d$, and $U_{ij} \sim$ Bernoulli(0.02), $1 \leqslant i < j \leqslant d$. If $U_{ij} = 1$, let $\tilde{\Omega}_{ij} = \tilde{\Omega}_{ji} \sim$ Unif$[-3, 3]$; otherwise, set $\tilde{\Omega}_{ij} = \tilde{\Omega}_{ji} = 0$. Let $\mathcal{S}_{\tilde{\Omega}} = \{(i, j) \in [d] \times [d] : \tilde{\Omega}_{ij} \neq 0\}$ be the support of $\tilde{\Omega}$. For given h and $0 \leqslant k \leqslant K$, uniformly choose h entries ($h + 1$ if h is odd) from $[d] \times [d] \backslash \tilde{\Omega}$, denoted as $\mathcal{S}_{\Gamma^{(k)}}$, such that $(i, j) \in \mathcal{S}_{\Gamma^{(k)}}$ if and only if $(j, i) \in \mathcal{S}_{\Gamma^{(k)}}$. Let $\Gamma^{(k)\star}_{ij} = u_{ij} \mathbb{1}\{(i, j) \in \mathcal{S}_{\Gamma^{(k)}}\}$, $1 \leqslant i, j \leqslant d$, where $u_{ij} \sim$ Unif$[-3, 3]$. Finally, let $\Omega^{(k)} = \tilde{\Omega} + \Gamma^{(k)\star} + \sigma I_d$, where σ ensures $\gamma_{\min}(\Omega^{(k)}) \geqslant 0.1$ for $0 \leqslant k \leqslant K$.

6.2 Experimental Design

We conduct four main experiments per model to investigate Trans-Glasso's performance under various conditions:

1. **Experiment 1**: Vary the target sample size n_0 while keeping the source sample size n_{source} and sparsity h fixed.
2. **Experiment 2**: Vary the source sample size n_{source} while keeping n_0 and h fixed.
3. **Experiment 3**: Increase both n_0, n_{source}, and h proportionally to examine scalability.
4. **Experiment 4**: Fix n_0 and n_{source} while increasing h, assessing performance as sparsity in deviations increases.

Each experiment is repeated 30 times to obtain reliable averages and standard errors.

6.3 Comparison Methods

We compare Trans-Glasso with the following baseline methods:

- **Glasso-Target**: Applies graphical lasso [Friedman et al., 2008] only to the target data.
- **Glasso-Pooled**: Combines all target and source data, applying graphical lasso on the pooled dataset.
- **Trans-CLIME**: A transfer learning approach for precision matrix estimation by Li et al. [2023a], which assumes a sparse divergence matrix across sources.

6.4 Results

The results for Models I – III are shown in Figure 2–4. Trans-Glasso generally outperforms baseline methods. In Experiment 3, Trans-Glasso shows consistency across all models, whereas Glasso-Pooled and Trans-CLIME do not. Trans-CLIME performs better with small h, but its performance deteriorates as h increases. This is because a small increase in h can significantly increase the sparsity of the divergence matrix defined in (2) when the covariance matrix is not sparse, as discussed in Appendix A. Therefore, when the precision matrix is sparse but the covariance matrix is not, Trans-Glasso is more reliable and robust.

6.5 Experiments with Unknown Informative Set

We perform simulation experiments with unknown $\mathcal{A}$, using the same three models. We divide $[K]$ into $[K] = \mathcal{A} \cup \mathcal{A}^c$. For $k \in \mathcal{A}$, we set the sparsity level h to be small, and for $k \in \mathcal{A}^c$, h to be large. Specifically, for Model I, $h = 20$ for $k \in \mathcal{A}$ and $h = 600$ for $k \in \mathcal{A}^c$; for Model II, $h = 30$ for $k \in \mathcal{A}$ and $h = 600$ for $k \in \mathcal{A}^c$; for Model III, $h = 10$ for $k \in \mathcal{A}$ and $h = 300$ for $k \in \mathcal{A}^c$. We implement the Trans-Glasso-CV algorithm (Section 4.3) and compare it with other methods. We vary $|\mathcal{A}|$ from 0 to K to observe performance changes. Each experiment is repeated 30 times with different random seeds.

Figure 5 shows that Trans-Glasso-CV generally outperforms baseline methods. Notably, it never performs worse than Glasso-Target, indicating no "negative transfer" of knowledge. In contrast, both Glasso-Pooled and Trans-CLIME can underperform compared to Glasso-Target. Additionally, as $|\mathcal{A}|$ increases, Trans-Glasso-CV achieves the best performance.

7 Real-World Data Analysis

We apply the Trans-Glasso algorithm to two real-world datasets. In Section 7.1, we use it on gene networks with different brain tissues. In Section 7.2, we use it on protein networks for various cancer subtypes.

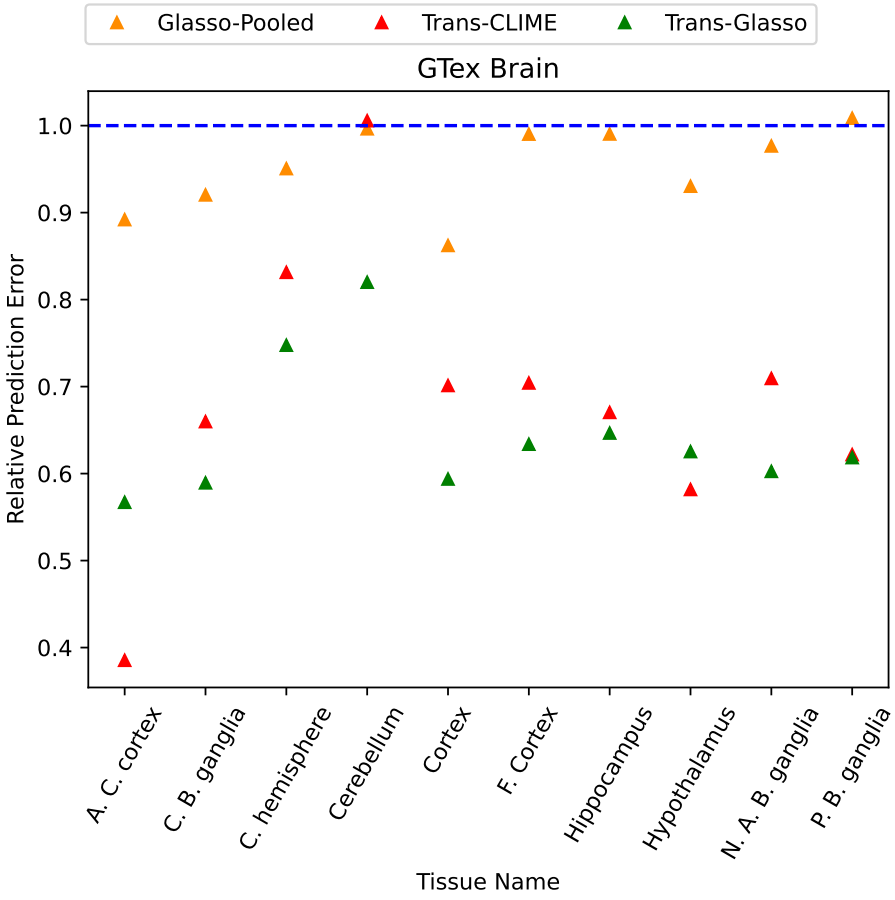

Figure 6: Cross-validation prediction error of different methods on GTEx brain tissue datasets, relative to Glasso-Target.

7.1 Gene Networks Data for Brain Tissues

We apply Trans-Glasso to detect gene networks in different tissues using the Genotype-Tissue Expression (GTEx) data[1]. Following Li et al. [2023a], we focus on genes related to central nervous system neuron differentiation (GO:0021953). We use the same 13 brain tissues as Li et al. [2023a], treating one as the target and the other 12 as sources. We only use 10 out of 13 tissues as targets, avoiding 3 due to small sample sizes. See Table 3 of the supplementary materials of Li et al. [2023a] for the complete list of tissues and Table 1 for the target tissues. We remove genes with missing values in these tissues and compare Trans-Glasso with baseline methods by computing cross-validation prediction error as defined in (16).

Figure 6 presents the final result, using Glasso-Target, Glasso-Pooled, and Trans-CLIME as baselines as in Section 6. To compare results across tissues, we report prediction errors relative to Glasso-Target. Figure 6 shows that Trans-Glasso performs best on most tissues. The relative prediction error of Trans-Glasso is always much smaller than 1, indicating that it performs significantly better than using target data alone and is robust to negative transfer. In comparison, Glasso-Pooled and Trans-CLIME can perform worse or similar to Glasso-Target on some tissues.

[1] https://gtexportal.org/home/

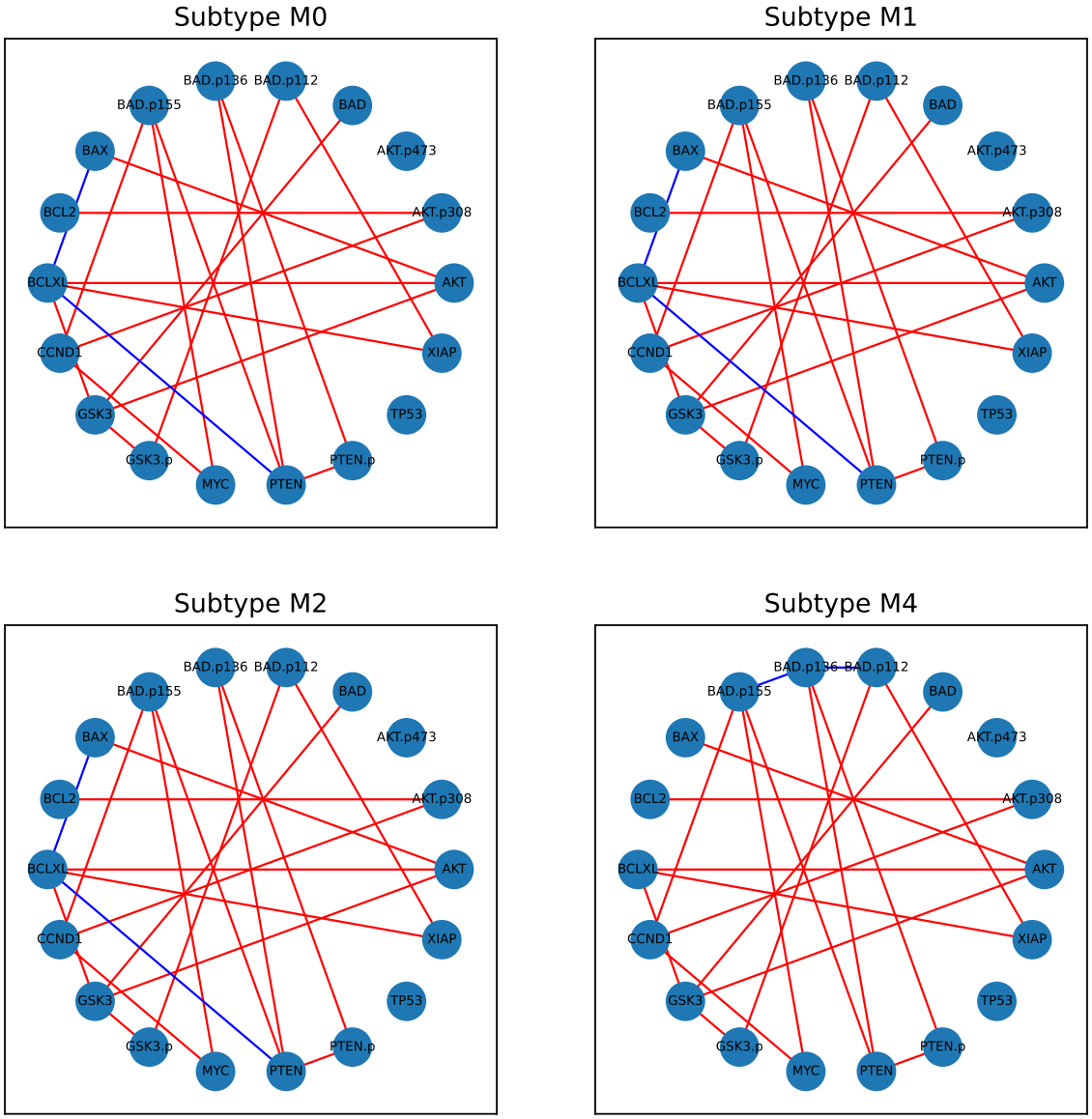

Figure 7: Protein networks for four AML subtypes. Red edges are shared by all, blue edges are not.

7.2 Protein Networks Data for AML

We apply our method to a protein networks dataset for Acute Myeloid Leukemia (AML) subtypes. Understanding protein relationships in cells is crucial in cancer studies, and graphical models help build these networks. Following Peterson et al. [2015], we analyze protein levels for 213 newly diagnosed AML patients,[2] classified by the FAB system. Although protein interactions may differ across subtypes, common AML-related processes suggest shared connections. Thus, transfer learning can enhance subtype estimation using data from other subtypes. We focus on 18 proteins involved in apoptosis and cell cycle regulation, studying four subtypes: M0 (17 subjects), M1 (34 subjects), M2 (68 subjects), and M4 (59 subjects) Peterson et al. [2015], Kanehisa et al. [2012].

For each subtype, after applying Trans-Glasso, we obtain the final estimated graph by choosing 20 edges with the largest absolute values in the estimated precision matrix, resulting in graphs with similar edge numbers as in Peterson et al. [2015]. The final result is shown in Figure 7. Comparing with Peterson et al. [2015], many edges are discovered in both studies. However, our estimated graphs are more similar across subtypes. M0, M1, and M2 have the same structures, while M4 differs in two edges and has stronger connections between proteins in BAD families. This is supported by Tzifi et al. [2012], which observed higher expression levels of BAD family proteins in AML subtypes M4, M5, and M6 [Tzifi et al., 2012, Table 1], indicating more active interactions.

8 Conclusion

We introduce Trans-Glasso, a novel transfer learning approach to precision matrix estimation, which addresses the limitations of small target sample sizes by leveraging related source data. Trans-Glasso operates through a two-step process: an initial estimation using multi-task learning, followed by refinement via differential network estimation. This methodology achieves minimax optimality in a wide range of parameter regimes. Through extensive simulations, we demonstrate that Trans-Glasso consistently outperforms baseline methods, showcasing its robustness and adaptability, particularly in high-dimensional settings with limited target samples.

Future research directions include extending Trans-Glasso to estimate other graphical models, such as Gaussian copula [Liu et al., 2009, 2012a], transelliptical [Liu et al., 2012b], functional [Qiao et al., 2019, Tsai et al., 2024, Zhao et al., 2024], and Ising models [Kuang et al., 2017], and studying inferential methods within the transfer learning framework.

[2]The dataset is provided as the supplement to Kornblau et al. [2009].

References

Hamsa Bastani. Predicting with proxies: Transfer learning in high dimension. *Management Science*, 67(5):2964–2984, 2021.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2:183–202, 2009.

Michael J. Best and Robert R. Grauer. Positively weighted minimum-variance portfolios and the structure of asset expected returns. *Journal of Financial and Quantitative Analysis*, 27(4):513–537, 1992.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

Changxiao Cai, T Tony Cai, and Hongzhe Li. Transfer learning for contextual multi-armed bandits. *The Annals of Statistics*, 52(1):207–232, 2024a.

T. Tony Cai and Hongji Wei. Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *The Annals of Statistics*, 49(1):100 – 128, 2021.

T. Tony Cai, Weidong Liu, and Xi Luo. A constrained l1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

T. Tony Cai, Hongzhe Li, Weidong Liu, and Jichun Xie. Joint estimation of multiple high-dimensional precision matrices. *Statistica Sinica*, 26(2):445–464, 2016a.

T. Tony Cai, Weidong Liu, and Harrison H. Zhou. Estimating sparse precision matrix: optimal rates of convergence and adaptive estimation. *The Annals of Statistics*, 44(2):455–488, 2016b.

T Tony Cai, Dongwoo Kim, and Hongming Pu. Transfer learning for functional mean estimation: Phase transition and adaptive algorithms. *The Annals of Statistics*, 52(2):654–678, 2024b.

Patrick Danaher, Pei Wang, and Daniela M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 76(2):373–397, 2014.

Farideh Fazayeli and Arindam Banerjee. Generalized direct change estimation in ising model structure. In *International Conference on Machine Learning*, 2016.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Jiawei Ge, Shange Tang, Jianqing Fan, Cong Ma, and Chi Jin. Maximum likelihood estimation is all you need for well-specified covariate shift. In *International Conference on Learning Representations*, 2024.

Jian Guo, Elizaveta Levina, George Michailidis, and Ji Zhu. Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15, 2011.

Ehsan Hajiramezanali, Siamak Zamani Dadaneh, Alireza Karbalayghareh, Mingyuan Zhou, and Xiaoning Qian. Bayesian multi-domain learning for cancer subtype discovery from next-generation sequencing count data. *Advances in Neural Information Processing Systems*, 2018.

Steve Hanneke and Samory Kpotufe. On the value of target data in transfer learning. *Advances in Neural Information Processing Systems*, 2019.

Steve Hanneke and Samory Kpotufe. A no-free-lunch theorem for multitask learning. *The Annals of Statistics*, 50(6):3119–3143, 2022.

Zelin He, Ying Sun, and Runze Li. Transfusion: Covariate-shift robust transfer learning for high-dimensional regression. In *International Conference on Artificial Intelligence and Statistics*, pages 703–711, 2024.

Minoru Kanehisa, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. Kegg for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*, 40(D1):D109–D114, 2012.

Steven M Kornblau, Raoul Tibes, Yi Hua Qiu, Wenjing Chen, Hagop M Kantarjian, Michael Andreeff, Kevin R Coombes, and Gordon B Mills. Functional proteomic profiling of aml predicts response and survival. *Blood, The Journal of the American Society of Hematology*, 113(1):154–164, 2009.

Zhaobin Kuang, Sinong Geng, and David Page. A screening rule for l1-regularized ising model estimation. In *Advances in Neural Information Processing Systems*, 2017.

Clifford Lam and Jianqing Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of statistics*, 37(6B):4254, 2009.

Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

Wonyul Lee and Yufeng Liu. Joint estimation of multiple precision matrices with common structures. *The Journal of Machine Learning Research*, 16(1):1035–1062, 2015.

Sai Li, T Tony Cai, and Hongzhe Li. Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):149–173, 2022.

Sai Li, T Tony Cai, and Hongzhe Li. Transfer learning in large-scale gaussian graphical models with false discovery rate control. *Journal of the American Statistical Association*, 118(543):2171–2183, 2023a.

Sai Li, Linjun Zhang, T Tony Cai, and Hongzhe Li. Estimation and inference for high-dimensional generalized linear models with knowledge transfer. *Journal of the American Statistical Association*, 0(0):1–12, 2023b.

Haotian Lin and Matthew Reimherr. Transfer learning for functional linear regression with structural interpretability. *arXiv preprint arXiv:2206.04277*, 2022.

Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10(80):2295–2328, 2009.

Han Liu, Fang Han, Ming Yuan, John Lafferty, and Larry Wasserman. High-dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293 – 2326, 2012a.

Han Liu, Fang Han, and Cun-hui Zhang. Transelliptical graphical models. *Advances in neural information processing systems*, 25, 2012b.

Shuo Shuo Liu. Unified transfer learning models for high-dimensional linear regression. *arXiv preprint arXiv:2307.00238*, 2023.

Song Liu, John A Quinn, Michael U Gutmann, Taiji Suzuki, and Masashi Sugiyama. Direct learning of sparse changes in markov networks by density ratio estimation. *Neural computation*, 26(6): 1169–1197, 2014.

Cong Ma, Junwei Lu, and Han Liu. Inter-subject analysis: A partial gaussian graphical model approach. *Journal of the American Statistical Association*, 116(534):746–755, 2021.

Cong Ma, Reese Pathak, and Martin J Wainwright. Optimally tackling covariate shift in rkhs-based nonparametric regression. *The Annals of Statistics*, 51(2):738–761, 2023.

Jing Ma and George Michailidis. Joint structural estimation of multiple graphical models. *Journal of Machine Learning Research*, 17:166:1–166:48, 2016.

Karthik Mohan, Palma London, Maryam Fazel, Daniela Witten, and Su-In Lee. Node-based learning of multiple gaussian graphical models. *The Journal of Machine Learning Research*, 15(1):445–488, 2014.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer New York, 2006.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

Neal Parikh and Stephen P. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, 2014.

Reese Pathak, Cong Ma, and Martin Wainwright. A new similarity measure for covariate shift with applications to nonparametric regression. In *International Conference on Machine Learning*, 2022.

Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. Technical report, Technical University of Denmark, 2008.

Christine Peterson, Francesco C Stingo, and Marina Vannucci. Bayesian inference of multiple gaussian graphical models. *Journal of the American Statistical Association*, 110(509):159–174, 2015.

Xinghao Qiao, Shaojun Guo, and Gareth M James. Functional graphical models. *Journal of the American Statistical Association*, 114(525):211–222, 2019.

Pradeep Ravikumar, Martin J. Wainwright, Garvesh Raskutti, and Bin Yu. High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935 – 980, 2011.

Zhao Ren, Tingni Sun, Cun-Hui Zhang, and Harrison H. Zhou. Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *The Annals of Statistics*, 43(3):991–1026, 2015.

Adam J. Rothman, Peter J. Bickel, Elizaveta Levina, and Ji Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494 – 515, 2008.

Takumi Saegusa and Ali Shojaie. Joint estimation of precision matrices in heterogeneous populations. *Electronic Journal of Statistics*, 10(1):1341, 2016.

George Saon and Jen-Tzung Chien. Bayesian sensing hidden markov models for speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2011.

Ye Tian and Yang Feng. Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association*, 118(544):2684–2697, 2023.

Katherine Tsai, Oluwasanmi Koyejo, and Mladen Kolar. Joint gaussian graphical model estimation: A survey. *Wiley Interdisciplinary Reviews: Computational Statistics*, 14(6):e1582, 2022.

Katherine Tsai, Boxin Zhao, Sanmi Koyejo, and Mladen Kolar. Latent multimodal functional graphical model estimation. *Journal of the American Statistical Association*, 0(0):1–13, 2024.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer New York, NY, 1st edition, 2008.

Jitendra K Tugnait. Estimation of high-dimensional differential graphs from multi-attribute data. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023.

Turki Turki, Zhi Wei, and Jason TL Wang. Transfer learning approaches to improve drug sensitivity prediction in multiple myeloma patients. *IEEE Access*, 5:7381–7393, 2017.

Flora Tzifi, Christina Economopoulou, Dimitrios Gourgiotis, Alexandros Ardavanis, Sotirios Papageorgiou, and Andreas Scorilas. The role of bcl2 family of apoptosis regulator proteins in acute and chronic leukemias. *Advances in Hematology*, 2012, 2012.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Kaizheng Wang. Pseudo-labeling for kernel ridge regression under covariate shift. *arXiv preprint arXiv:2302.10160*, 2023.

Huili Yuan, Ruibin Xi, Chong Chen, and Minghua Deng. Differential network analysis via lasso penalized d-trace loss. *Biometrika*, 104(4):755–770, 2017.

Ming Yuan and Yi Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

Boxin Zhao, Y. Samuel Wang, and Mladen Kolar. Direct estimation of differential functional graphical models. In *Advances in Neural Information Processing Systems*, 2019.

Boxin Zhao, Y. Samuel Wang, and Mladen Kolar. Fudge: A method to estimate a functional differential graph in a high-dimensional setting. *Journal of Machine Learning Research*, 23:82:1–82:82, 2022.

Boxin Zhao, Percy S Zhai, Y Samuel Wang, and Mladen Kolar. High-dimensional functional graphical model structure learning via neighborhood selection approach. *Electronic Journal of Statistics*, 18(1):1042–1129, 2024.

Sihai Dave Zhao, T Tony Cai, and Hongzhe Li. Direct estimation of differential networks. *Biometrika*, 101(2):253–268, 2014.

Yunzhang Zhu, Xiaotong Shen, and Wei Pan. Structural pursuit over multiple undirected graphs. *Journal of the American Statistical Association*, 109(508):1683–1696, 2014.

A Comparison with Li et al. [2023a]

In this section, we make a more detailed comparison of the similarity assumption made in Li et al. [2023a] and our Assumption 1. Let $\Upsilon^{(k)} = \left(\Omega^{(0)} - \Omega^{(k)}\right)\Sigma^{(k)} = -\Psi^{(k)}\Sigma^{(k)}$, or equivalently $\Psi^{(k)} = -\Upsilon^{(k)}\Omega^{(k)}$. Li et al. [2023a] assumes $\Upsilon^{(k)}$ is column-wise sparse in L_p-norm. Our Assumption 1 and the similarity assumption in Li et al. [2023a] do not imply each other and are generally incomparable. However, our Assumption 1 can be preferable in some applications. First, while the divergence matrix $\Upsilon^{(k)}$ is motivated by the KL divergence between Gaussian distributions, Assumption 1 is structural and applies to any distribution. Second, Assumption 1 is naturally interpreted in Gaussian graphical models, unlike the similarity assumption in Li et al. [2023a]. Finally, a technical advantage of our assumption is that, while sparsity in $\Psi^{(k)}$ does not generally imply sparsity in $\Upsilon^{(k)}$, or vice versa, further assumptions on $\Sigma^{(k)}$ or $\Omega^{(k)}$ can establish one or both directions. For example, note that

$$\left|\Psi^{(k)}\right|_0 \leqslant \sum_{i,j} \mathbb{1}\left\{\sum_{l=1}^{d} \mathbb{1}\left\{\Upsilon_{il}^{(k)} \neq 0\right\} \mathbb{1}\left\{\Omega_{lj}^{(k)} \neq 0\right\} \geqslant 1\right\}.$$

If both $\Upsilon^{(k)}$ and $\Omega^{(k)}$ are sparse, then $\Psi^{(k)}$ is also sparse. Similarly, if $\Psi^{(k)}$ and $\Sigma^{(k)}$ are sparse, then $\Upsilon^{(k)}$ is sparse. The key difference is that $\Upsilon^{(k)} \implies \Psi^{(k)}$ relies on $\Omega^{(k)}$ being sparse, while $\Psi^{(k)} \implies \Upsilon^{(k)}$ relies on $\Sigma^{(k)}$ being sparse. In graphical models, sparsity assumptions on the precision matrix are more common than on the covariance matrix, making the sparsity assumptions on $\Psi^{(k)}$ weaker than those on $\Upsilon^{(k)}$.

B Preliminary Lemmas

We first collect several inequalities related to matrix norms.

Lemma 1. *For any two matrices $A, B \in \mathbb{R}^{d\times d}$, we have*

$$|AB|_\infty \leqslant \min\left\{\|A\|_\infty |B|_\infty, \|B\|_1 |A|_\infty\right\}. \tag{25}$$

Proof. For $1 \leqslant j \leqslant d$, denote by $B_{:j}$ the j-th column of B. We then have

$$\begin{aligned}
|AB|_\infty &= \max_{1\leqslant j\leqslant d} \|AB_{:j}\|_\infty \\
&\leqslant \max_{1\leqslant j\leqslant d} \|A\|_\infty \|B_{:j}\|_\infty \\
&= \|A\|_\infty \max_{1\leqslant j\leqslant d} \|B_{:j}\|_\infty \\
&= \|A\|_\infty |B|_\infty.
\end{aligned}$$

The other claim follows from the facts that $|AB|_\infty = |B^\top A^\top|_\infty$, and $\|B^\top\|_\infty = \|B\|_1$. ∎

Lemma 2. *For any matrix $A \in \mathbb{R}^{m\times n}$ and any vector $v \in \mathbb{R}^n$, we have $\|Av\|_\infty \leqslant |A|_\infty \|v\|_1$.*

Proof. Let $A_{i\cdot}$ denote the i-th row of A, we then have

$$\|Av\|_\infty = \max_{1\leqslant i\leqslant m} |\langle A_{i\cdot}, v\rangle| \leqslant \max_{1\leqslant i\leqslant m} \|A_{i\cdot}\|_\infty \|v\|_1 = |A|_\infty \|v\|_1,$$

which completes the proof. ∎

The next lemma requires the following definition of the sub-Gaussian random variable.

Definition B.1 (Sub-Gaussian Random Variable, Definition 2 of Ravikumar et al. [2011])**.** We say that a random variable $X \in \mathbb{R}$ is sub-Gaussian with parameter σ if

$$\mathbb{E}\left[\exp\left(\lambda X\right)\right] \leqslant \exp\left(\sigma^2\lambda/2\right) \quad \text{for all } \lambda \in \mathbb{R}.$$

Lemma 3. *Assume that we obtain samples from K distributions, each with a mean of zero and a covariance matrix of $\Sigma^{(k)}$. Let $\{X_i^{(k)} = (X_{i1}^{(k)}, \ldots, X_{id}^{(k)})^\top\}_{i=1}^{n_k}$ represent n_k independently distributed samples from the k-th distribution. In addition, we assume that $X_{ij}^{(k)}/\sqrt{\Sigma_{jj}^{(k)}}$ is sub-Gaussian with parameter σ as defined in Definition B.1. Let*

$$\widehat{\Sigma}^{(k)} = \frac{1}{n_k}\sum_{i=1}^{n_k} X_i^{(k)} X_i^{(k)\top}, \quad \widehat{\Sigma} = \sum_{k=1}^{K} \alpha_k \widehat{\Sigma}^{(k)}, \quad \text{and} \quad \Sigma^\star = \sum_{k=1}^{K} \alpha_k \Sigma^{(k)},$$

where $\alpha_k = n_k/N$ and $N = \sum_{k=1}^K n_k$. For a fixed $\delta \in (0,1]$, we have

$$\mathbb{P}\left\{\left|\widehat{\Sigma} - \Sigma^\star\right|_\infty > 16(1+4\sigma^2)\left\{\max_{1\leqslant k\leqslant K}\left|\Sigma^{(k)}\right|_\infty\right\}\max\left\{\sqrt{\frac{\log(2d^2/\delta)}{2N}}, \frac{\log(2d^2/\delta)}{N}\right\}\right\} \leqslant \delta.$$

Proof. Let $\bar{X}_{ij}^{(k)} = X_{ij}^{(k)}/\sqrt{\Sigma_{jj}^{(k)}}$, $i = 1, \ldots, n_k$, $k = 1, \ldots, K$, $j = 1, \ldots, d$. Define $U_{i,jl}^{(k)} = \bar{X}_{ij}^{(k)} + \bar{X}_{il}^{(k)}$, $V_{i,jl}^{(k)} = \bar{X}_{ij}^{(k)} - \bar{X}_{il}^{(k)}$, and $\rho_{jl}^{(k)} = \Sigma_{jl}^{(k)}/\sqrt{\Sigma_{jj}^{(k)} \cdot \Sigma_{ll}^{(k)}}$. Then it follows from Proposition 2.9 and (2.18) in Wainwright [2019], that

$$\mathbb{E}\left[\exp\left\{\lambda\left[\left(U_{i,jl}^{(k)}\right)^2 - 2\left(1+\rho_{jl}^{(k)}\right)\right]\right\}\right] \leqslant \exp\left(\frac{\lambda^2\nu^2}{2}\right), \quad \text{and}$$
$$\mathbb{E}\left[\exp\left\{\lambda\left[\left(V_{i,jl}^{(k)}\right)^2 - 2\left(1-\rho_{jl}^{(k)}\right)\right]\right\}\right] \leqslant \exp\left(\frac{\lambda^2\nu^2}{2}\right),$$

for all $|\lambda| < 1/\phi$ where $\nu = \phi = 16(1+4\sigma^2)$. Since

$$\sum_{i=1}^{n_k}\left(\bar{X}_{ij}^{(k)}\bar{X}_{il}^{(k)} - \rho_{jl}^{(k)}\right) = \frac{1}{4}\sum_{i=1}^{n_k}\left(\left(U_{i,jl}^{(k)}\right)^2 - 2\left(1+\rho_{jl}^{(k)}\right)\right) - \frac{1}{4}\sum_{i=1}^{n_k}\left(\left(V_{i,jl}^{(k)}\right)^2 - 2\left(1-\rho_{jl}^{(k)}\right)\right),$$

we have

$$\begin{aligned}
&\mathbb{E}\left[\exp\left\{\lambda\left[\sum_{i=1}^{n_k}\left(\bar{X}_{ij}^{(k)}\bar{X}_{il}^{(k)} - \rho_{jl}^{(k)}\right)\right]\right\}\right] \\
&= \mathbb{E}\left[\exp\left\{\frac{\lambda}{4}\left[\sum_{i=1}^{n_k}\left(\left(U_{i,jl}^{(k)}\right)^2 - 2\left(1+\rho_{jl}^{(k)}\right)\right)\right]\right\}\cdot\exp\left\{-\frac{\lambda}{4}\left[\sum_{i=1}^{n_k}\left(\left(V_{i,jl}^{(k)}\right)^2 - 2\left(1-\rho_{jl}^{(k)}\right)\right)\right]\right\}\right] \\
&\leqslant \left\{\mathbb{E}\left[\exp\left\{\frac{\lambda}{2}\left[\sum_{i=1}^{n_k}\left(\left(U_{i,jl}^{(k)}\right)^2 - 2\left(1+\rho_{jl}^{(k)}\right)\right)\right]\right\}\right]\right\}^{\frac{1}{2}} \times \\
&\quad \left\{\mathbb{E}\left[\exp\left\{-\frac{\lambda}{2}\left[\sum_{i=1}^{n_k}\left(\left(V_{i,jl}^{(k)}\right)^2 - 2\left(1-\rho_{jl}^{(k)}\right)\right)\right]\right\}\right]\right\}^{\frac{1}{2}} \\
&\leqslant \exp\left(\frac{n_k\nu^2\lambda^2}{8}\right) \quad \text{for all } |\lambda| < \frac{2}{\phi}.
\end{aligned}$$

Let $\tau_\infty = \max_{1\leqslant k\leqslant K}|\Sigma^{(k)}|_\infty$. It then follows that

$$\begin{aligned}
&\mathbb{E}\left[\exp\left\{\lambda\left[\sum_{i=1}^{n_k}\left(X_{ij}^{(k)}X_{il}^{(k)} - \Sigma_{jl}^{(k)}\right)\right]\right\}\right] \\
&\quad = \mathbb{E}\left[\exp\left\{\lambda\sqrt{\Sigma_{jj}^{(k)}}\sqrt{\Sigma_{ll}^{(k)}}\left[\sum_{i=1}^{n_k}\left(\bar{X}_{ij}^{(k)}\bar{X}_{il}^{(k)} - \rho_{jl}^{(k)}\right)\right]\right\}\right]
\end{aligned}$$

$$
\begin{aligned}
&\leqslant \mathbb{E}\left[\exp\left\{\lambda|\Sigma^{(k)}|_\infty\left[\sum_{i=1}^{n_k}\left(\bar{X}_{ij}^{(k)}\bar{X}_{il}^{(k)}-\rho_{jl}^{(k)}\right)\right]\right\}\right]\\
&\leqslant \mathbb{E}\left[\exp\left\{\lambda\tau_\infty\left[\sum_{i=1}^{n_k}\left(\bar{X}_{ij}^{(k)}\bar{X}_{il}^{(k)}-\rho_{jl}^{(k)}\right)\right]\right\}\right]\\
&\leqslant \exp\left(\frac{n_k\nu^2\tau_\infty^2\lambda^2}{8}\right) \quad \text{for all } |\lambda|<\frac{2}{\phi\tau_\infty},
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{E}\left[\exp\left\{\lambda N\left(\widehat{\Sigma}_{jl}-\Sigma_{jl}^\star\right)\right\}\right] &= \mathbb{E}\left[\exp\left\{\lambda\left(\sum_{k=1}^{K}\sum_{i=1}^{n_k}\left(X_{ij}^{(k)}X_{il}^{(k)}-\Sigma_{jl}^{(k)}\right)\right)\right\}\right]\\
&= \prod_{k=1}^{K}\mathbb{E}\left[\exp\left\{\lambda\left(\sum_{i=1}^{n_k}\left(X_{ij}^{(k)}X_{il}^{(k)}-\Sigma_{jl}^{(k)}\right)\right)\right\}\right]\\
&\leqslant \prod_{k=1}^{K}\exp\left(\frac{n_k\nu^2\tau_\infty^2\lambda^2}{8}\right)\\
&= \exp\left(\frac{N\nu^2\tau_\infty^2\lambda^2}{8}\right)
\end{aligned}
$$

when $|\lambda|<2/(\tau_\infty\phi)$. Using Proposition 2.9 in Wainwright [2019], for $t\geqslant 0$, we have that

$$
\mathbb{P}\left\{\left|\widehat{\Sigma}_{jl}-\Sigma_{jl}^\star\right|>t\right\} = \mathbb{P}\left\{\left|N\left(\widehat{\Sigma}_{jl}-\Sigma_{jl}^\star\right)\right|>Nt\right\} \leqslant 2\exp\left\{-\frac{1}{2}\min\left(\frac{4Nt^2}{\nu^2\tau_\infty^2},\frac{2Nt}{\phi\tau_\infty}\right)\right\}.
$$

A union bound then gives us

$$
\mathbb{P}\left\{\left|\widehat{\Sigma}-\Sigma^\star\right|_\infty>t\right\} \leqslant 2d^2\exp\left\{-\frac{1}{2}\min\left(\frac{4Nt^2}{\nu^2\tau_\infty^2},\frac{2Nt}{\phi\tau_\infty}\right)\right\}.
$$

Rewriting the above equation, we complete the proof. ■

Lemma 4. *Consider a zero mean random vector $X=(X_1,\ldots,X_d)^\top$ with covariance $\Sigma^\star$ such that each $X_j/\sqrt{\Sigma_{jj}^\star}$ is sub-Gaussian with parameter σ as defined in Definition B.1. Let $\{X(i)\}_{i=1}^n$ be n i.i.d. copies of X and let $\widehat{\Sigma}_n=\frac{1}{n}\sum_{i=1}^n X(i)X(i)^\top$. For $\delta\in(0,1]$, we have*

$$
\mathbb{P}\left\{\left|\widehat{\Sigma}_n-\Sigma^\star\right|_\infty>16(1+4\sigma^2)\left|\Sigma^\star\right|_\infty\max\left\{\sqrt{\frac{\log(2d^2/\delta)}{2n}},\frac{\log(2d^2/\delta)}{n}\right\}\right\}\leqslant\delta.
$$

Proof. The proof follows directly from Lemma 3. ■

In the next lemma, we use $\mathbb{S}_+^d$ to denote the set of positive definite matrices with dimension d.

Lemma 5. *Let $A\in\mathbb{S}_+^d$. Define $f(\Omega):=\langle\Omega,A\rangle-\log\det\Omega$ for $\Omega\in\mathbb{S}_+^d$. Given $\Omega_0\in\mathbb{S}_+^d$ and $\Delta_\Omega\in\mathbb{R}^{d\times d}$ such that $\Omega_0+\Delta_\Omega\in\mathbb{S}_+^d$, there exists $t\in(0,1)$ such that*

$$
f(\Omega_0+\Delta_\Omega)-f(\Omega_0)-\langle\nabla f(\Omega_0),\Delta_\Omega\rangle=\frac{1}{2}\mathrm{vec}(\Delta_\Omega)^\top\left\{(\Omega_0+t\Delta_\Omega)^{-1}\otimes(\Omega_0+t\Delta_\Omega)^{-1}\right\}\mathrm{vec}(\Delta_\Omega)
$$

Proof. One can view f as a function of $\mathrm{vec}(\Omega)$. Note that $\nabla^2 f(\mathrm{vec}(\Omega))=\Omega^{-1}\otimes\Omega^{-1}$. The rest of the proof then follows Taylor's Theorem [Nocedal and Wright, 2006, Theorem 2.1]. ■

Lemma 6. *Let* $\mathbb{S}^p(s) := \{\theta \in \mathbb{R}^p : \|\theta\|_0 \leqslant s, \|\theta\|_2 \leqslant 1\}$. *There exists* $\{\theta^0, \theta^1, \ldots, \theta^M\} \subseteq \mathbb{S}^p(s)$ *such that*

(i) $\theta^0 = 0$;

(ii) $\left\|\theta^j - \theta^k\right\|_2 \geqslant \frac{1}{4}$ *for all* $0 \leqslant j \neq k \leqslant M$;

(iii) $\log(M+1) \geqslant \frac{s}{2} \log\left(\frac{d-s}{s}\right)$.

Proof. It follows from Example 15.16 in Wainwright [2019] that there exists a 1/2-packing of $\mathbb{S}^p(s)$, which we denote as $\{\tilde{\theta}^0, \tilde{\theta}^1, \ldots, \tilde{\theta}^M\}$, such that

$$\left\|\tilde{\theta}^j - \tilde{\theta}^k\right\|_2 \geqslant \frac{1}{2} \qquad \text{for all } 0 \leqslant j \neq k \leqslant M \qquad \text{and} \qquad \log(M+1) \geqslant \frac{s}{2} \log\left(\frac{d-s}{s}\right). \tag{26}$$

Without loss of generality, assume that $\|\tilde{\theta}^0\|_2 \leqslant \|\tilde{\theta}^j\|_2$ for $1 \leqslant j \leqslant M$. Let $\theta^0 = 0$ and let $\theta^j = \tilde{\theta}^j$ for $1 \leqslant j \leqslant M$. Then the set $\{\theta^0, \theta^1, \ldots, \theta^M\}$ satisfies (i)–(iii). To prove the claim, by our construction of the set and (26), we only need to verify that $\|\theta^j - \theta^0\|_2 = \|\theta^j\|_2 \geqslant \frac{1}{4}$ for all $1 \leqslant j \leqslant M$. We prove the result in two cases.

Case 1: $\|\tilde{\theta}^0\|_2 \geqslant \frac{1}{4}$. Since $\|\tilde{\theta}^0\|_2 \leqslant \|\tilde{\theta}^j\|_2$, we have that $\|\theta^j\|_2 = \|\tilde{\theta}^j\|_2 \geqslant \|\tilde{\theta}^0\|_2 \geqslant \frac{1}{4}$ for $1 \leqslant j \leqslant M$.

Case 2: $\|\tilde{\theta}^0\|_2 < \frac{1}{4}$. We have

$$\left\|\theta^j\right\|_2 = \left\|\tilde{\theta}^j\right\|_2 \geqslant \left\|\tilde{\theta}^j - \tilde{\theta}^0\right\|_2 - \left\|\tilde{\theta}^0\right\|_2 > \frac{1}{2} - \frac{1}{4} = \frac{1}{4} \quad \text{for } 1 \leqslant j \leqslant M,$$

where the last inequality follows from (26).

Combining Case 1 and Case 2, we have proved that $\|\theta^j\|_2 \geqslant \frac{1}{4}$ for $1 \leqslant j \leqslant M$, which completes the proof. ■

Lemma 7. *Let* $m \geqslant 2$ *be an even integer and set* $d = 2m$. *Let* $B \in \mathbb{R}^{m \times m}$ *be such that* $\|B\|_{\mathrm{F}} \leqslant 1/4$ *and define*

$$\Omega = \begin{bmatrix} I_m & B \\ B^\top & I_m \end{bmatrix}.$$

Then $\Omega \succ 0$ *and* $D_{KL}\left(N\left(0, \Omega^{-1}\right) \| N\left(0, I_d\right)\right) \leqslant \frac{16}{15} \|B\|_{\mathrm{F}}^2$.

Proof. We begin with proving $\Omega \succ 0$. By Weyl's inequality, we know that

$$\begin{aligned} \gamma_{\min}(\Omega) &\geqslant 1 - \left\|\begin{bmatrix} 0 & B \\ B^\top & 0 \end{bmatrix}\right\| \geqslant 1 - \left\|\begin{bmatrix} 0 & B \\ B^\top & 0 \end{bmatrix}\right\|_{\mathrm{F}} \\ &= 1 - \sqrt{2}\|B\|_{\mathrm{F}} \geqslant \frac{1}{2}. \end{aligned}$$

As a result, we have $\Omega \succ 0$.

Now we move on to the second claim regarding the KL divergence. Recall that

$$D_{\mathrm{KL}}\left(N\left(0, \Omega^{-1}\right) \| N\left(0, I_d\right)\right) = \frac{1}{2}\left[\log\det(\Omega) - d + \operatorname{tr}\left(\Omega^{-1}\right)\right], \tag{27}$$

which motivates us to compute $\log\det(\Omega)$ and $\operatorname{tr}\left(\Omega^{-1}\right)$. Let $B = UDV^\top$ be the singular value decomposition of B with $U, D, V \in \mathbb{R}^{m \times m}$, $U^\top U = UU^\top = I_m$, $V^\top V = VV^\top = I_m$, and $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$. We have the following identities.

$$\log\det(\Omega) = \sum_{i=1}^{m} \log\left(1 - \lambda_i^2\right). \tag{28a}$$

$$\operatorname{tr}\left(\Omega^{-1}\right) = \sum_{i=1}^{m} \frac{2}{1-\lambda_i^2}. \tag{28b}$$

Combining (27), (28a), and (28b), we obtain

$$\begin{aligned} D_{\mathrm{KL}}\left(N\left(0,\Omega^{-1}\right) \| N\left(0, I_d\right)\right) &= \frac{1}{2}\left[\sum_{i=1}^{m} \log\left(1-\lambda_i^2\right) + \sum_{i=1}^{m} \frac{2}{1-\lambda_i^2} - d\right] \\ &= \sum_{i=1}^{m}\left[\frac{\log\left(1-\lambda_i^2\right)}{2} + \frac{1}{1-\lambda_i^2} - 1\right]. \end{aligned}$$

Since $\lambda_j^2 \leqslant \sum_{i=1}^{d} \lambda_i^2 = \|B\|_{\mathrm{F}}^2 \leqslant 1/16$, we have $-\lambda_j^2 \geqslant -1/16$. Also note that $\log(1+x) \leqslant x$ for all $x \geqslant -1$. We thus have

$$\begin{aligned} D_{\mathrm{KL}}\left(N\left(0,\Omega^{-1}\right) \| N\left(0, I_d\right)\right) &\leqslant \sum_{i=1}^{m}\left[\frac{-\lambda_i^2}{2} + \frac{1}{1-\lambda_i^2} - 1\right] \\ &= \sum_{i=1}^{m} \frac{\lambda_i^2 + \lambda_i^4}{2\left(1-\lambda_i^2\right)} \leqslant \sum_{i=1}^{m} \frac{\lambda_i^2}{1-\lambda_i^2} \leqslant \frac{16}{15} \sum_{i=1}^{m} \lambda_i^2 = \frac{16}{15}\|B\|_{\mathrm{F}}^2. \end{aligned}$$

This completes the proof.

Proof of Equation (28a). Using Section 9.1.2 of Petersen et al. [2008], we have $\det(\Omega) = \det\left(I_m - B^\top B\right)$. Since

$$I_m - B^\top B = I_m - V D^2 V^\top = V\left(I_m - D^2\right) V^\top, \tag{29}$$

we have $\det(\Omega) = \det\left(I_m - D^2\right) = \prod_{i=1}^{m}\left(1-\lambda_i^2\right)$ and hence $\log\det(\Omega) = \sum_{i=1}^{m} \log\left(1-\lambda_i^2\right)$.

Proof of Equation (28b). Using Section 9.1.3 of Petersen et al. [2008], we have

$$\Omega^{-1} = \begin{bmatrix} I_m & B \\ B^\top & I_m \end{bmatrix}^{-1} = \begin{bmatrix} \left(I_m - BB^\top\right)^{-1} & -B\left(I_m - B^\top B\right)^{-1} \\ -\left(I_m - B^\top B\right)^{-1} B^\top & \left(I_m - B^\top B\right)^{-1} \end{bmatrix},$$

which implies

$$\operatorname{tr}\left(\Omega^{-1}\right) = \operatorname{tr}\left\{\left(I_m - BB^\top\right)^{-1}\right\} + \operatorname{tr}\left\{\left(I_m - B^\top B\right)^{-1}\right\}. \tag{30}$$

It follows from (29) that $\left(I_m - B^\top B\right)^{-1} = V\left(I_m - D^2\right)^{-1} V^\top$, and therefore,

$$\operatorname{tr}\left\{\left(I_m - BB^\top\right)^{-1}\right\} = \operatorname{tr}\left\{V\left(I_m - D^2\right)^{-1} V^\top\right\} = \operatorname{tr}\left\{\left(I_m - D^2\right)^{-1}\right\} = \sum_{i=1}^{m} \frac{1}{1-\lambda_i^2}. \tag{31}$$

Similarly, we have $\operatorname{tr}\{\left(I_m - B^\top B\right)^{-1}\} = \sum_{i=1}^{m} \frac{1}{1-\lambda_i^2}$. Take the previous results collectively to yield the claim (28b). ■

C Proof of Theorem 1

We adopt the proof strategy laid out in Wainwright [2019, Chapter 9].

C.1 Additional Notation

We first introduce some additional notation that will be helpful in the proof. We define the Hilbert space for the parameters

$$\mathbb{H} := \left\{ \Theta = \left(\Omega, \left\{ \Gamma^{(k)} \right\}_{k=0}^{K} \right) : \Omega, \Gamma^{(k)} \in \mathbb{R}^{d \times d} \text{ for all } 0 \leqslant k \leqslant K \right\},$$

with the associated inner product

$$\left\langle \Theta, \Theta' \right\rangle_{\mathbb{H}} = \left\langle \Theta, \Theta' \right\rangle := \left\langle \Omega, \Omega' \right\rangle + \sum_{k=0}^{K} \left\langle \Gamma^{(k)}, \Gamma^{(k)\prime} \right\rangle.$$

The space $\mathbb{H}$ endowed with the inner product $\langle \cdot, \cdot \rangle$ is indeed a Hilbert space. Correspondingly, we have

$$\|\Theta\|_{\mathbb{H}}^2 = \|\Omega\|_{\mathrm{F}}^2 + \sum_{k=0}^{K} \left\| \Gamma^{(k)} \right\|_{\mathrm{F}}^2.$$

We also need the dual norm of $\Phi(\cdot)$, which is defined as

$$\Phi^\star(\Theta) := \sup_{\Theta' : \Phi(\Theta') \leqslant 1} \left\langle \Theta, \Theta' \right\rangle. \tag{32}$$

Recall that $\Theta^\star = (\Omega^\star, \{\Gamma^{(k)\star}\}_{k=0}^K)$ is the true parameter. Let $\mathcal{S}_\Omega$ and $\mathcal{S}_{\Gamma^{(k)}}$ be the supports of $\Omega^\star$ and $\Gamma^{(k)\star}$ for all $0 \leqslant k \leqslant K$, respectively. Under Assumption 1, the true parameter $\Theta^\star$ lies in the following subspace of $\mathbb{H}$:

$$\mathbb{M} := \left\{ \Theta = \left(\Omega, \left\{ \Gamma^{(k)} \right\}_{k=0}^{K} \right) : \operatorname{supp}(\Omega) \subseteq \mathcal{S}_\Omega, \operatorname{supp}\left[\Gamma^{(k)} \right] \subseteq \mathcal{S}_{\Gamma^{(k)}} \text{ for all } 0 \leqslant k \leqslant K \right\}. \tag{33}$$

The orthogonal complement of $\mathbb{M}$ is given by

$$\mathbb{M}^\perp := \left\{ \Theta = \left(\Omega, \left\{ \Gamma^{(k)} \right\}_{k=0}^{K} \right) : \operatorname{supp}(\Omega) \subseteq \mathcal{S}_\Omega^c, \operatorname{supp}\left[\Gamma^{(k)} \right] \subseteq \mathcal{S}_{\Gamma^{(k)}}^c \text{ for all } 0 \leqslant k \leqslant K \right\}. \tag{34}$$

Clearly, for any $\Theta \in \mathbb{M}$ and $\Theta' \in \mathbb{M}^\perp$, we have that $\langle \Theta, \Theta' \rangle = 0$.

We also need to define the projection of a parameter onto a subspace. For any matrix $B \in \mathbb{R}^{d \times d}$ and any subspace $\mathcal{F} \subseteq \mathbb{R}^{d \times d}$, we define

$$[B]_{\mathcal{F}} = \arg\min_{\tilde{B} \in \mathcal{F}} \left\| \tilde{B} - B \right\|_{\mathrm{F}}.$$

Similarly, for any $\Theta \in \mathbb{H}$ and any subspace $\mathbb{F} \subseteq \mathbb{H}$, we define

$$[\Theta]_{\mathbb{F}} = \arg\min_{\tilde{B} \in \mathbb{F}} \left\| \tilde{B} - B \right\|_{\mathbb{H}}.$$

In addition, for $S \subseteq [d] \times [d]$, we define

$$\mathcal{M}(S) := \left\{ B \in \mathbb{R}^{d \times d} : \operatorname{supp}(B) \subseteq S \right\}.$$

For any $B \in \mathbb{R}^{d \times d}$, we define $[B]_S := [B]_{\mathcal{M}(S)}$. This way, for any $\Theta = (\Omega, \{\Gamma^{(k)}\}_{k=0}^K) \in \mathbb{H}$, we have

$$[\Theta]_{\mathbb{M}} = \left([\Omega]_{\mathcal{S}_\Omega}, \left\{ \left[\Gamma^{(k)} \right]_{\mathcal{S}_{\Gamma^{(k)}}} \right\}_{k=0}^{K} \right) \qquad \text{and} \qquad [\Theta]_{\mathbb{M}^\perp} = \left([\Omega]_{\mathcal{S}_\Omega^c}, \left\{ \left[\Gamma^{(k)} \right]_{\mathcal{S}_{\Gamma^{(k)}}^c} \right\}_{k=0}^{K} \right).$$

Finally, we define our metric on the estimation error. For $\Theta = (\Omega, \{\Gamma^{(k)}\}_{k=0}^K) \in \mathbb{H}$, let

$$H(\Theta) := \sum_{k=0}^{K} \alpha_k \left\| \Omega + \Gamma^{(k)} \right\|_{\mathrm{F}}^2. \tag{35}$$

We define

$$\widehat{\Delta} := \widehat{\Theta} - \Theta^\star = \left(\widehat{\Omega} - \Omega^\star, \left\{ \widehat{\Gamma}^{(k)} - \Gamma^{(k)\star} \right\}_{k=0}^K \right),$$

where $\widehat{\Theta}$ is the Trans-Glasso estimator (3). Then our goal is to control

$$H(\widehat{\Delta}) = \sum_{k=0}^{K} \alpha_k \left\| \widehat{\Omega} - \Omega^\star + \widehat{\Gamma}^{(k)} - \Gamma^{(k)\star} \right\|_{\mathrm{F}}^2 = \sum_{k=0}^{K} \alpha_k \left\| \breve{\Omega}^{(k)} - \Omega^{(k)} \right\|_{\mathrm{F}}^2.$$

C.2 Useful lemmas

Now we collect several useful lemmas, whose proofs are deferred to the end of the section.

We begin by demonstrating that $\Phi(\cdot)$, defined in (5), is decomposable with respect to $(\mathbb{M}, \mathbb{M}^\perp)$.

Lemma 8. *We have*

$$\Phi(\Theta + \Theta') = \Phi(\Theta) + \Phi(\Theta') \quad \textit{for any } \Theta \in \mathbb{M},\ \Theta' \in \mathbb{M}^\perp.$$

The following lemma relates $\Phi(\Theta)$ with $H(\Theta)$ defined in (35) for any $\Theta \in \mathbb{M}$.

Lemma 9. *For any $\Theta \in \mathbb{M}$, we have*

$$\Phi(\Theta) \leqslant \sqrt{2}\sqrt{s + (K+1)h}\sqrt{H(\Theta)}.$$

For $\Delta_\Theta = (\Delta_\Omega, \{\Delta_{\Gamma^{(k)}}\}_{k=0}^K) \in \mathbb{H}$ such that $\Delta_\Omega + \Delta_{\Gamma^{(k)}} + \Omega^{(k)} \succ 0$ for all $0 \leqslant k \leqslant K$, we define

$$\mathcal{R}\left(\Delta_\Theta\right) := \mathcal{L}\left(\Theta^\star + \Delta_\Theta\right) - \mathcal{L}\left(\Theta^\star\right) - \left\langle \nabla \mathcal{L}\left(\Theta^\star\right), \Delta_\Theta \right\rangle$$

to be the residual of $\mathcal{L}(\cdot)$ around $\Theta^\star$, where $\mathcal{L}(\cdot)$ is defined in (4). The following lemma claims that $\mathcal{R}(\cdot)$ is locally strongly convex with respect to the geometry defined by $H(\cdot)$.

Lemma 10. *Let $\Delta_\Theta = (\Delta_\Omega, \{\Delta_{\Gamma^{(k)}}\}_{k=0}^K) \in \mathbb{H}$ be such that $\Delta_\Omega + \Delta_{\Gamma^{(k)}} + \Omega^{(k)} \succ 0$ and assume that $\|\Delta_\Omega + \Delta_{\Gamma^{(k)}}\|_2 \leqslant M_{\mathrm{op}} + M_\Omega$ for all $0 \leqslant k \leqslant K$. Then, we have*

$$\mathcal{R}\left(\Delta_\Theta\right) \geqslant \frac{\kappa}{2} H\left(\Delta_\Theta\right),$$

where $\kappa = (2M_\Omega + M_{\mathrm{op}})^{-2}$.

Last but not least, define the event

$$\mathbb{G}(\lambda_{\mathrm{M}}) := \left\{ \frac{\lambda_{\mathrm{M}}}{2} \geqslant \Phi^\star\left(\nabla \mathcal{L}\left(\Theta^\star\right)\right) \right\}. \tag{36}$$

The following lemma asserts that this event happens with high probability.

Lemma 11. *When (18) holds and λ_M satisfies (19), we have that $\mathbb{P}\{\mathbb{G}(\lambda_M)\} \geqslant 1 - \delta$.*

C.3 Remaining proof

Now we are ready to prove Theorem 1. As Lemma 11 asserts that $\mathbb{G}(\lambda_{\mathrm{M}})$ holds with high probability, thus we only need to prove the conclusion under the assumption that $\mathbb{G}(\lambda_{\mathrm{M}})$ holds. Throughout the proof, we assume the event $\mathbb{G}(\lambda_{\mathrm{M}})$ holds.

For any $\Delta_\Theta = (\Delta_\Omega, \{\Delta_{\Gamma^{(k)}}\}_{k=0}^K) \in \mathbb{H}$, we define the objective difference

$$F(\Delta_\Theta) := \mathcal{L}(\Theta^\star + \Delta_\Theta) - \mathcal{L}(\Theta^\star) + \lambda_{\mathrm{M}} \{\Phi(\Theta^\star + \Delta_\Theta) - \Phi(\Theta^\star)\}.$$

Note that by definition, we have $F\left(\widehat{\Delta}_\Theta\right) \leqslant 0$. It suffices to show that for all Δ_Θ such that $\Delta_\Theta + \Theta^\star \in \mathcal{C}(M_{\mathrm{op}})$ and $H(\Delta_\Theta) > \frac{18(s+(K+1)h)\lambda_{\mathrm{M}}^2}{\kappa^2}$, we have $F(\Delta_\Theta) > 0$.

It is easy to see that any such Δ_Θ obeys

$$\Delta\Omega + \Delta\Gamma_k + \Omega^{(k)} = \Omega - \Omega^\star + \Gamma_k - \Gamma^{(k)\star} + \Omega^{(k)} = \Omega + \Gamma_k \succ 0,$$

and for all $0 \leqslant k \leqslant K$

$$\|\Delta\Omega + \Delta\Gamma_k\|_2 = \left\|\Omega - \Omega^\star + \Gamma_k - \Gamma^{(k)\star}\right\|_2 = \left\|\Omega + \Gamma_k - \Omega^{(k)}\right\|_2 \leqslant \|\Omega + \Gamma_k\|_2 + \left\|\Omega^{(k)}\right\|_2 \leqslant M_{\mathrm{op}} + M_\Omega,$$

where the last inequality follows the definition of M_{op} and M_Ω. These two taken together allows us to invoke Lemma 10 to obtain

$$F(\Delta_\Theta) \geqslant \langle \nabla\mathcal{L}(\Theta^\star), \Delta_\Theta \rangle + \frac{\kappa}{2} H(\Delta_\Theta) + \lambda_{\mathrm{M}} \{\Phi(\Theta^\star + \Delta_\Theta) - \Phi(\Theta^\star)\},$$

where $\kappa = (2M_\Omega + M_{\mathrm{op}})^{-2}$.

In addition, combining Lemma 8 and the fact that $[\Theta^\star]_{\mathbb{M}^\perp} = 0$ with Wainwright [2019, Lemma 9.14], we have

$$\Phi(\Theta^\star + \Delta_\Theta) - \Phi(\Theta^\star) \geqslant \Phi([\Delta_\Theta]_{\mathbb{M}^\perp}) - \Phi([\Delta_\Theta]_{\mathbb{M}}) \tag{37}$$

for any $\Delta_\Theta \in \mathbb{H}$. This way, we have obtained a lower bound for $\Phi(\Theta^\star + \Delta_\Theta) - \Phi(\Theta^\star)$.

Take the previous two displays together to reach

$$F(\Delta_\Theta) \geqslant \langle \nabla\mathcal{L}(\Theta^\star), \Delta_\Theta \rangle + \frac{\kappa}{2} H(\Delta_\Theta) + \lambda_{\mathrm{M}} \{\Phi([\Delta_\Theta]_{\mathbb{M}^\perp}) - \Phi([\Delta_\Theta]_{\mathbb{M}})\}.$$

Recall the definition of $\Phi^\star$ in (32), under the assumption that $\mathbb{G}(\lambda_{\mathrm{M}})$ is true, we have

$$\begin{aligned}
F(\Delta_\Theta) &\geqslant \frac{\kappa}{2} H(\Delta_\Theta) - |\langle \nabla\mathcal{L}(\Theta^\star), \Delta_\Theta \rangle| + \lambda_{\mathrm{M}} \{\Phi([\Delta_\Theta]_{\mathbb{M}^\perp}) - \Phi([\Delta_\Theta]_{\mathbb{M}})\} \\
&\geqslant \frac{\kappa}{2} H(\Delta_\Theta) - \Phi(\Delta_\Theta) \Phi^\star(\nabla\mathcal{L}(\Theta^\star)) + \lambda_{\mathrm{M}} \{\Phi([\Delta_\Theta]_{\mathbb{M}^\perp}) - \Phi([\Delta_\Theta]_{\mathbb{M}})\} \\
&\geqslant \frac{\kappa}{2} H(\Delta_\Theta) - \frac{\lambda_{\mathrm{M}}}{2} \Phi(\Delta_\Theta) + \lambda_{\mathrm{M}} \{\Phi([\Delta_\Theta]_{\mathbb{M}^\perp}) - \Phi([\Delta_\Theta]_{\mathbb{M}})\} \\
&= \frac{\kappa}{2} H(\Delta_\Theta) - \frac{\lambda_{\mathrm{M}}}{2} \{\Phi([\Delta_\Theta]_{\mathbb{M}}) + \Phi([\Delta_\Theta]_{\mathbb{M}^\perp})\} + \lambda_{\mathrm{M}} \{\Phi([\Delta_\Theta]_{\mathbb{M}^\perp}) - \Phi([\Delta_\Theta]_{\mathbb{M}})\} \\
&= \frac{\kappa}{2} H(\Delta_\Theta) - \frac{\lambda_{\mathrm{M}}}{2} \{3\Phi([\Delta_\Theta]_{\mathbb{M}}) - \Phi([\Delta_\Theta]_{\mathbb{M}^\perp})\} \\
&\geqslant \frac{\kappa}{2} H(\Delta_\Theta) - \frac{3\lambda_{\mathrm{M}}}{2} \Phi([\Delta_\Theta]_{\mathbb{M}}).
\end{aligned} \tag{38}$$

By Lemma 9, we have

$$\Phi([\Delta_\Theta]_{\mathbb{M}}) \leqslant \sqrt{s + (K+1)h} \sqrt{H([\Delta_\Theta]_{\mathbb{M}})} \leqslant \sqrt{2} \sqrt{s + (K+1)h} \sqrt{H(\Delta_\Theta)},$$

where the last inequality follows from $H(\Delta_\Theta) = H\left([\Delta_\Theta]_{\mathbb{M}}\right) + H\left([\Delta_\Theta]_{\mathbb{M}^\perp}\right) \geqslant H\left([\Delta_\Theta]_{\mathbb{M}}\right)$. Plugging the above inequality into (38), we arrive at the conclusion that

$$\begin{aligned} F\left(\Delta_\Theta\right) &\geqslant \frac{\kappa}{2} H\left(\Delta_\Theta\right) - \frac{3\lambda_{\mathrm{M}}}{\sqrt{2}} \sqrt{s + (K+1)h} \sqrt{H\left(\Delta_\Theta\right)} \\ &= \frac{1}{2} \sqrt{H\left(\Delta_\Theta\right)} \left\{ \kappa \sqrt{H\left(\Delta_\Theta\right)} - 3\sqrt{2}\lambda_{\mathrm{M}} \sqrt{s + (K+1)h} \right\} > 0, \end{aligned}$$

where the last relation arises from the assumption

$$H\left(\Delta_\Theta\right) > \frac{18\left(s + (K+1)h\right)\lambda_{\mathrm{M}}^2}{\kappa^2}. \tag{39}$$

This finishes the proof.

C.4 Proof of Useful Lemmas

In this section, we collect the proof of useful lemmas.

C.4.1 Proof of Lemma 8

By the definition of $\Phi(\cdot)$ and the fact that $\Theta \in \mathbb{M}$ and $\Theta' \in \mathbb{M}^\perp$, we have

$$\begin{aligned} \Phi(\Theta + \Theta') &= \Phi\left([\Theta]_{\mathbb{M}} + \left[\Theta'\right]_{\mathbb{M}^\perp}\right) \\ &= \left|[\Omega]_{\mathcal{S}_\Omega} + \left[\Omega'\right]_{\mathcal{S}_\Omega^c}\right|_1 + \sum_{k=0}^{K} \sqrt{\alpha_k} \left| \left[\Gamma^{(k)}\right]_{\mathcal{S}_{\Gamma^{(k)}}} + \left[\Gamma^{(k)\prime}\right]_{\mathcal{S}^c_{\Gamma^{(k)}}} \right|_1 \\ &= \left|[\Omega]_{\mathcal{S}_\Omega}\right|_1 + \left|\left[\Omega'\right]_{\mathcal{S}_\Omega^c}\right|_1 + \sum_{k=0}^{K} \sqrt{\alpha_k} \left| \left[\Gamma^{(k)}\right]_{\mathcal{S}_{\Gamma^{(k)}}} \right|_1 + \sum_{k=0}^{K} \sqrt{\alpha_k} \left| \left[\Gamma^{(k)\prime}\right]_{\mathcal{S}^c_{\Gamma^{(k)}}} \right|_1 \\ &= \Phi\left([\Theta]_{\mathbb{M}}\right) + \Phi\left(\left[\Theta'\right]_{\mathbb{M}^\perp}\right) \\ &= \Phi(\Theta) + \Phi(\Theta'). \end{aligned}$$

C.4.2 Proof of Lemma 9

For any $\Theta \in \mathbb{M}$, we have $\Omega = [\Omega]_{\mathcal{S}_\Omega}$ and $\Gamma^{(k)} = \left[\Gamma^{(k)}\right]_{\mathcal{S}_{\Gamma^{(k)}}}$. Thus

$$\begin{aligned} \Phi(\Theta) &= \left|[\Omega]_{\mathcal{S}_\Omega}\right|_1 + \sum_{k=0}^{K} \sqrt{\alpha_k} \left| \left[\Gamma^{(k)}\right]_{\mathcal{S}_{\Gamma^{(k)}}} \right|_1 \\ &\leqslant |\mathcal{S}_\Omega|^{\frac{1}{2}} \left\|[\Omega]_{\mathcal{S}_\Omega}\right\|_{\mathrm{F}} + \sum_{k=0}^{K} \sqrt{\alpha_k} \left|\mathcal{S}_{\Gamma^{(k)}}\right|^{\frac{1}{2}} \left\| \left[\Gamma^{(k)}\right]_{\mathcal{S}_{\Gamma^{(k)}}} \right\|_{\mathrm{F}} && \text{(Jensen's inequality)} \\ &\overset{(ii)}{\leqslant} \sqrt{s} \left\|[\Omega]_{\mathcal{S}_\Omega}\right\|_{\mathrm{F}} + \sqrt{h} \sum_{k=0}^{K} \sqrt{\alpha_k} \left\| \left[\Gamma^{(k)}\right]_{\mathcal{S}_{\Gamma^{(k)}}} \right\|_{\mathrm{F}}. && \text{(by Assumption 1)} \end{aligned}$$

Furthermore, by Jensen's inequality, we have

$$\begin{aligned} \frac{1}{2}\Phi^2(\Theta) &\leqslant s \left\|[\Omega]_{\mathcal{S}_\Omega}\right\|_{\mathrm{F}}^2 + h \left(\sum_{k=0}^{K} \sqrt{\alpha_k} \left\| \left[\Gamma^{(k)}\right]_{\mathcal{S}_{\Gamma^{(k)}}} \right\|_{\mathrm{F}} \right)^2 \\ &\leqslant s \left\|[\Omega]_{\mathcal{S}_\Omega}\right\|_{\mathrm{F}}^2 + h(K+1) \sum_{k=0}^{K} \alpha_k \left\| \left[\Gamma^{(k)}\right]_{\mathcal{S}_{\Gamma^{(k)}}} \right\|_{\mathrm{F}}^2 \end{aligned}$$

$$\leqslant (s+(K+1)h)\left\{\left\|[\Omega]_{\mathcal{S}_\Omega}\right\|_{\mathrm{F}}^2+\sum_{k=0}^{K}\alpha_k\left\|\left[\Gamma^{(k)}\right]_{\mathcal{S}_{\Gamma^{(k)}}}\right\|_{\mathrm{F}}^2\right\}$$
$$=(s+(K+1)h)\sum_{k=0}^{K}\alpha_k\left(\left\|[\Omega]_{\mathcal{S}_\Omega}\right\|_{\mathrm{F}}^2+\left\|\left[\Gamma^{(k)}\right]_{\mathcal{S}_{\Gamma^{(k)}}}\right\|_{\mathrm{F}}^2\right).$$

By the assumption that the supports of $\Omega^\star$ and $\Gamma^{(k)\star}$ are disjoint and the fact that $\Omega=[\Omega]_{\mathcal{S}_\Omega}$ and $\Gamma^{(k)}=\left[\Gamma^{(k)}\right]_{\mathcal{S}_{\Gamma^{(k)}}}$, we then have

$$\begin{aligned}\Phi^2(\Theta)&\leqslant 2\,(s+(K+1)h)\sum_{k=0}^{K}\alpha_k\left\|[\Omega]_{\mathcal{S}_\Omega}+\left[\Gamma^{(k)}\right]_{\mathcal{S}_{\Gamma^{(k)}}}\right\|_{\mathrm{F}}^2\\&\leqslant 2\,(s+(K+1)h)\sum_{k=0}^{K}\alpha_k\left\|\Omega+\Gamma^{(k)}\right\|_{\mathrm{F}}^2\\&=2\,(s+(K+1)h)\,H\,(\Theta)\,.\end{aligned}$$

C.4.3 Proof of Lemma 10

By Lemma 5, we have

$$\begin{aligned}\mathcal{R}\,(\Delta_\Theta)=\sum_{k=0}^{K}\frac{\alpha_k}{2}&\mathrm{vec}\,(\Delta_\Omega+\Delta_{\Gamma^{(k)}})\\&\times\left\{\left(\Omega^{(k)}+t_k\,(\Delta_\Omega+\Delta_{\Gamma^{(k)}})\right)^{-1}\otimes\left(\Omega^{(k)}+t_k\,(\Delta_\Omega+\Delta_{\Gamma^{(k)}})\right)^{-1}\right\}\\&\times\mathrm{vec}\,(\Delta_\Omega+\Delta_{\Gamma^{(k)}})\,,\end{aligned}\tag{40}$$

where $t_k\in(0,1)$, $0\leqslant k\leqslant K$. Since $\gamma_{\min}(A^{-1}\otimes A^{-1})=\|A\|_2^{-2}$ for any $A\succ 0$, we have

$$\begin{aligned}&\gamma_{\min}\left(\left(\Omega^{(k)}+t_k\,(\Delta_\Omega+\Delta_{\Gamma^{(k)}})\right)^{-1}\otimes\left(\Omega^{(k)}+t_k\,(\Delta_\Omega+\Delta_{\Gamma^{(k)}})\right)^{-1}\right)\\&=\left\{\left\|\Omega^{(k)}+t_k\,(\Delta_\Omega+\Delta_{\Gamma^{(k)}})\right\|_2\right\}^{-2}\\&\geqslant\left\{\left\|\Omega^{(k)}\right\|_2+t_k\,\|\Delta_\Omega+\Delta_{\Gamma^{(k)}}\|_2\right\}^{-2}\\&\geqslant\left\{\left\|\Omega^{(k)}\right\|_2+\|\Delta_\Omega+\Delta_{\Gamma^{(k)}}\|_2\right\}^{-2}\\&\geqslant(2M_\Omega+M_{\mathrm{op}})^{-2}\,,\end{aligned}$$

where the last line follows the definition of M_Ω in (17) and the assumption that $\|\Delta_\Omega+\Delta_{\Gamma^{(k)}}\|_2\leqslant M_\Omega+M_{\mathrm{op}}$ for all $0\leqslant k\leqslant K$. Let $\kappa=(2M_\Omega+M_{\mathrm{op}})^{-2}$. Then

$$\mathcal{R}\,(\Delta_\Theta)\geqslant\frac{\kappa}{2}\sum_{k=0}^{K}\alpha_k\,\|\mathrm{vec}\,(\Delta_\Omega+\Delta_{\Gamma^{(k)}})\|_2^2\,.$$

The final result follows by noting that $\|\mathrm{vec}(\Delta_\Omega+\Delta_{\Gamma^{(k)}})\|_2=\|\Delta_\Omega+\Delta_{\Gamma^{(k)}}\|_{\mathrm{F}}$.

C.4.4 Proof of Lemma 11

We first state and prove the following lemma that gives the closed-form expression of $\Phi^\star(\cdot)$.

Lemma 12. *For the dual norm defined in (32), we have*

$$\Phi^{\star}(\Theta) = \max\left\{ |\Omega|_{\infty}, \max_{0\leqslant k\leqslant K} \frac{\left|\Gamma^{(k)}\right|_{\infty}}{\sqrt{\alpha_k}} \right\}.$$

Proof. For any $\Theta, \Theta' \in \mathbb{H}$, we have

$$\begin{aligned}
\left\langle \Theta, \Theta' \right\rangle &= \left\langle \Omega, \Omega' \right\rangle + \sum_{k=0}^{K} \left\langle \Gamma^{(k)}, \Gamma^{(k)\prime} \right\rangle \\
&\leqslant |\Omega|_{\infty} \left|\Omega'\right|_1 + \sum_{k=0}^{K} \left|\Gamma^{(k)}\right|_{\infty} \left|\Gamma'\right|_1 \\
&= |\Omega|_{\infty} \left|\Omega'\right|_1 + \sum_{k=0}^{K} \sqrt{\alpha_k} \left|\Gamma'\right|_1 \frac{|\Gamma^{(k)}|_{\infty}}{\sqrt{\alpha_k}} \\
&\leqslant |\Omega|_{\infty} \left|\Omega'\right|_1 + \left\{ \max_{0\leqslant k\leqslant K} \frac{\left|\Gamma^{(k)}\right|_{\infty}}{\sqrt{\alpha_k}} \right\} \sum_{k=0}^{K} \sqrt{\alpha_k} \left|\Gamma'\right|_1 \\
&\leqslant \max\left\{ |\Omega|_{\infty}, \max_{0\leqslant k\leqslant K} \frac{\left|\Gamma^{(k)}\right|_{\infty}}{\sqrt{\alpha_k}} \right\} \left(\left|\Omega'\right|_1 + \sum_{k=0}^{K} \sqrt{\alpha_k} \left|\Gamma'\right|_1 \right) \\
&= \max\left\{ |\Omega|_{\infty}, \max_{0\leqslant k\leqslant K} \frac{\left|\Gamma^{(k)}\right|_{\infty}}{\sqrt{\alpha_k}} \right\} \Phi\left(\Theta'\right).
\end{aligned}$$

Thus, we have

$$\Phi^{\star}(\Theta) := \sup_{\Theta':\Phi(\Theta')\leqslant 1} \left\langle \Theta, \Theta' \right\rangle \leqslant \max\left\{ |\Omega|_{\infty}, \max_{0\leqslant k\leqslant K} \frac{\left|\Gamma^{(k)}\right|_{\infty}}{\sqrt{\alpha_k}} \right\}.$$

Finally, it is easy to see that the equality is achievable. ∎

Now we are ready to prove Lemma 11. Note that

$$\begin{aligned}
\nabla_{\Omega} \mathcal{L}(\Theta) &= \sum_{k=0}^{K} \alpha_k \left(\widehat{\Sigma}^{(k)} - \left(\Omega + \Gamma^{(k)} \right)^{-1} \right), \\
\nabla_{\Gamma^{(k)}} \mathcal{L}(\Theta) &= \alpha_k \left(\widehat{\Sigma}^{(k)} - \left(\Omega + \Gamma^{(k)} \right)^{-1} \right) \quad \text{for all } 0 \leqslant k \leqslant K.
\end{aligned}$$

Then by Lemma 12, we have

$$\mathbb{G}(\lambda_{\mathrm{M}}) = \left\{ \frac{\lambda_{\mathrm{M}}}{2} \geqslant \max\left\{ \left| \sum_{k=0}^{K} \alpha_k \left(\widehat{\Sigma}^{(k)} - \Sigma^{(k)} \right) \right|_{\infty}, \max_{0\leqslant k\leqslant K} \sqrt{\alpha_k} \left| \widehat{\Sigma}^{(k)} - \Sigma^{(k)} \right|_{\infty} \right\} \right\}. \tag{41}$$

By Lemma 3 and the union bound, we have

$$\left| \sum_{k=0}^{K} \alpha_k \left(\widehat{\Sigma}^{(k)} - \Sigma^{(k)} \right) \right|_{\infty} \leqslant 80 M_{\Sigma} \max\left\{ \sqrt{\frac{\log(2(K+2)d^2/\delta)}{2N}}, \frac{\log(2(K+2)d^2/\delta)}{N} \right\}$$

and

$$\left| \widehat{\Sigma}^{(k)} - \Sigma^{(k)} \right|_{\infty} \leqslant 80 M_{\Sigma} \max\left\{ \sqrt{\frac{\log(2(K+2)d^2/\delta)}{2n_k}}, \frac{\log(2(K+2)d^2/\delta)}{n_k} \right\}$$

for all $0 \leqslant k \leqslant K$ hold simultaneously with probability at least $1-\delta$. When $\min_{0\leqslant k\leqslant K} n_k$ is large enough such that

$$\frac{\log(2\,(K+2)\,d^2/\delta)}{\min_{0\leqslant k\leqslant K} n_k} \leqslant \frac{1}{2},$$

we then have that

$$\begin{aligned}
&\left|\sum_{k=0}^{K} \alpha_k \left(\widehat{\Sigma}^{(k)} - \Sigma^{(k)}\right)\right|_\infty \leqslant 80 M_\Sigma \sqrt{\frac{\log(2\,(K+2)\,d^2/\delta)}{2N}}, \\
&\left|\widehat{\Sigma}^{(k)} - \Sigma^{(k)}\right|_\infty \leqslant 80 M_\Sigma \sqrt{\frac{\log(2\,(K+2)\,d^2/\delta)}{2n_k}} \quad \text{for all } 0 \leqslant k \leqslant K. \qquad (42)
\end{aligned}$$

hold simultaneously with probability at least $1-\delta$. Note that (42) implies that

$$\sqrt{\alpha_k}\left|\widehat{\Sigma}^{(k)} - \Sigma^{(k)}\right|_\infty \leqslant 80 M_\Sigma \sqrt{\frac{\log(2\,(K+2)\,d^2/\delta)}{2N}} \quad \text{for all } 0 \leqslant k \leqslant K.$$

Thus, when

$$\lambda_{\mathrm{M}} \geqslant 160 M_\Sigma \sqrt{\frac{\log(2\,(K+2)\,d^2/\delta)}{2N}}, \qquad (43)$$

by (41), we have $\mathbb{G}(\lambda_{\mathrm{M}})$ hold with probability $1-\delta$.

D Theorem 7 and Its Proof

The following theorem provides a high probability error bound for the D-Trace Loss estimator.

Theorem 7. *Suppose that Assumption 1 and Assumption 2 hold. Let $C_\gamma = M_\Omega^{-1}$. Furthermore, suppose that for a given $\delta \in (0,1]$ and all $k \in [K]$ we have*

$$\frac{\log\left(2(K+1)d^2/\delta\right)}{\min\{n_k, n_0\}} \leqslant \min\left\{1, \frac{\gamma_{\min}^2(\Sigma^{(k)})\gamma_{\min}^2(\Sigma^{(0)})}{C' h^2 M_\Sigma^4}\right\}. \qquad (44)$$

If

$$\lambda_\Gamma^{(k)} \geqslant 2C_\gamma\left[(C_\gamma+2)M_\Sigma M_\Gamma + 2\right] M_\Sigma \sqrt{\frac{\log\left(2(K+1)d^2/\delta\right)}{2\min\{n_k, n_0\}}}$$

for all $k \in [K]$, then

$$\left\|\widehat{\Psi}^{(k)} - \Psi^{(k)}\right\|_{\mathrm{F}} \leqslant \frac{C''\sqrt{h}\lambda_\Gamma^{(k)}}{\gamma_{\min}(\Sigma^{(k)})\gamma_{\min}(\Sigma^{(0)})} \quad \textit{for all } 0 \leqslant k \leqslant K$$

hold simultaneously with probability at least $1-\delta$, where C', C'' are universal constants that depend on C_γ.

Proof. The D-Trace Loss estimator is a special case of the FuDGE estimator proposed in Zhao et al. [2019, 2022]. We adapt the proof of Theorem 10 in Zhao et al. [2022] here.

Given $\delta \in (0,1]$, we define the event

$$\mathcal{E}^1(\delta) := \left\{\left|\widehat{\Sigma}^{(k)} - \Sigma^{(k)}\right|_\infty \leqslant C_\gamma M_\Sigma \cdot \max\left\{\sqrt{\frac{\log\left(2(K+1)d^2/\delta\right)}{2n_k}}, \frac{\log\left(2(K+1)d^2/\delta\right)}{n_k}\right\}\right.$$

$$\text{for all } 0 \leqslant k \leqslant K \Bigg\},$$

where C_γ is a universal constant. By Lemma 4 and the union bound, we have $\mathbb{P}(\mathcal{E}^1(\delta)) \geqslant 1-\delta$. In the following, we work on the event $\mathcal{E}^1(\delta)$.

Let

$$\psi_k = C_\gamma M_\Sigma \sqrt{\frac{\log\left(2(K+1)d^2/\delta\right)}{2\min\{n_k, n_0\}}}\,.$$

On the event $\mathcal{E}^1$ and when the condition (44) holds with $C' = 4096 C_\gamma^2 (C_\gamma+2)^2$, we have

$$\max\left\{\left|\widehat{\Sigma}^{(k)} - \Sigma^{(k)}\right|_\infty, \left|\widehat{\Sigma}^{(0)} - \Sigma^{(0)}\right|_\infty\right\} \leqslant \psi_k.$$

When (44) holds, we have

$$\psi_k \leqslant C_\gamma \cdot \max_{0\leqslant k\leqslant K}\left|\Sigma^{(k)}\right|_\infty \quad \text{and} \quad \psi_k \leqslant \frac{\gamma_{\min}(\Sigma^{(k)})\gamma_{\min}(\Sigma^{(0)})}{64(C_\gamma+2)hM_\Sigma}.$$

Thus, we have

$$\kappa_{\mathcal{L}}^{(k)} := \frac{1}{2}\gamma_{\min}(\Sigma^{(k)})\gamma_{\min}(\Sigma^{(0)}) - 16h\left(\psi_k^2 + 2\psi_k M_\Sigma\right) \geqslant \frac{1}{4}\gamma_{\min}(\Sigma^{(k)})\gamma_{\min}(\Sigma^{(0)}),$$

where $\kappa_{\mathcal{L}}^{(k)}$ is the restricted convexity parameter defined in the proof of Theorem 10 of Zhao et al. [2022]. Set

$$\lambda_\Gamma^{(k)} \geqslant 4\left[(C_\gamma+2)M_\Sigma M_\Gamma + 1\right]\psi_k\,.$$

Following the proof of Theorem 10 of Zhao et al. [2022], we have

$$\left\|\widehat{\Psi}^{(k)} - \Psi^{(k)}\right\|_{\mathrm{F}}^2 \leqslant \frac{18h}{\frac{1}{16}\gamma_{\min}^2(\Sigma^{(k)})\gamma_{\min}^2(\Sigma^{(0)})}\left(\lambda_\Gamma^{(k)}\right)^2,$$

which completes the proof. ■

E Proof of Theorem 4

Let $R = \lfloor s/d \rfloor$ and let $r = s - Rd$. We follow the construction of the hard-case collection of precision matrices in Step 2 of Proof of Theorem 4.1 in Cai et al. [2016b] to get $\mathcal{F}_\star$. We set $M_{n,p}$ and $c_{n,p}$ therein as $M_{n,p}$ = some constant and $c_{n,p} = R$. By (4.13) and (4.15) in Cai et al. [2016b], for any $\Omega \in \mathcal{F}_\star$, we have

$$0 < c_1 \leqslant \gamma_{\min}(\Omega) \leqslant \gamma_{\max}(\Omega) \leqslant c_2 < \infty\,.$$

Furthermore, we have $|\Omega|_0 \leqslant c_{n,p} d \leqslant s$. Thus, $\mathcal{F}_\star \subseteq \mathcal{G}_1$. Following Theorem 6.1 of Cai et al. [2016b], we have

$$\inf_{\widehat{\Omega}} \sup_{\Omega\in\mathcal{G}_1} \mathbb{E}\left[\left\|\widehat{\Omega} - \Omega\right\|_{\mathrm{F}}^2\right] \geqslant \inf_{\widehat{\Omega}} \sup_{\Omega\in\mathcal{F}_\star} \mathbb{E}\left[\left\|\widehat{\Omega} - \Omega\right\|_{\mathrm{F}}^2\right] \gtrsim d c_{n,p}\frac{\log d}{n} = \frac{dR}{s}\cdot\frac{s\log d}{n} \geqslant \frac{1}{2}\frac{s\log d}{n}\,,$$

which completes the proof.

F Proof of Theorem 5

Without loss of generality, we assume that d and h are even positive numbers. Let $m = d/2$. By Lemma 6, there exist $\{b^0, b^1, \ldots, b^M\} \subseteq \mathbb{R}^{m^2}$ such that

(i) $\|b^j\|_0 \leqslant \frac{h}{2}$ and $\|b^j\|_2 \leqslant 1$ for all $0 \leqslant j \leqslant M$;

(ii) $b^0 = 0$;

(iii) $\|b^j - b^k\|_2 \geqslant \frac{1}{4}$ for all $0 \leqslant j \neq k \leqslant M$;

(iv) $\log(M+1) \geqslant \frac{h}{4} \log\left(\frac{d^2/4 - h/2}{h/2}\right)$.

For $j = 0, \ldots, M$, let $B^j \in \mathbb{R}^{m \times m}$ be such that $\mathrm{vec}(B^j) = \delta \cdot b^j$, where δ is a positive number that depends on n, d, h and will be specified later. By (24), $d \geqslant 4h$ and

$$\log(M+1) \geqslant \frac{h}{4} \log\left(\frac{d^2/4 - h/2}{h/2}\right) = \frac{h}{4} \log\left(\frac{d^2}{2h} - 1\right) \geqslant \frac{h}{4} \log\left(\frac{d^2}{4h}\right) \geqslant \frac{h}{4} \log d.$$

Furthermore, we have

$$\left|B^j\right|_0 \leqslant \frac{h}{2}, \quad \left\|B^j\right\|_{\mathrm{F}} \leqslant \delta, \quad \text{for } 0 \leqslant j \leqslant M,$$
$$B^0 = 0, \quad \left\|B^j - B^k\right\|_{\mathrm{F}} \geqslant \frac{\delta}{4}, \quad \text{for } 0 \leqslant j \neq k \leqslant M.$$

For $j = 0, \ldots, M$, let

$$\Omega^j = \begin{bmatrix} I_m & B^j \\ (B^j)^\top & I_m \end{bmatrix}.$$

We next verify that $\Omega^j \in \mathcal{G}_2$ when

$$\delta \leqslant \min\left\{\frac{C_\Gamma}{\sqrt{2h}}, \frac{1-c_1}{2}, \frac{c_2 - 1}{2}\right\}, \tag{45}$$

where $\mathcal{G}_2$ is defined in (23). First note that $\mathrm{diag}(\Omega^j) = I_d$, $\left|\Omega^j - I_d\right|_0 = 2\left|B^j\right|_0 \leqslant h$, and

$$\left|\Omega^j - I_d\right|_1 = 2\left|B^j\right|_1 \leqslant 2\sqrt{\frac{h}{2}} \left\|B^j\right\|_{\mathrm{F}} \leqslant \sqrt{2h}\delta \leqslant C_\Gamma,$$

using the choice of δ in (45). Furthermore, we have

$$\begin{aligned}
\gamma_{\min}\left(\Omega^j\right) &= \gamma_{\min}\left(\begin{bmatrix} I_m & 0 \\ 0 & I_m \end{bmatrix} + \begin{bmatrix} 0 & B^j \\ (B^j)^\top & 0 \end{bmatrix}\right) \\
&\geqslant 1 - \left\|\begin{bmatrix} 0 & B^j \\ (B^j)^\top & 0 \end{bmatrix}\right\|_2 \\
&\geqslant 1 - \left\|\begin{bmatrix} 0 & B^j \\ (B^j)^\top & 0 \end{bmatrix}\right\|_{\mathrm{F}} \\
&= 1 - 2\left\|B^j\right\|_{\mathrm{F}} \\
&\geqslant c_1,
\end{aligned}$$

again using the choice of δ in (45). Similarly, we have $\gamma_{\max}\left(\Omega^j\right) \leqslant c_2$. Therefore, when (45) holds, we have that $\Omega^j \in \mathcal{G}_2$.

Let $\mathbb{P}_j$ denote the probability measure of $N\left(0, \{\Omega^j\}^{-1}\right)$ and let $\mathbb{P}_j^n$ denote the product probability measure of $(X_1, \ldots, X_n)$ where $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathbb{P}_j$. Note that $\mathbb{P}_0 = N(0, I_d)$. When $\delta \leqslant 1/4$, Lemma 7 gives us

$$D_{\mathrm{KL}}\left(\mathbb{P}_j^n \,\|\, \mathbb{P}_0^n\right) = nD_{\mathrm{KL}}\left(\mathbb{P}_j \,\|\, \mathbb{P}_0\right) \leqslant \frac{16}{15} n \left\|B^j\right\|_{\mathrm{F}}^2 \leqslant \frac{16}{15} n\delta^2,$$

which implies that

$$\frac{1}{M}\sum_{j=1}^{M} D_{\mathrm{KL}}\left(\mathbb{P}_j^n \,\|\, \mathbb{P}_0^n\right) \leqslant \frac{16}{15} n\delta^2. \tag{46}$$

From condition (24), we have $h \log d \geqslant 8 \log 3$ and thus $\log(M+1) \geqslant \log 9$, which implies that $M \geqslant 8$. Furthermore, we have

$$\log M = \frac{\log M}{\log(M+1)} \log(M+1) \geqslant \frac{\log 8}{\log 9} \log(M+1) \geqslant \frac{4}{5} \log(M+1) \geqslant \frac{h}{5} \log d.$$

We set

$$\delta = \frac{1}{4\sqrt{2}} \sqrt{\frac{h \log d}{n}}.$$

By (24), we have that (45) holds and that $\delta \leqslant 1/4$. Thus (46) holds, implying

$$\frac{1}{M}\sum_{j=1}^{M} D_{\mathrm{KL}}\left(\mathbb{P}_j^n \,\|\, \mathbb{P}_0^n\right) \leqslant \frac{1}{30} h \log d \leqslant \frac{1}{6} \log M. \tag{47}$$

In addition, we have

$$\left\|B^j - B^k\right\|_{\mathrm{F}} \geqslant 2 \cdot \frac{1}{32\sqrt{2}} \sqrt{\frac{h \log d}{n}} \quad \text{for all } 0 \leqslant j \neq k \leqslant M. \tag{48}$$

Note that

$$\frac{\log(M+1) - \log 2}{\log M} - \frac{1}{6} \geqslant \frac{\log M - \log 2}{\log M} - \frac{1}{6} \geqslant \frac{\log 8 - \log 2}{\log 8} - \frac{1}{6} = \frac{1}{2}. \tag{49}$$

Following Section 2.2 and Corollary 2.6 in Tsybakov [2008], combined with (47)–(49), we have

$$\min_{\widehat{\Omega}} \sup_{\Omega \in \mathcal{G}_2} \mathbb{E}\left[\left\|\widehat{\Omega} - \Omega\right\|_{\mathrm{F}}^2\right] \geqslant \frac{1}{2}\left(\frac{1}{32\sqrt{2}} \sqrt{\frac{h \log d}{n}}\right)^2 = \frac{1}{4096} \cdot \frac{h \log d}{n},$$

which completes the proof.

G Proof of Theorem 6

The upper bound follows directly from Corollary 1. To prove the lower bound, we assume that $\mathcal{P}_k = N\left(0, \left(\Omega^{(k)}\right)^{-1}\right)$.

First, when $h = 0$, we have N i.i.d. samples from $\Omega^{(0)}$ where

$$\Omega^{(0)} \in \left\{\Omega \in \mathbb{R}^{d \times d} : \Omega \succ 0,\ |\Omega|_0 \leqslant s,\ \|\Omega\|_2 \leqslant M_\Omega,\ \left|\Omega^{-1}\right|_\infty \leqslant M_\Sigma\right\} := \tilde{\mathcal{G}}'$$

for some positive universal constants M_Ω and M_Σ. Let

$$\mathcal{G}' = \left\{\Omega \in \mathbb{R}^{d \times d} : \Omega \succ 0,\ |\Omega|_0 \leqslant s,\ 0 < \frac{1}{M_\Sigma} \leqslant \gamma_{\min}(\Omega) \leqslant \gamma_{\max}(\Omega) \leqslant M_\Omega < \infty\right\}.$$

Since $\left|\Omega^{-1}\right|_\infty \leqslant \|\Omega^{-1}\|_2 = \{\gamma_{\min}(\Omega)\}^{-1} \leqslant M_\Sigma$ for any $\Omega \in \mathcal{G}'$, we have $\mathcal{G}' \subseteq \tilde{\mathcal{G}}'$. Thus, when $s \geqslant d \geqslant c' N^\beta$, for some universal constants $c' > 0$ and $\beta > 1$ and

$$\lceil s/d \rceil = o\left(\frac{N}{(\log d)^{\frac{3}{2}}}\right)$$

it follows by Theorem 4 that

$$\inf_{\widehat{\Omega}} \sup_{\Omega^{(0)} \in \tilde{\mathcal{G}}'} \mathbb{E}\left[\left\|\widehat{\Omega} - \Omega^{(0)}\right\|_{\mathrm{F}}^2\right] \geqslant \inf_{\widehat{\Omega}} \sup_{\Omega^{(0)} \in \mathcal{G}'} \mathbb{E}\left[\left\|\widehat{\Omega} - \Omega^{(0)}\right\|_{\mathrm{F}}^2\right] \gtrsim \frac{s \log d}{N}. \tag{50}$$

If $\Omega^{(k)} = \Omega^\star = I_d$ and $\Gamma^{(k)\star} = 0$ for every $k \in [K]$, then samples from source distributions cannot be used to estimate $\Omega^{(0)}$. Therefore, we must depend solely on samples from the target distribution to estimate $\Omega^{(0)}$. Note that now we have

$$\begin{aligned}\Omega^{(0)} \in \big\{\Omega \in \mathbb{S}^{d\times d} : \Omega \succ 0,\, \Omega = I_d + \Delta,\, \Delta_{jj} = 0 \text{ for all } 1 \leqslant j \leqslant d,\\ |\Delta|_0 \leqslant h,\, |\Delta|_1 \leqslant M_\Gamma,\, \|\Omega\|_2 \leqslant M_\Omega,\, \left|\Omega^{-1}\right|_\infty \leqslant M_\Sigma\big\} := \tilde{\mathcal{G}}''.\end{aligned}$$

Let

$$\begin{aligned}\mathcal{G}'' = \big\{\Omega \in \mathbb{S}^{d\times d} : \Omega \succ 0,\, \Omega = I_d + \Delta,\, \Delta_{jj} = 0 \text{ for all } 1 \leqslant j \leqslant d,\\ |\Delta|_0 \leqslant h,\, |\Delta|_1 \leqslant M_\Gamma,\, 0 < \frac{1}{M_\Sigma} \leqslant \gamma_{\min}(\Omega) \leqslant \gamma_{\max}(\Omega) \leqslant M_\Omega < \infty\big\}.\end{aligned}$$

Given that for any $\Omega \in \mathcal{G}''$, it holds that $\left|\Omega^{-1}\right|_\infty \leqslant \|\Omega^{-1}\|_2 = \{\gamma_{\min}(\Omega)\}^{-1} \leqslant M_\Sigma$, we can deduce that $\mathcal{G}'' \subseteq \tilde{\mathcal{G}}''$. By Theorem 5, when $M_\Omega, M_\Sigma > 1$ and

$$d \geqslant 4h,\, h \log d \geqslant 8 \log 3,\, \frac{h \log d}{n_0} \leqslant \min\left\{2, 8\left(1 - \frac{1}{M_\Sigma}\right)^2, 8\left(1 - M_\Omega\right)^2\right\},\, h\sqrt{\frac{\log d}{n_0}} \leqslant 4M_\Gamma,$$

we have

$$\inf_{\widehat{\Omega}} \sup_{\Omega^{(0)} \in \tilde{\mathcal{G}}''} \mathbb{E}\left[\left\|\widehat{\Omega} - \Omega^{(0)}\right\|_{\mathrm{F}}^2\right] \geqslant \inf_{\widehat{\Omega}} \sup_{\Omega^{(0)} \in \mathcal{G}''} \mathbb{E}\left[\left\|\widehat{\Omega} - \Omega^{(0)}\right\|_{\mathrm{F}}^2\right] \gtrsim \frac{h \log d}{n_0}. \tag{51}$$

The final result follows from (50) and (51).

H Additional Optimization Details

In this section, we provide more details about the numerical algorithms introduced in Section 4.1.

We first discuss how to compute the updating steps (9) and (10). Note that (9) is equivalent to

$$\begin{aligned}\Omega_k^{(t)} &= \arg\min_{\Omega_k \succ 0}\left\{f_k(\Omega_k) + \rho\left\langle Z_k^{(t-1)}, \Omega_k - \Omega^{(t-1)} - \Gamma_k^{(t-1)}\right\rangle + \frac{\rho}{2}\left\|\Omega_k - \Omega^{(t-1)} - \Gamma_k^{(t-1)}\right\|_{\mathrm{F}}^2\right\}\\ &= \arg\min_{\Omega_k \succ 0}\left\{-\alpha_k \log\det(\Omega_k) + \frac{\rho}{2}\left\|\Omega_k - \tilde{C}_k^{(t-1)}\right\|_{\mathrm{F}}^2\right\},\end{aligned} \tag{52}$$

where $\tilde{C}_k^{(t-1)} = -Z_k^{(t-1)} + \left(\Omega^{(t-1)} + \Gamma_k^{(t-1)}\right) - (\alpha_k/\rho)\widehat{\Sigma}^{(k)}$ for $0 \leqslant k \leqslant K$. Taking the gradient with respect to Ω_k in (52) and setting it to zero gives

$$-\alpha_k \Omega_k^{-1} + \rho\Omega_k - \rho\tilde{C}_k^{(t-1)} = 0. \tag{53}$$

The matrix $\Omega_k^{(t)}$ is obtained by finding $\Omega_k \succ 0$ that satisfies (53). Let $\rho\tilde{C}_k^{(t-1)} = U\Lambda U^\top$, $\Lambda = \mathrm{diag}\left(\{\lambda_i\}_{i=1}^d\right)$, be the eigenvalue decomposition of $\rho\tilde{C}_k^{(t-1)}$. Following [Boyd et al., 2011, Section 6.5], we have

$$\Omega_k^{(t)} = U\mathrm{diag}\left(\left\{\frac{\lambda_i + \sqrt{\lambda_i^2 + 4\rho\alpha_k}}{2\rho}\right\}_{i=1}^d\right)U^\top.$$

On ther other hand, computing (10) is equivalent to solving

$$\begin{aligned}\Omega^{(t)}, \left\{\Gamma_k^{(t)}\right\} &\in \arg\min_{\Omega,\{\Gamma^{(k)}\}}\left\{\rho\sum_{k=0}^K\left\langle Z_k^{(t-1)}, \Omega_k^{(t)} - \Omega - \Gamma^{(k)}\right\rangle + \frac{\rho}{2}\sum_{k=0}^K\left\|\Omega_k^{(t)} - \Omega - \Gamma^{(k)}\right\|_{\mathrm{F}}^2\right.\\ &= \arg\min_{\Omega,\{\Gamma^{(k)}\}}\left\{\frac{\rho}{2}\sum_{k=0}^K\left\|\Omega + \Gamma^{(k)} - \check{C}_k^{(t)}\right\|_{\mathrm{F}}^2 + \lambda_{\mathrm{M}}\left|\Omega\right|_1 + \lambda_{\mathrm{M}}\sum_{k=0}^K\sqrt{\alpha_k}\left|\Gamma^{(k)}\right|_1\right\}\end{aligned}$$

where $\check{C}_k^{(t)} = \Omega_k^{(t)} + Z_k^{(t-1)}$. Given $c = (c_1, \ldots, c_K)$, let $\mathcal{S}(c)$ be defined as $\mathcal{S}(c) = (x^\star, y^\star)$ where

$$(x^\star, y^\star) \in \arg\min_{(x,y)}\left\{\frac{\rho}{2}\sum_{k=0}^K(x + y_k - c_k)^2 + \lambda_{\mathrm{M}}|x| + \lambda_{\mathrm{M}}\sum_{k=0}^K\sqrt{\alpha_k}|y_k|\right\} \tag{54}$$

with $y = (y_1, \ldots, y_K)$. With $\Omega^{(t)} = (\Omega_{jl}^{(t)})_{1\leqslant j,l\leqslant d}$, $\Gamma_k^{(t)} = (\Gamma_{k,jl}^{(t)})_{1\leqslant j,l\leqslant d}$, and $\check{C}_k^{(t)} = (\check{C}_{k,jl}^{(t)})_{1\leqslant j,l\leqslant d}$, we have

$$\Omega_{jl}^{(t)}, \left\{\Gamma_{k,jl}^{(t)}\right\}_{k=1}^K = \mathcal{S}\left(\left\{\check{C}_{k,jl}^{(t)}\right\}_{k=1}^K\right), \quad \text{for } 1 \leqslant j, l \leqslant d.$$

To solve (54), we iteratively update x or y while fixing the other until convergence. For $c \in \mathbb{R}$ and $\lambda \geqslant 0$, let

$$\mathrm{ST}_\lambda(c) = \begin{cases} c - \lambda & \text{if } c > \lambda, \\ 0 & \text{if } |c| \leqslant \lambda, \\ c + \lambda & \text{if } c < -\lambda \end{cases} \tag{55}$$

be the soft-thresholding function. After initializing $x^{(0)}, y^{(0)}$, we repeat the following process until convergence:

$$\begin{aligned}x^{(r)} &= \arg\min_x\left\{\frac{\rho}{2}\sum_{k=0}^K\left(x + y_k^{(r-1)} - c_k\right)^2 + \lambda_{\mathrm{M}}|x| + \lambda_{\mathrm{M}}\sum_{k=0}^K\sqrt{\alpha_k}\left|y_k^{(r-1)}\right|\right\}\\ &= \mathrm{ST}_{\lambda_{\mathrm{M}}/(\rho(K+1))}\left(\frac{1}{K+1}\sum_{k=0}^K\left(c_k - y_k^{(r-1)}\right)\right),\end{aligned}$$

and

$$\begin{aligned}y^{(r)} &= \arg\min_y\left\{\frac{\rho}{2}\sum_{k=0}^K\left(x^{(r)} + y_k - c_k\right)^2 + \lambda_{\mathrm{M}}\left|x^{(r)}\right| + \lambda_{\mathrm{M}}\sum_{k=0}^K\sqrt{\alpha_k}\left|y_k\right|\right\}\\ \Leftrightarrow y_k^{(r)} &= \mathrm{ST}_{\lambda_{\mathrm{M}}\sqrt{\alpha_k}/\rho}\left(c_k - x^{(r)}\right) \quad \text{for } 0 \leqslant k \leqslant K.\end{aligned}$$

We then discuss the stopping criterion for the ADMM algorithm to solve Trans-MT-Glasso objective (3).

Stopping criterion for ADMM. Following [Boyd et al., 2011, Section 3.3.1], let $\epsilon^{\text{abs}} > 0$ be an absolute tolerance and $\epsilon^{\text{rel}} > 0$ be a relative tolerance, we then define the feasibility tolerance for primal feasibility condition $\epsilon^{\text{pri}} > 0$ and the feasibility tolerance for dual feasibility condition $\epsilon^{\text{dual}} > 0$ at iteration t as

$$\epsilon^{\text{pri}} = \epsilon^{\text{abs}} d\sqrt{K+1} + \epsilon^{\text{rel}} \max\left\{ \left(\sum_{k=0}^{K} \left\| \Omega_k^{(t)} \right\|_{\text{F}}^2 \right)^{\frac{1}{2}}, \left(\sum_{k=0}^{K} \left\| \Omega^{(t)} + \Gamma_k^{(t)} \right\|_{\text{F}}^2 \right)^{\frac{1}{2}} \right\},$$

$$\epsilon^{\text{dual}} = \epsilon^{\text{abs}} d\sqrt{K+1} + \epsilon^{\text{rel}} \left(\sum_{k=0}^{K} \left\| Z_k^{(t)} \right\|_{\text{F}}^2 \right)^{\frac{1}{2}}.$$

Besides, let

$$r^{\text{pri}} = \left(\sum_{k=0}^{K} \left\| \Omega_k^{(t)} - \left(\Omega^{(t)} + \Gamma_k^{(t)} \right) \right\|_{\text{F}}^2 \right)^{\frac{1}{2}} \quad \text{and}$$

$$r^{\text{dual}} = \rho \left(\sum_{k=0}^{K} \left\| \left(\Omega^{(t)} + \Gamma_k^{(t)} \right) - \left(\Omega^{(t-1)} + \Gamma_k^{(t-1)} \right) \right\|_{\text{F}}^2 \right)^{\frac{1}{2}}$$

be the primal and dual residuals at iteration t. We then stop the iteration if

$$r^{\text{pri}} \leqslant \epsilon^{\text{pri}} \quad \text{and} \quad r^{\text{dual}} \leqslant \epsilon^{\text{dual}}.$$

Stopping criterion for sub problem (54). The optimality conditions of problem (54) are

$$0 \in \rho \sum_{k=0}^{K} (x^\star + y_k^\star - c_k) + \lambda_{\text{M}} \partial |x^\star|, \tag{56}$$

$$0 \in \rho (x^\star + y^\star - c_k) + \lambda_{\text{M}} \sqrt{\alpha_k} \partial |y_k^\star| \quad \text{for all } k = 0, 1, \ldots, K. \tag{57}$$

By the definition of $y^{(r)}$, we always have $x^{(r)}, y^{(r)}$ satisfy (57). Besides, by definition of $x^{(r)}$, we have

$$0 \in \rho \sum_{k=0}^{K} \left(x^{(r)} + y_k^{(r-1)} - c_k \right) + \lambda_{\text{M}} \partial |x^{(r)}|. \tag{58}$$

Thus, when

$$\rho \sum_{k=0}^{K} \left(y_k^{(r)} - y_k^{(r-1)} \right) = 0,$$

we will have $x^{(r)}, y^{(r)}$ satisfy (58). Let $\epsilon^{\text{abs}} > 0$ be an absolute tolerance and $\epsilon^{\text{rel}} > 0$ be a relative tolerance, we then stop at iteration r if

$$r^{\text{sub}} = \rho \left| \sum_{k=0}^{K} \left(y_k^{(r)} - y_k^{(r-1)} \right) \right| \leqslant (K+1)\epsilon^{\text{abs}} + \epsilon^{\text{rel}} \max\left\{ \sum_{k=0}^{K} \left| y_k^{(r)} \right|, \sum_{k=0}^{K} \left| y_k^{(r-1)} \right| \right\} := \epsilon^{\text{sub}}.$$

The final ADMM algorithm and the algorithm to solve the sub problem are summarized in Algorithm 1 and Algorithm 2.

For the proximal gradient descent algorithm to solve D-Trace loss objective (7), the stopping criterion is introduced as below and the detailed algorithm is described in Algorithm 3.

Algorithm 1 ADMM for Trans-MT-GLasso

Input: $\{\widehat{\Sigma}^{(k)}\}_{k=0}^{K}$, λ_{M}, ρ, ϵ^{abs}, ϵ^{rel} and $\{\alpha_k\}_{k=0}^{K}$.

Initialize: Let $\Omega^{(0)} = I_d$, $\Gamma_k^{(0)} = 0$ and $Z_k^{(0)} = I_d$ for all $0 \leqslant k \leqslant K$. Let $r^{\mathrm{pri}} = r^{\mathrm{dual}} = \infty$ and $\epsilon^{\mathrm{pri}} = \epsilon^{\mathrm{dual}} = 0$. Let $t = 0$.

while $r^{\mathrm{pri}} > \epsilon^{\mathrm{pri}}$ or $r^{\mathrm{dual}} > \epsilon^{\mathrm{dual}}$ **do**

$t \leftarrow t + 1$.

for $k = 0, 1, \ldots, K$ **do**

Let

$$\tilde{C}_k^{(t-1)} = -Z_k^{(t-1)} + \left(\Omega^{(t-1)} + \Gamma_k^{(t-1)}\right) - (\alpha_k/\rho)\widehat{\Sigma}^{(k)},$$

and compute the eigenvalue decomposition of $\rho\tilde{C}_k^{(t-1)}$ as

$$\rho\tilde{C}_k^{(t-1)} = U\Lambda U^\top, \quad \Lambda = \mathrm{diag}\left(\{\lambda_i\}_{i=1}^{d}\right).$$

Let

$$\Omega_k^{(t)} = U\mathrm{diag}\left(\left\{\frac{\lambda_i + \sqrt{\lambda_i^2 + 4\rho\alpha_k}}{2\rho}\right\}_{i=1}^{d}\right)U^\top.$$

Let

$$\check{C}_k^{(t)} = \Omega_k^{(t)} + Z_k^{(t-1)}.$$

end for

Solve

$$\Omega_{jl}^{(t)}, \left\{\Gamma_{k,jl}^{(t)}\right\}_{k=1}^{K} = \mathcal{S}\left(\left\{\check{C}_{k,jl}^{(t)}\right\}_{k=1}^{K}\right) \quad \text{for all } 1 \leqslant j, l \leqslant d.$$

by Algorithm 1.

Let

$$Z_k^{(t)} = Z_k^{(t-1)} + \rho\left(\Omega_k^{(t)} - \Omega^{(t)} - \Gamma_k^{(t)}\right) \text{ for all } 0 \leqslant k \leqslant K.$$

Let

$$r^{\mathrm{pri}} = \left(\sum_{k=0}^{K}\left\|\Omega_k^{(t)} - \left(\Omega^{(t)} + \Gamma_k^{(t)}\right)\right\|_{\mathrm{F}}^2\right)^{\frac{1}{2}},$$

$$r^{\mathrm{dual}} = \rho\left(\sum_{k=0}^{K}\left\|\left(\Omega^{(t)} + \Gamma_k^{(t)}\right) - \left(\Omega^{(t-1)} + \Gamma_k^{(t-1)}\right)\right\|_{\mathrm{F}}^2\right)^{\frac{1}{2}},$$

and

$$\epsilon^{\mathrm{pri}} = \epsilon^{\mathrm{abs}} d\sqrt{K+1} + \epsilon^{\mathrm{rel}} \max\left\{\left(\sum_{k=0}^{K}\left\|\Omega_k^{(t)}\right\|_{\mathrm{F}}^2\right)^{\frac{1}{2}}, \left(\sum_{k=0}^{K}\left\|\Omega^{(t)} + \Gamma_k^{(t)}\right\|_{\mathrm{F}}^2\right)^{\frac{1}{2}}\right\},$$

$$\epsilon^{\mathrm{dual}} = \epsilon^{\mathrm{abs}} d\sqrt{K+1} + \epsilon^{\mathrm{rel}}\left(\sum_{k=0}^{K}\left\|Z_k^{(t)}\right\|_{\mathrm{F}}^2\right)^{\frac{1}{2}}.$$

end while

Output: $\breve{\Omega}^{(k)} = \Omega^{(t)} + \Gamma_k^{(t)}$ for all $k = 0, 1, \ldots, K$.

Algorithm 2 Solver of sub problem (54)

Input: $c = (c_0, c_1, \ldots, c_K)$; Initial $x^{(0)}$ and $\{y_k^{(0)}\}_{k=0}^K$; λ_{M}, ρ and $\{\alpha_k\}_{k=0}^K$.
Initialize: $r^{\mathrm{sub}} = \infty$ and $\epsilon^{\mathrm{sub}} = 0$. Define $\mathrm{ST}_\lambda(\cdot)$ as in (55). Let $r = 0$.
while $r^{\mathrm{sub}} > \epsilon^{\mathrm{sub}}$ **do**
  $r \leftarrow r + 1$.
  Let

$$x^{(r)} = \mathrm{ST}_{\lambda_{\mathrm{M}}/(\rho(K+1))}\left(\frac{1}{K+1}\sum_{k=0}^{K}\left(c_k - y_k^{(r-1)}\right)\right).$$

  Let

$$y_k^{(r)} = \mathrm{ST}_{\lambda_{\mathrm{M}}\sqrt{\alpha_k}/\rho}\left(c_k - x^{(r)}\right) \quad \text{for all } 0 \leqslant k \leqslant K.$$

  Let

$$r^{\mathrm{sub}} = \rho\left|\sum_{k=0}^{K}\left(y_k^{(r)} - y_k^{(r-1)}\right)\right|,$$
$$\epsilon^{\mathrm{sub}} = (K+1)\epsilon^{\mathrm{abs}} + \epsilon^{\mathrm{rel}}\rho\max\left\{\sum_{k=0}^{K}\left|y_k^{(r)}\right|, \sum_{k=0}^{K}\left|y_k^{(r-1)}\right|\right\}.$$

end while
Output: $x^{(r)}$ and $\{y_k^{(r)}\}_{k=0}^K$.

Stopping criterion for proximal gradient. If $\Gamma^{(t)}$ is the solution for (7), the optimization criterion requires that

$$0 \in \nabla L_D(\Gamma^{(t)}) + \lambda_\Gamma^{(k)} \cdot \partial\left|\Gamma^{(t)}\right|_1.$$

Note that by the definition of $\Gamma^{(t)}$ in (12), we have

$$0 \in \Gamma^{(t)} - \left(\Gamma^{(t-1)} - \eta\nabla L_D\left(\Gamma^{(t-1)}\right)\right) + \eta\lambda_\Gamma^{(k)} \cdot \partial\left|\Gamma^{(t)}\right|_1, \tag{59}$$

which implies that

$$0 \in \nabla L_D(\Gamma^{(t)}) + \lambda_\Gamma^{(k)} \cdot \partial\left|\Gamma^{(t)}\right|_1 + \frac{1}{\eta}\left(\Gamma^{(t)} - \Gamma^{(t-1)}\right) - \left[\nabla L_D\left(\Gamma^{(t)}\right) - \nabla L_D\left(\Gamma^{(t-1)}\right)\right].$$

Thus, when

$$\begin{aligned} r^D &= \left\|\frac{1}{\eta}\left(\Gamma^{(t)} - \Gamma^{(t-1)}\right) - \left[\nabla L_D\left(\Gamma^{(t)}\right) - \nabla L_D\left(\Gamma^{(t-1)}\right)\right]\right\|_{\mathrm{F}} \\ &= \left\|\frac{1}{\eta}\left(\Gamma^{(t)} - \Gamma^{(t-1)}\right) - \frac{1}{2}\widehat{\Sigma}^{(0)}\left(\Gamma^{(t)} - \Gamma^{(t-1)}\right)\widehat{\Sigma}^{(k)} - \frac{1}{2}\widehat{\Sigma}^{(k)}\left(\Gamma^{(t)} - \Gamma^{(t-1)}\right)\widehat{\Sigma}^{(0)}\right\|_{\mathrm{F}} \end{aligned}$$

is close to zero, we then have (59) hold approximately. Therefore, given an absolute tolerance $\epsilon^{\mathrm{abs}} > 0$ and a relative tolerance $\epsilon^{\mathrm{rel}} > 0$, we then stop at iteration t if

$$r^D \leqslant \epsilon^{\mathrm{abs}} d + \epsilon^{\mathrm{rel}} \times \max\left\{\frac{1}{\eta}\left\|\Gamma^{(t)} - \Gamma^{(t-1)}\right\|_{\mathrm{F}}, \left\|\frac{1}{2}\widehat{\Sigma}^{(0)}\left(\Gamma^{(t)} - \Gamma^{(t-1)}\right)\widehat{\Sigma}^{(k)}\right\|_{\mathrm{F}}\right\}.$$

Algorithm 3 Proximal Gradient Method for D-Trace Loss

1: **Input:** $\widehat{\Sigma}^{(0)}$, $\widehat{\Sigma}^{(k)}$, $\lambda_\Gamma^{(k)}$, η, ϵ^{abs} and ϵ^{rel}.
2: **Initialize:** $\Gamma^{(0)} = 0$, $r^D = \infty$, $\epsilon^D = 0$ and $t = 0$.
3: **while** $r^D > \epsilon^D$ **do**
4: $t \leftarrow t + 1$
5: Let

$$\begin{aligned} A^{(t-1)} &= \Gamma^{(t-1)} - \eta \nabla L_D\left(\Gamma^{(t-1)}\right) \\ &= \Gamma^{(t-1)} - \eta\left\{\frac{1}{2}\left(\widehat{\Sigma}^{(k)}\Gamma^{(t-1)}\widehat{\Sigma}^{(0)} + \widehat{\Sigma}^{(0)}\Gamma^{(t-1)}\widehat{\Sigma}^{(k)}\right) - \left(\widehat{\Sigma}^{(0)} - \widehat{\Sigma}^{(k)}\right)\right\}. \end{aligned}$$

6: Let

$$\Gamma_{jl}^{(t)} = \left[|A_{jl}^{(t-1)}| - \lambda_\Gamma^{(k)}\eta\right]_+ \cdot A_{jl}^{(t-1)}/|A_{jl}^{(t-1)}|, \qquad 1 \leqslant j, l \leqslant d.$$

7: Let

$$r^D = \left\|\frac{1}{\eta}\left(\Gamma^{(t)} - \Gamma^{(t-1)}\right) - \frac{1}{2}\widehat{\Sigma}^{(0)}\left(\Gamma^{(t)} - \Gamma^{(t-1)}\right)\widehat{\Sigma}^{(k)} - \frac{1}{2}\widehat{\Sigma}^{(k)}\left(\Gamma^{(t)} - \Gamma^{(t-1)}\right)\widehat{\Sigma}^{(0)}\right\|_{\text{F}},$$

and

$$\epsilon^D = \epsilon^{\text{abs}} d + \epsilon^{\text{rel}} \times \max\left\{\frac{1}{\eta}\left\|\Gamma^{(t)} - \Gamma^{(t-1)}\right\|_{\text{F}}, \left\|\frac{1}{2}\widehat{\Sigma}^{(0)}\left(\Gamma^{(t)} - \Gamma^{(t-1)}\right)\widehat{\Sigma}^{(k)}\right\|_{\text{F}}\right\}.$$

8: **end while**
9: **Output:** $\widehat{\Psi}^{(k)} = \Gamma^{(t)}$.

I Minimax Optimal Rate for Differential Network Estimation

Differential network estimation aims to estimate the difference between two precision matrices without first estimating the individual precision matrices. Although existing studies focus on providing upper bounds for this problem [Zhao et al., 2014, Yuan et al., 2017, Zhao et al., 2022], there is no known matching lower bound, making the minimax optimal rate an open question. As a byproduct of our analysis, in this section, we provide a minimax optimal rate for differential network estimation problem under certain conditions. To the best of our knowledge, this is the first minimax optimal guarantee towards this direction.

We start by formulating the problem setup. Note that we reintroduce some notation used in this section to make it self-contained, and one should not confuse it with the notation used in the other parts of the paper. Suppose that we have n_X i.i.d. samples $X_1, \ldots, X_{n_X} \sim N(0, \Omega_X^{-1})$ and n_Y i.i.d. samples $Y_1, \ldots, Y_{n_Y} \sim N(0, \Omega_Y^{-1})$. Let $\Delta := \Omega_Y - \Omega_X$ be the differential network between Ω_X and Ω_Y. Our goal is to use samples from two populations to estimate Δ. In addition, we assume that Ω_X and Ω_Y belong to the following parameter space.

$$\begin{aligned} (\Omega_X, \Omega_Y) \in \mathcal{M} := \big\{\Omega_X, \Omega_Y \in \mathbb{S}^{d\times d} : \; & c_1 \leqslant \gamma_{\min}(\Omega_X) \leqslant \gamma_{\max}(\Omega_X) \leqslant c_2, \\ & c_1 \leqslant \gamma_{\min}(\Omega_Y) \leqslant \gamma_{\max}(\Omega_Y) \leqslant c_2, \, |\Delta|_0 \leqslant h, \, |\Delta|_1 \leqslant C_\Gamma\big\}, \end{aligned} \tag{60}$$

where $c_1, c_2, C_\Gamma > 0$ are positive universal constants. The parameter space defined in (60) requires that the smallest eigenvalues and the largest eigenvalues of Ω_X and Ω_Y are lower bounded and upper bounded, respectively. Besides, we also require that the sparsity level of the differential network is bounded by h, and the L_1 norm of the differential network is bounded by universal constant.

We first present the minimax lower bound for the above estimation problem. We state the result in the following theorem.

Theorem 8. *Assume that*

$$d \geqslant 4h,\ h\log d \geqslant 8\log 3,\ \frac{h\log d}{n} \leqslant \min\left\{2, 8(1-c_1)^2, 8(1-c_2)^2\right\},\ h\sqrt{\frac{\log d}{n}} \leqslant 4C_\Gamma. \tag{61}$$

We then have

$$\min_{\widehat{\Delta}} \sup_{(\Omega_X,\Omega_Y)\in\mathcal{M}} \mathbb{E}\left[\left\|\widehat{\Delta}-\Delta\right\|_{\mathrm{F}}^2\right] \gtrsim \frac{h\log d}{\min\{n_X,n_Y\}}$$

Proof. Let $\Omega_X = I_d$, note that

$$(I_d,\Omega_Y)\in\mathcal{M} \Leftarrow \Omega_Y\in\mathcal{G}_2,$$

where $\mathcal{G}_2$ is defined in (23). Now that samples from population X are useless for estimating Ω_Y, and we can only rely on samples from population Y to estimate Ω_Y. By Theorem 5, we then have

$$\min_{\widehat{\Delta}} \sup_{(\Omega_X,\Omega_Y)\in\mathcal{M}} \mathbb{E}\left[\left\|\widehat{\Delta}-\Delta\right\|_{\mathrm{F}}^2\right] \geqslant \min_{\widehat{\Omega}_Y} \sup_{\Omega_Y\in\mathcal{G}_2} \mathbb{E}\left[\left\|\widehat{\Omega}_Y-\Omega_Y\right\|_{\mathrm{F}}^2\right] \gtrsim \frac{h\log d}{n_Y}.$$

Similarly, we can show that

$$\min_{\widehat{\Delta}} \sup_{(\Omega_X,\Omega_Y)\in\mathcal{M}} \mathbb{E}\left[\left\|\widehat{\Delta}-\Delta\right\|_{\mathrm{F}}^2\right] \gtrsim \frac{h\log d}{n_X}.$$

Combine the above two inequalities, we have the final result. ■

Next, we derive the matching upper bound. Let $\widehat{\Delta}$ be the D-Trace loss estimator defined in (7), and $\check{\Delta}(\tau)$ be the truncated version of $\widehat{\Delta}$ as defined in (62). Following directly from Theorem 13, we then have the following theorem.

Theorem 9. *Assume that $C_\Gamma \leqslant d^{\tau_3}$, where C_Γ is the same universal constant used in (60). Let*

$$\lambda_\Gamma \asymp \sqrt{\frac{\log d}{\min\{n_X,n_Y\}}},$$

where λ_Γ is the penalization parameter of D-Trace loss. Then for any $\tau \geqslant \tau_3$, we have

$$\sup_{(\Omega_X,\Omega_Y)\in\mathcal{M}} \mathbb{E}\left[\left\|\check{\Delta}(\tau)-\Delta\right\|_{\mathrm{F}}^2\right] \lesssim \frac{h\log d}{\min\{n_X,n_Y\}}.$$

Combine Theorem 8 and Theorem 9, we then have the final main result of this section.

Theorem 10. *Assume that the conditions of Theorem 8 and Theorem 9 hold. We have*

$$\min_{\widehat{\Delta}} \sup_{(\Omega_X,\Omega_Y)\in\mathcal{M}} \mathbb{E}\left[\left\|\widehat{\Delta}-\Delta\right\|_{\mathrm{F}}^2\right] \asymp \frac{h\log d}{\min\{n_X,n_Y\}}.$$

J Upper Bound for Expected Error

In this section, we develop theoretical guarantees for the expected error. We start with the analysis of the Trans-MT-Glasso. The subsequent theorem provides the expected error measured in the Frobenius norm.

Theorem 11. *Suppose that Assumption 1 and Assumption 2 hold. Assume that* $2(K+2) \leqslant d^{\tau_1}$ *and* $N \leqslant d^{\tau_2}$*, for some universal constants* $\tau_1, \tau_2 > 0$*. In addition, assume that*

$$\frac{\log d}{\min_{0 \leqslant k \leqslant K} n_k} \lesssim 1.$$

Let $\bar{n} = N/(K+1)$*,* $M_{\text{op}} \geqslant M_\Omega$ *and* $M_{\text{op}} = O(1)$*. If* $\lambda_M \asymp \sqrt{\log d / N}$*, we have*

$$\mathbb{E}\left[\sum_{k=0}^{K} \alpha_k \left\|\widehat{\Omega}_k - \Omega^{(k)}\right\|_{\text{F}}^2\right] \lesssim \left(\frac{s}{N} + \frac{h}{\bar{n}}\right) \log d.$$

Refer to the proof in Appendix K. The rate described in Theorem 11 consists of two parts. The first part, which is of the order $(s \log d)/N$, refers to the estimation of the shared component. The second part, of the order $(h \log d)/\bar{n}$, relates to the estimation of the individual components.

As discussed in Section 5.2, the differential network estimate $\widehat{\Psi}^{(k)}$ is considered the result of a black-box algorithm, with the presumption that its estimation errors are appropriately controlled. Let $\mathcal{B}_{\text{F}}(R) := \left\{A \in \mathbb{R}^{d \times d} : \|A\|_{\text{F}} \leqslant R\right\}$ be the ball with radius R. For $\tau > 0$ and $0 \leqslant k \leqslant K$, we define

$$\widehat{\Psi}^{(k)}_{\text{proj}}(\tau) \in \arg\min_{\Gamma \in \mathcal{B}_{\text{F}}(d^\tau)} \left\|\Gamma - \widehat{\Psi}^{(k)}\right\|_{\text{F}}. \tag{62}$$

$\widehat{\Psi}^{(k)}_{\text{proj}}(\tau)$ is utilized solely for theoretical reasons. Suppose that by choosing τ appropriately, we have

$$\mathbb{E}\left[\left\|\widehat{\Psi}^{(k)}_{\text{proj}}(\tau) - \Psi^{(k)}\right\|_{\text{F}}^2\right] \lesssim \bar{g}^{(k)}_{\text{F}}(n_0, n_k, d, h, M_\Gamma) := \bar{g}^{(k)}_{\text{F}} \quad \text{for all } 0 \leqslant k \leqslant K. \tag{63}$$

We define

$$\widehat{\Omega}^{(0)}_{\text{proj}}(\tau) = \sum_{k=0}^{K} \alpha_k \left(\breve{\Omega}^{(k)} - \widehat{\Psi}^{(k)}_{\text{proj}}(\tau)\right), \tag{64}$$

$\widehat{\Psi}^{(k)}_{\text{proj}}(\tau)$ is defined in (62). Combining (63) with Theorem 11 yields the following theorem with expected error upper bound for Trans-Glasso.

Theorem 12. *If the conditions of Theorem 11 are satisfied and* τ *is chosen so that* (63) *holds, we have*

$$\mathbb{E}\left[\left\|\widehat{\Omega}^{(0)}_{\text{proj}}(\tau) - \Omega^{(0)}\right\|_{\text{F}}^2\right] \lesssim \left(\frac{s}{N} + \frac{h}{\bar{n}}\right) \log d + \sum_{k=0}^{K} \alpha_k \bar{g}^{(k)}_{\text{F}}.$$

We then characterize $\bar{g}^{(k)}_{\text{F}}$ in the case where the D-Trace loss estimator in (7) is used. Recall that $\widehat{\Psi}^{(k)}_{\text{proj}}(\tau)$ is defined by (63).

Theorem 13. *Suppose that Assumption 1 and Assumption 2 hold. Assume that* $2(K+1) \leqslant d^{\tau_1}$*,* $N \leqslant d^{\tau_2}$ *and* $M_\Gamma \leqslant d^{\tau_3}$ *for some universal constants* $\tau_1, \tau_2, \tau_3 > 0$*. Let*

$$\lambda^{(k)}_\Gamma \asymp M_\Gamma \sqrt{\frac{\log d}{\min\{n_k, n_0\}}}.$$

Then, for any $\tau \geqslant \tau_3$*, we have*

$$\mathbb{E}\left[\left\|\widehat{\Psi}^{(k)}_{\text{proj}}(\tau) - \Psi^{(k)}\right\|_{\text{F}}^2\right] \lesssim \frac{M_\Gamma^2 \, h \log d}{\min\{n_k, n_0\}}.$$

Proof. Note that $\|A\|_{\text{F}} \leqslant |A|_1$ for any matrix A. Thus we have $\|\Psi^{(k)}\|_{\text{F}} \leqslant \|\Psi^{(k)}\|_1 \leqslant 2M_\Gamma$. The rest of the proof is similar to the proof of Theorem 11. ∎

By Theorem 13, we have

$$\bar{g}_{\mathrm{F}}^{(k)} = \frac{M_\Gamma^2\, h \log d}{\min\{n_k, n_0\}}$$

for D-Trace loss estimator. Plugging the above results into Theorem 12, we then have the following corollary.

Corollary 2. *Let $\widehat{\Omega}^{(0)}$ be obtained by Trans-Glasso (8) with the D-Trace loss estimator used in Step 1. Suppose that Assumption 1 and Assumption 2 hold, and that the conditions in Theorem 11 and Theorem 13 are satisfied. If $2(K+2) \leqslant d^{\tau_1}$, $N \leqslant d^{\tau_2}$, $M_\Gamma \leqslant d^{\tau_3}$ for some universal constants $\tau_1, \tau_2, \tau_3 > 0$, and*

$$\lambda_M \asymp \sqrt{\frac{\log d}{N}}, \quad \lambda_\Gamma^{(k)} \asymp M_\Gamma \sqrt{\frac{\log d}{\min\{n_k, n_0\}}} \quad \text{for all } k \in [K],$$

then for any $\tau \geqslant \tau_3$, we have

$$\mathbb{E}\left[\left\|\widehat{\Omega}_{\mathrm{proj}}^{(0)}(\tau) - \Omega^{(0)}\right\|_{\mathrm{F}}^2\right] \lesssim \left(\frac{s}{N} + (1+M_\Gamma^2)\cdot\frac{h}{\bar{n}} + M_\Gamma^2 \cdot \frac{h}{n_0}\right)\log d\,. \tag{65}$$

The estimation error in (65) is comprised of three parts: shared component estimation, individual component estimation, and differential network estimation. If $\bar{n} \geqslant n_0$ and M_Γ is bounded by a universal constant, the error scales as $\frac{s\log d}{N} + \frac{h\log d}{n_0}$. When using only target samples, the lowest error rate achievable is $\frac{(s+h)\log d}{n_0}$ as stated in Theorem 4. Therefore, if $N \gg n_0$, the error rate can be significantly reduced compared to the optimal rate obtained with only the target samples. Moreover, as demonstrated in Section 5.3, the rate $\frac{s\log d}{N} + \frac{h\log d}{n_0}$ is minimax optimal under certain conditions.

K Proof of Theorem 11

Note that for any matrix $A \in \mathbb{R}^{d\times d}$, we have $\|A\|_{\mathrm{F}} \leqslant \sqrt{d}\,\|A\|_2$. Since by our assumptions that $\left\|\Omega^{(k)}\right\|_2 = O(1)$ and $\left\|\breve{\Omega}^{(k)}\right\|_2 = O(1)$, we thus have

$$\left\|\Omega^{(k)}\right\|_{\mathrm{F}} = O\left(d^{\frac{1}{2}}\right) \quad \text{and} \quad \left\|\breve{\Omega}^{(k)}\right\|_{\mathrm{F}} = O\left(d^{\frac{1}{2}}\right).$$

For $\delta \in (0,1]$, let

$$\lambda_{\mathrm{M}} \geqslant C_1 C_3 \sqrt{\frac{\log(2(K+2)d^2/\delta)}{2N}},$$

where $C_1 = 160$ and $C_3 = M_\Sigma$. Besides, by the proof of Theorem 1, when

$$\frac{\log(2(K+2)d^2/\delta)}{\min_{0\leqslant k\leqslant K} n_k} \leqslant \frac{1}{2},$$

then by (43), we have $\mathbb{P}\{\mathbb{G}(\lambda_{\mathrm{M}})\} \geqslant 1-\delta$; or equivalently, we have $\mathbb{P}\left\{\bar{\mathbb{G}}(\lambda_{\mathrm{M}})\right\} \leqslant \delta$, where $\bar{\mathbb{G}}(\lambda_{\mathrm{M}})$ denotes the event that $\mathbb{G}(\lambda_{\mathrm{M}})$ does not hold.

Recall that by assumption we have $2(K+2) \leqslant d^{\tau_1}$. Let $\delta = d^{-\tau'}$, where τ' will be specified later, then by letting

$$\lambda_{\mathrm{M}} = C_1 C_3 \sqrt{\frac{(\tau'+\tau_1+2)\log d}{2N}}, \tag{66}$$

$$\frac{(\tau'+\tau_1+2)\log d}{\min_{0\leqslant k\leqslant K} n_k} \leqslant \frac{1}{2}, \tag{67}$$

we have $\mathbb{P}\left\{\bar{\mathbb{G}}(\lambda_{\mathrm{M}})\right\} \leqslant d^{-\tau'}$. Besides, by Section C.3, when (66)–(67) are true, $\mathbb{G}(\lambda_{\mathrm{M}})$ then implies that

$$\begin{aligned}\sum_{k=0}^{K} \alpha_k \left\|\breve{\Omega}^{(k)} - \Omega^{(k)}\right\|_{\mathrm{F}}^2 &\leqslant \frac{9\left(s + (K+1)h\right)\lambda_{\mathrm{M}}^2}{4\kappa^2} \\ &\leqslant \frac{9C_1^2C_3^2(\tau' + \tau_1 + 2)}{8}\left(\frac{s}{N} + \frac{h}{\bar{n}}\right)\log d.\end{aligned}$$

Note that

$$\begin{aligned}&\mathbb{E}\left[\sum_{k=0}^{K} \alpha_k \left\|\breve{\Omega}^{(k)} - \Omega^{(k)}\right\|_{\mathrm{F}}^2\right] \\ &= \mathbb{P}\left\{\mathbb{G}(\lambda_{\mathrm{M}})\right\}\mathbb{E}\left[\sum_{k=0}^{K} \alpha_k \left\|\breve{\Omega}^{(k)} - \Omega^{(k)}\right\|_{\mathrm{F}}^2 \middle| \mathbb{G}(\lambda_{\mathrm{M}})\right] + \mathbb{P}\left\{\bar{\mathbb{G}}(\lambda_{\mathrm{M}})\right\}\mathbb{E}\left[\sum_{k=0}^{K} \alpha_k \left\|\breve{\Omega}^{(k)} - \Omega^{(k)}\right\|_{\mathrm{F}}^2 \middle| \bar{\mathbb{G}}(\lambda_{\mathrm{M}})\right].\end{aligned} \tag{68}$$

Given (66)–(67) are true, we have

$$\mathbb{E}\left[\sum_{k=0}^{K} \alpha_k \left\|\breve{\Omega}^{(k)} - \Omega^{(k)}\right\|_{\mathrm{F}}^2 \middle| \mathbb{G}(\lambda_{\mathrm{M}})\right] \leqslant \frac{9C_1^2C_3^2(\tau' + \tau_1 + 2)}{8}\left(\frac{s}{N} + \frac{h}{\bar{n}}\right)\log d. \tag{69}$$

Besides, since $\|\breve{\Omega}^{(k)}\|_{\mathrm{F}}, \|\Omega^{(k)}\|_{\mathrm{F}} = O(d^{\frac{1}{2}})$ for all $0 \leqslant k \leqslant K$, we have

$$\sum_{k=0}^{K} \alpha_k \left\|\breve{\Omega}^{(k)} - \Omega^{(k)}\right\|_{\mathrm{F}}^2 \leqslant C'd,$$

for some constant $C' > 0$, and thus we have

$$\mathbb{P}\left\{\bar{\mathbb{G}}(\lambda_{\mathrm{M}})\right\}\mathbb{E}\left[\sum_{k=0}^{K} \alpha_k \left\|\breve{\Omega}^{(k)} - \Omega^{(k)}\right\|_{\mathrm{F}}^2 \middle| \bar{\mathbb{G}}(\lambda_{\mathrm{M}})\right] \leqslant C'd \cdot d^{-\tau'} = C'd^{-(\tau'-1)}.$$

By Assumption that we have $N \leqslant d^{\tau_2}$ where $\tau_2 > 0$, then when we choose τ' such that

$$\tau' \geqslant \tau_2 + 1 + \frac{\log\left(8C'/9C_1^2C_3^2(\tau' + \tau_1 + 2)\right)}{\log d},$$

we then have

$$\begin{aligned}C'd^{-(\tau'-1)} &\leqslant \frac{9C_1^2C_3^2(\tau' + \tau_1 + 2)}{8} \cdot \frac{1}{N} \\ &\leqslant \frac{9C_1^2C_3^2(\tau' + \tau_1 + 2)}{8}\left(\frac{s}{N} + \frac{h}{\bar{n}}\right)\log d,\end{aligned}$$

which then implies that

$$\mathbb{P}\left\{\bar{\mathbb{G}}(\lambda_{\mathrm{M}})\right\}\mathbb{E}\left[\sum_{k=0}^{K} \alpha_k \left\|\breve{\Omega}^{(k)} - \Omega^{(k)}\right\|_{\mathrm{F}}^2 \middle| \bar{\mathbb{G}}(\lambda_{\mathrm{M}})\right] \leqslant \mathbb{E}\left[\sum_{k=0}^{K} \alpha_k \left\|\breve{\Omega}^{(k)} - \Omega^{(k)}\right\|_{\mathrm{F}}^2 \middle| \mathbb{G}(\lambda_{\mathrm{M}})\right].$$

Combine the above inequality with (68) and (69), we finally have

$$\begin{aligned}\mathbb{E}\left[\sum_{k=0}^{K} \alpha_k \left\|\breve{\Omega}^{(k)} - \Omega^{(k)}\right\|_{\mathrm{F}}^2\right] &\leqslant 2\mathbb{E}\left[\sum_{k=0}^{K} \alpha_k \left\|\breve{\Omega}^{(k)} - \Omega^{(k)}\right\|_{\mathrm{F}}^2 \middle| \mathbb{G}(\lambda_{\mathrm{M}})\right] \\ &\leqslant \frac{18C_1^2C_3^2(\tau' + \tau_1 + 2)}{8}\left(\frac{s}{N} + \frac{h}{\bar{n}}\right)\log d \\ &\lesssim \left(\frac{s}{N} + \frac{h}{\bar{n}}\right)\log d.\end{aligned}$$