Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Transfer Learning in Large-Scale Gaussian Graphical Models with False Discovery Rate Control

Sai Li, T. Tony Cai & Hongzhe Li

To cite this article: Sai Li, T. Tony Cai & Hongzhe Li (2022): Transfer Learning in Large-Scale Gaussian Graphical Models with False Discovery Rate Control, Journal of the American Statistical Association, DOI: 10.1080/01621459.2022.2044333

To link to this article: https://doi.org/10.1080/01621459.2022.2044333

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2022, VOL. 00, NO. 0, 1–13: Theory and Methods
https://doi.org/10.1080/01621459.2022.2044333

Transfer Learning in Large-Scale Gaussian Graphical Models with False Discovery Rate Control

Sai Li[a,b], T. Tony Cai[c], and Hongzhe Li[b]

[a]Institute of Statistics and Big Data, Renmin University of China, Beijing, China; [b]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; [c]Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA

ABSTRACT

Transfer learning for high-dimensional Gaussian graphical models (GGMs) is studied. The target GGM is estimated by incorporating the data from similar and related auxiliary studies, where the similarity between the target graph and each auxiliary graph is characterized by the sparsity of a divergence matrix. An estimation algorithm, Trans-CLIME, is proposed and shown to attain a faster convergence rate than the minimax rate in the single-task setting. Furthermore, we introduce a universal debiasing method that can be coupled with a range of initial graph estimators and can be analytically computed in one step. A debiased Trans-CLIME estimator is then constructed and is shown to be element-wise asymptotically normal. This fact is used to construct a multiple testing procedure for edge detection with false discovery rate control. The proposed estimation and multiple testing procedures demonstrate superior numerical performance in simulations and are applied to infer the gene networks in a target brain tissue by leveraging the gene expressions from multiple other brain tissues. A significant decrease in prediction errors and a significant increase in power for link detection are observed. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received October 2020
Accepted February 2022

KEYWORDS
Debiased estimator; Inverse covariance matrix; Meta learning; Multiple testing

1. Introduction

Gaussian graphical models (GGMs), which represent the dependence structure among a set of random variables, have been widely used to model the conditional dependence relationships in many applications, including gene regulatory networks and brain connectivity maps (Varoquaux et al. 2010; Zhao, Cai, and Li 2014; Drton and Maathuis 2017; Glymour, Zhang, and Spirtes 2019). In the classical setting with data from a single study, the estimation of high-dimensional GGMs has been well studied in a series of articles, including penalized likelihood methods (Yuan and Lin 2007; Friedman, Hastie, and Tibshirani 2008; Rothman et al. 2008; Lam and Fan 2009), convex optimization-based methods (Cai, Liu, and Luo 2011; Cai, Liu, and Zhou 2016; Liu and Wang 2017), and penalized ℓ_1 log-determinant divergence (Ravikumar et al. 2011). Model selection has been considered in Ravikumar et al. (2008). The minimax optimal rates are studied in Cai, Liu, and Zhou (2016). Ren et al. (2015) studies the estimation optimality and inference for individual entries. A survey of optimal estimation of the structured high-dimensional GGMs can be found in Cai, Ren, and Zhou (2016). Liu (2013) considers the inference for GGMs based on a node-wise regression approach and introduces a multiple testing procedure for the partial correlations with the FDP and FDR control and Xia, Cai, and Cai (2015) studies simultaneous testing for the differential networks.

Methods for estimating a single GGM have also been extended to simultaneously estimating multiple graphs when data from multiple studies are available. For example, Guo et al. (2011), Chiquet, Grandvalet, and Ambroise (2011), Danaher, Wang, and Witten (2014), and Cai et al. (2016) consider jointly estimating multiple GGMs by employing some penalties in order to induce common structures among different graphs. This problem falls in the category of multi-task learning (Lounici et al. 2009; Agarwal, Negahban, and Wainwright 2012), whose goal is to jointly estimate several related graphs.

Due to high dimensionality and relatively small sample sizes in many modern applications, estimation of GGMs based on a single study often has large uncertainty and low power in link detection. However, the blessing is that samples from some different but related studies can be abundant. Particularly, for a given target study, there might be other studies where we expect some similar dependence structures among the same set of variables. One example is to infer the gene regulatory networks among a set of genes for a given issue. Although gene regulatory networks are expected to vary from tissue to tissue, certain shared regulatory structures are expected and have indeed been observed (Pierson et al. 2015; Fagny et al. 2017). This article introduces a transfer learning approach to improve the estimation and inference accuracy for the gene regulatory network in one target tissue by incorporating the data in other tissues.

CONTACT Hongzhe Li *hongzhe@upenn.edu* Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104.

Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

Transfer learning techniques have been developed in a range of applications, including pattern recognition, natural language processing, and drug discovery (Pan and Yang 2009; Turki, Wei, and Wang 2017; Bastani 2021). Transfer learning has been studied in different settings with various similarity measures, but only a few of them offer statistical guarantees. Cai and Wei (2021) investigates nonparametric classification in transfer learning and proposes minimax and adaptive classifiers. In linear regression models, Li, Cai, and Li (2021) considers the estimation of high-dimensional regression coefficient vectors when the difference between the auxiliary model and the target model is sufficiently sparse and proves the minimax optimal rate. Tripuraneni, Jin, and Jordan (2021) proposes an algorithm that assumes all the auxiliary studies and the target study share a common, low-dimensional linear representation. Transfer learning in general functional classes has been studied in Tripuraneni, Jordan, and Jin (2020) and Hanneke and Kpotufe (2020). Loosely speaking, transfer learning aims to improve the learning accuracy for the target study by transferring information from multiple related studies. This is different from the multi-task learning outlined above, where the goal is to simultaneously estimate multiple graphs. In terms of theoretical results, the error criteria for these two learning frameworks are different and are incomparable in general. The convergence rate for estimating the target graph in transfer learning can be faster than the corresponding rate in multi-task learning.

1.1. Model Set-Up

Suppose that we observe iid samples $x_i \in \mathbb{R}^p$ generated from $N(\mu, \Sigma)$, $i = 1, \ldots, n$, and the parameter of interest is the precision matrix $\Omega = \Sigma^{-1}$. Indeed, Ω uniquely determines the conditional dependence structure and the corresponding graph. If the i-th and j-th variables are conditionally dependent in the target study, there is an undirected edge between the i-th and j-th nodes in the Gaussian graph and, equivalently, the (i, j)-th and (j, i)-th entries of Ω are nonzero. Our focus is on the estimation and inference for high-dimensional sparse Gaussian graphs where p can be much larger than n and Ω is sparse such that each column of Ω has at most s nonzero elements with $s \ll p$.

In the transfer learning setting, besides the observations $\{x_1, \ldots, x_n\}$ from the target distribution $N(\mu, \Sigma)$, we also observe samples from K auxiliary studies. For $k = 1, \ldots, K$, the observations $x_i^{(k)} \in \mathbb{R}^p$ are independently generated from $N(\mu^{(k)}, \Sigma^{(k)})$, $i = 1, \ldots, n_k$. Let $\Omega^{(k)} = \{\Sigma^{(k)}\}^{-1}$ be the precision matrix of the kth study, $k = 1, \ldots, K$. If some knowledge can be transferred to the target study, a certain level of similarity needs to be possessed by the auxiliary models and the target one.

To motivate our proposed similarity measure, consider the relative entropy, or equivalently the Kullback–Leibler (KL) divergence, between the kth auxiliary model and the target model. That is,

$$\mathcal{D}_{KL}(N_{\Sigma^{(k)}} \,\|\, N_\Sigma) = \frac{1}{2}\mathrm{Tr}(\Delta^{(k)}) - \frac{1}{2}\log\det(I_p + \Delta^{(k)}) \text{ for } \Delta^{(k)} = \Omega\Sigma^{(k)} - I_p, \tag{1}$$

where $N_{\Sigma^{(k)}}$ and N_Σ denote the normal distributions with mean zero and covariance matrix $\Sigma^{(k)}$ and Σ, respectively. The KL-divergence is parametrized by the matrix $\Delta^{(k)}$ and we call $\Delta^{(k)}$ the kth divergence matrix. We characterize the difference between Ω and $\Omega^{(k)}$ via

$$\mathcal{D}_q(\Omega^{(k)}, \Omega) = \max_{1\le j\le p} ||\Delta^{(k)}_{j,\cdot}||_q + \max_{1\le j\le p} ||\Delta^{(k)}_{\cdot,j}||_q \tag{2}$$

for some fixed $q \in [0, 1]$. In words, $\mathcal{D}_q(\Omega, \Omega^{(k)})$ is the maximum row-wise ℓ_q-sparsity of $\Delta^{(k)}$ plus the maximum column-wise ℓ_q-sparsity of $\Delta^{(k)}$. Both the row-wise and column-wise norms are taken into account because $\Delta^{(k)}$ is nonsymmetric. The quantity $\mathcal{D}_q(\Omega^{(k)}, \Omega)$ measures the "relative distance" between Ω and $\Omega^{(k)}$ in the sense that $\mathcal{D}_q(\Omega^{(k)}, \Omega) = \mathcal{D}_q(c\Omega^{(k)}, c\Omega)$ for any constant $c > 0$. Notice that the spectral norm of $\Delta^{(k)}$ is upper bounded by $\mathcal{D}_1(\Omega^{(k)}, \Omega)$, which further provides an upper bound on the KL-divergence.

In this work, we develop estimation and inference procedures for GGMs under the similarity characterization (2) for any fixed $q \in [0, 1]$. We focus on the methods and theory when $q = 1$ in the main article and provide matching minimax upper and lower bounds for $q \in [0, 1)$ in the supplementary materials.

1.2. Our Contributions

A transfer learning algorithm, called Trans-CLIME, is proposed for estimating the target GGM. Inspired by the CLIME in the single-task setting (Cai, Liu, and Luo 2011), the proposed algorithm includes additional steps to incorporate auxiliary information. Furthermore, edge detection with uncertainty quantification is considered. We introduce a universal debiasing method that can be coupled with many initial graph estimators, including both the single-task and the transfer learning estimators. The debiasing step can be analytically computed in one step. We demonstrate the asymptotic normality under certain conditions. Applying this procedure, we construct the confidence interval for an edge of interest and propose a multiple testing procedure for all the edges with false discovery rate (FDR) control.

Theoretically, we establish the minimax optimal rate of convergence for estimating the GGMs with transfer learning in the Frobenius norm by providing matching minimax upper and lower bounds. We also establish the optimal rate of convergence for estimating individual entries in the graph. These convergence rates are faster than the corresponding minimax rates in the classical single-task setting, where no auxiliary samples are available or used. Our proposed Trans-CLIME and debiased Trans-CLIME are shown to be rate optimal for different error criteria under proper conditions.

1.3. Organization and Notation

The rest of this article is organized as follows. In Section 2, we propose an algorithm for graph estimation with transfer learning given $q = 1$ in the similarity characterization. In Section 3, we study statistical inference for each edge of the graph. In Section 4, we consider multiple testing of all the edges in the graph with false discovery rate guarantee. In Section 5, we establish the minimax lower bounds for any fixed $q \in [0, 1]$.

In Section 6, we study the numerical performance of Trans-CLIME in comparison to some other relevant methods. We then present an application of the proposed methods to estimate gene regulatory graphs based on data from multiple brain tissues in Section 7. Section 8 concludes the article. The proofs and other supporting information are given in the supplementary materials.

For a matrix $A \in \mathbb{R}^{p\times p}$, let A_j denote the jth column of A. For any fixed $j \le p$, we call $||A_j||_2$ the column-wise ℓ_2-norm of A. Let $||A||_{\infty,q} = \max_{j\le p} ||A_j||_q$ for $q > 0$ and $||A||_{\infty,\infty} = \max_{i,j\le p} |A_{i,j}|$, and $||A||_1 = \sum_{j=1}^{p} ||A_j||_1$. Let $||A||_2$ denote the spectral norm of A and $||A||_F$ denote the Frobenius norm of A. For a vector $v \in \mathbb{R}^p$, let $||v||_0$ denote the number of nonzero elements of v. For a symmetric matrix A, let $\Lambda_{\max}(A)$ and $\Lambda_{\min}(A)$ denote the largest and smallest eigenvalues of A, respectively. Let $\Phi(t)$ denote the standard normal probability function. Let z_q denote the qth quantile of standard normal distribution. We use $c_0, c_1, \ldots$ and $C_0, C_1, \ldots$ as generic constants which can be different at different places.

2. GGM Estimation with Transfer Learning

In this section, we study GGM estimation based on transfer learning. In Section 2.1, we introduce the rationale for the proposed algorithm. The proposal is introduced in Section 2.2 and its theoretical properties are studied in Section 2.3.

2.1. Rationale and Moment Equations

Moment equations provide a powerful tool for deriving estimation methods in parametric models. By the definition of Ω, it is natural to consider the following moment equation:

$$\Sigma\Omega - I_p = 0. \tag{3}$$

The idea of CLIME (Cai, Liu, and Luo 2011) is to solve an empirical version of (3) and to encourage the sparsity of the estimator. In the context of transfer learning, we re-express the moment equation (3) to incorporate auxiliary information. Specifically, for $k = 1, \ldots, K$,

$$I_p = \Sigma^{(k)}\Omega^{(k)} = \Sigma^{(k)}\Omega - (\Delta^{(k)})^\intercal, \tag{4}$$

where $\Delta^{(k)}$ is the divergence matrix defined in (1). We see from (4) that $\Delta^{(k)}$ corresponds to the bias in the moment equations when we try to identify Ω based on the k-th study. To simultaneously leverage all the auxiliary studies, we further define the weighted average of the covariance and divergence matrices

$$\Sigma^{\mathcal{K}} = \sum_{k=1}^{K} \alpha_k \Sigma^{(k)} \text{ and } \Delta^{\mathcal{K}} = \sum_{k=1}^{K} \alpha_k \Delta^{(k)}, \tag{5}$$

where $\alpha_k = n_k/N$ for $N = \sum_{k=1}^{K} n_k$. For knowledge transfer, the moment equation considered for Ω is

$$\Sigma^{\mathcal{K}}\Omega - (\Delta^{\mathcal{K}})^\intercal - I_p = 0, \tag{6}$$

where $\Sigma^{\mathcal{K}}$ is an average parameter over the K source studies and it incorporates the auxiliary information. The moment equation (6) motivates our procedure. First, we will estimate $\Delta^{\mathcal{K}}$ based on the following moment equation:

$$\Sigma\Delta^{\mathcal{K}} - (\Sigma^{\mathcal{K}} - \Sigma) = 0. \tag{7}$$

Once $\Delta^{\mathcal{K}}$ is identified, we can estimate our target Ω via (6).

In some practical scenarios, the similarity between $\Omega^{(k)}$ and Ω can be weak, for some $1 \le k \le K$, that is, $\mathcal{D}_q(\Omega^{(k)}, \Omega)$ can be large. In fact, the similarity is often unknown a priori in practice. In this case, information transfer may negatively affect the learning performance of the target problem, which is also known as the "negative transfer" (Hanneke and Kpotufe 2020). To address this issue, we will further perform an aggregation step. The aggregation methods and theory have been extensively studied in the existing literature in regression problems, to name a few, Rigollet and Tsybakov (2011), Tsybakov (2014) and Lecué and Rigollet (2014). This type of methods can guarantee that, loosely speaking, the aggregated estimator has prediction performance comparable to the best prediction performance which could be achieved by the initial estimators.

2.2. Trans-CLIME Algorithm

We introduce our proposed transfer learning algorithm, Trans-CLIME. We randomly split the data from the target study into two disjoint folds. Specifically, let $\mathcal{I}$ be a random subset of $\{1, \ldots, n\}$ such that $|\mathcal{I}| = cn$ for some constant $0 < c < 1$. Let $\mathcal{I}^c$ denote the complement of $\mathcal{I}$. Let $\bar{x} = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} x_i$, $\tilde{x} = \frac{1}{|\mathcal{I}^c|}\sum_{i\in\mathcal{I}^c} x_i$,

$$\widehat{\Sigma} = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} x_i x_i^\intercal - \bar{x}\bar{x}^\intercal \text{ and } \widetilde{\Sigma} = \frac{1}{|\mathcal{I}^c|}\sum_{i\in\mathcal{I}^c} x_i x_i^\intercal - \tilde{x}\tilde{x}^\intercal. \tag{8}$$

We will use $\widehat{\Sigma}$ for constructing estimators and will use $\widetilde{\Sigma}$ for aggregation. Let $\tilde{n} = |\mathcal{I}^c|$. Let $\bar{x}^{\mathcal{K}} = \sum_{k=1}^{K}\sum_{i=1}^{n_k} x_i^{(k)}/N$ denote the sample mean and $\widehat{\Sigma}^{\mathcal{K}} = \sum_{k=1}^{K}\sum_{i=1}^{n_k} x_i^{(k)}(x_i^{(k)})^\intercal/N - \bar{x}^{\mathcal{K}}(\bar{x}^{\mathcal{K}})^\intercal$ denote the sample covariance based on the auxiliary samples. To begin with, we compute the single-task CLIME $\widehat{\Omega}^{(\mathrm{CL})}$ such that

$$\begin{aligned}\widehat{\Omega}^{(\mathrm{CL})} = &\arg\min_{\Omega\in\mathbb{R}^{p\times p}} ||\Omega||_1 \\ &\text{subject to } \left\|(\widehat{\Sigma} + |\mathcal{I}|^{-1}I_p)\Omega - I_p\right\|_{\infty,\infty} \le \lambda_{\mathrm{CL}},\end{aligned} \tag{9}$$

where $\widehat{\Sigma}$ is defined in (8) and $\lambda_{\mathrm{CL}} > 0$ is a tuning parameter. A diagonal matrix $|\mathcal{I}|^{-1}I_p$ is added to $\widehat{\Sigma}$ for making the sample covariance matrix positive definite. This modification has been considered in a refined version of the CLIME (Cai, Ren, and Zhou 2016).

Step 1. We estimate $\Delta^{\mathcal{K}}$ based on the moment equation (7). Compute

$$\begin{aligned}\widehat{\Delta}^{\mathcal{K}} = &\arg\min_{\Delta\in\mathbb{R}^{p\times p}} ||\Delta||_1 \\ &\text{subject to } \quad ||\Delta - \{(\widehat{\Omega}^{(\mathrm{CL})})^\intercal\widehat{\Sigma}^{\mathcal{K}} - I_p\}||_{\infty,\infty} \le \lambda_\Delta.\end{aligned} \tag{10}$$

The optimization (10) can be understood as an adaptive thresholding of an initial estimate $\Delta^{\mathcal{K}}$, $(\widehat{\Omega}^{(\mathrm{CL})})^\intercal\widehat{\Sigma}^{\mathcal{K}} - I_p$. This initial estimate is inspired by the moment equation (7). We further explain that the CLIME estimator $\widehat{\Omega}^{(\mathrm{CL})}$ is not necessarily symmetric and we use $(\widehat{\Omega}^{(\mathrm{CL})})^\intercal$ in step 1 to directly leverage the

constraints in the CLIME optimization. The optimization in (10) is a more sophisticated version of hard thresholding and it is designed for the approximate sparse parameter $\Delta^{\mathcal{K}}$. The estimate $\widehat{\Delta}^{\mathcal{K}}$ is row-wise and column-wise ℓ_1-sparse and will be used in the next step.

Step 2. For $\widehat{\Delta}^{\mathcal{K}}$ defined in (10), compute

$$\widehat{\Theta} = \arg\min_{\Theta\in\mathbb{R}^{p\times p}} ||\Theta||_1 \quad (11)$$
$$\text{subject to } \left|\left|(\widehat{\Sigma}^{\mathcal{K}} + N^{-1}I_p)\Theta - (\widehat{\Delta}^{\mathcal{K}} + I_p)^{\intercal}\right|\right|_{\infty,\infty} \le \lambda_\Theta.$$

This step is a CLIME-type optimization based on the moment equation (6) and it outputs an estimator of the target graph Ω. When the similarities between auxiliary studies and the target study are sufficiently high, $\widehat{\Theta}$ is a desirable transfer learning estimator of Ω.

As we have discussed in Section 2.1, $\widehat{\Theta}$ may not be as good as the single-task estimator if the similarity is weak. Hence, we perform a model selection aggregation in Step 3 based on the single-task CLIME estimator and $\widehat{\Theta}$ to produce a final graph estimator. Loosely speaking, we compute a weight vector $\hat{v}_j$ motivated by the moment equation

$$\widetilde{\Sigma}(\widehat{\Omega}_j^{(\mathrm{CL})}, \widehat{\Theta}_j)v_j - e_j \approx 0.$$

Notice that the sample splitting step guarantees that both $\widehat{\Theta}$ and $\widehat{\Omega}^{(\mathrm{CL})}$ are independent of the samples used for aggregation, $\widetilde{\Sigma}$.

Step 3. For $j = 1, \ldots, p$, compute

$$\widehat{W}(j) = \begin{pmatrix} (\widehat{\Omega}_j^{(\mathrm{CL})})^{\intercal}\widetilde{\Sigma}\widehat{\Omega}_j^{(\mathrm{CL})} & (\widehat{\Omega}_j^{(\mathrm{CL})})^{\intercal}\widetilde{\Sigma}\widehat{\Theta}_j \\ (\widehat{\Omega}_j^{(\mathrm{CL})})^{\intercal}\widetilde{\Sigma}\widehat{\Theta}_j & \widehat{\Theta}_j^{\intercal}\widetilde{\Sigma}\widehat{\Theta}_j \end{pmatrix}$$

and

$$\hat{v}_j = \arg\min_{v\in\{(0,1)^{\intercal},(1,0)^{\intercal}\}} v^{\intercal}\widehat{W}(j)v - 2v^{\intercal}(\widehat{\Omega}_{j,j}^{(\mathrm{CL})}, \widehat{\Theta}_{j,j})$$

where $\widehat{\Omega}^{(\mathrm{CL})}$ is defined in (9) and $\widehat{\Theta}$ is defined in (11). For $j = 1, \ldots, p$, let

$$\widehat{\Omega}_j = (\widehat{\Omega}_j^{(\mathrm{CL})}, \widehat{\Theta}_j)\hat{v}_j. \quad (12)$$

We summarize the formal algorithm as follows.

Algorithm 1: Trans-CLIME algorithm

Input : Target data (after a random sample splitting) $\{\widehat{\Sigma}, \widetilde{\Sigma}\}$ and auxiliary samples $\widehat{\Sigma}^{\mathcal{K}}$.
Output: $\widehat{\Omega}$.
Step 1. Compute $\widehat{\Delta}^{\mathcal{K}}$ via (10).
Step 2. Compute $\widehat{\Theta}$ via (11).
Step 3. Aggregation for positive transfer: For $j = 1, \ldots, p$, compute $\widehat{\Omega}_j$ via (12).

Computationally, all the optimizations in these three steps can be separated into p independent optimizations, analogous to the original CLIME algorithm. This makes the computation scalable. While $\widehat{\Omega}$ is not symmetric in general, one can use $(\widehat{\Omega} + \widehat{\Omega}^{\intercal})/2$ as a symmetric estimate for Ω. It is not hard to show that $(\widehat{\Omega} + \widehat{\Omega}^{\intercal})/2$ has the same convergence rate as $\widehat{\Omega}$ in Frobenius norm.

2.3. Convergence Rate of Trans-CLIME

In this section, we provide theoretical guarantees for the Trans-CLIME algorithm. We assume the following condition in our theoretical analysis.

Condition 2.1 (Gaussian graphs). For $i = 1, \ldots, n$, $x_i \in \mathbb{R}^p$ are iid distributed as $N(\mu, \Sigma)$. For each $1 \le k \le K$, $x_i^{(k)}$ are iid distributed as $N(\mu^{(k)}, \Sigma^{(k)})$ for $i = 1, \ldots, n_k$. It holds that $1/C \le \Lambda_{\min}(\Sigma) \le \Lambda_{\max}(\Sigma) \le C$ and $1/C \le \min_{1\le k\le K}\Lambda_{\min}(\Sigma^{(k)}) \le \max_{1\le k\le K}\Lambda_{\max}(\Sigma^{(k)}) \le C$ for some constant $C > 0$.

The Gaussian assumption facilitates the justification of the restricted eigenvalue conditions of the empirical covariance matrices. The Gaussian distribution of the target data also simplifies the limiting distribution of our proposed estimator for inference.

The parameter space we consider is

$$\mathbb{G}_q(s,h) = \Big\{(\Omega, \Omega^{(1)}, \ldots, \Omega^{(K)}) : \max_{1\le j\le p} ||\Omega_j||_0 \le s, \max_{1\le k\le K} \mathcal{D}_q(\Omega, \Omega^{(k)}) \le h\Big\}. \quad (13)$$

We mention that the parameter space for GGMs in the single-task setting (Ren et al. 2015) can be written as $\mathbb{G}_q(s,\infty)$ under Condition 2.1 for any $q \in [0,1]$. This is because $\mathbb{G}_q(s,\infty)$ allows the auxiliary studies to be arbitrarily far away from the target study and the worst-case scenario is equivalent to the setting where only the target data is available.

In the following, we demonstrate the convergence rate of Trans-CLIME under Condition 2.1. Let $\delta_{h,\Omega} = \{(||\Omega||_{\infty,1} + h)h\sqrt{\log p/n}\} \wedge h^2$ and $\delta_h = h\sqrt{\log p/n} \wedge h^2$. We see that $\delta_{h,\Omega} \gtrsim \delta_h$ as $||\Omega||_{\infty,1} \ge c > 0$. If $\max\{||\Omega||_{\infty,1}, h\}$ is finite, then $\delta_{h,\Omega} \asymp \delta_h$.

Theorem 2.1 (Convergence rate of Trans-CLIME). Assume Condition 2.1. Let the Trans-CLIME estimator $\widehat{\Omega}$ be computed with

$$\lambda_\Delta = c_1(||\Omega||_{\infty,1} + h)\sqrt{\frac{\log p}{n}} \text{ and } \lambda_\Theta = c_1\sqrt{\frac{\log p}{N}},$$

where c_1 is a large enough constant. If $p \ge c_2 N$ for some positive constant c_2 and $s\log p/N + \delta_{h,\Omega} \wedge s\log p/n = o(1)$, then for any true models in $\mathbb{G}_1(s,h)$, we have

$$\mathbb{E}\left[\frac{1}{p}||\widehat{\Omega} - \Omega||_F^2 \vee ||\widehat{\Omega}_j - \Omega_j||_2^2\right] \le C\left(\frac{s\log p}{N+n} + \delta_{h,\Omega} \wedge \frac{s\log p}{n} + \frac{1}{n}\right) \quad (14)$$

for any fixed $1 \le j \le p$ and some positive constant C.

Theorem 2.1 demonstrates that under a proper choice of the tuning parameters, the upper bounds can be obtained in Frobenius norm and in column-wise ℓ_2-norm. We first illustrate the convergence rate of $\widehat{\Omega}$ in column-wise ℓ_2-norm. As all the $\Omega^{(k)}$, $k = 1, \ldots, K$, share the column-wise s-sparse matrix Ω, the term $s\log p/(N+n)$ comes from estimating Ω based on $N + n$ independent samples. The term $\delta_{h,\Omega}$ comes from the

errors of estimating $\Delta^{\mathcal{K}}$ in row-wise ℓ_2-norm. It goes to zero as the target sample size n goes to infinity. The term $\delta_{h,\Omega}$ is determined by the target sample size because the divergence matrix is defined and can only be identified based on the target samples. Nevertheless, $\delta_{h,\Omega}$ can still be a fast rate when the similarity among these studies is high, that is, h is small. The minimal term $\delta_{h,\Omega} \wedge s\log p/n$ is a consequence of the aggregation performed in Step 3, where $s\log p/n$ is the single-task convergence rate under the same distance measure. However, there is a mild cost of aggregation, which is of order n^{-1} shown in the last term in (14), and it is negligible in most parameter spaces of interest.

To understand the gain of transfer learning, we compare the current results with the convergence rate of CLIME in the single-task setting.

Remark 2.1 (*Convergence rate of single-task CLIME*). Assume Condition 2.1, $s^2\log p = o(n)$ and $p \geq c_1 n$ for some positive constant c_1. For the CLIME estimator $\widehat{\Omega}^{(\mathrm{CL})}$ defined in (9) with $\lambda_{\mathrm{CL}} = c_2\sqrt{\log p/n}$ with large enough constant c_2, then for any true models in $\mathbb{G}_1(s,\infty)$,

$$\mathbb{E}\left[\frac{1}{p}||\widehat{\Omega}^{(\mathrm{CL})} - \Omega||_F^2 \vee ||\widehat{\Omega}_j^{(\mathrm{CL})} - \Omega_j||_2^2\right] \leq \frac{Cs\log p}{n}$$

for any fixed $1 \leq j \leq p$ and some positive constant C.

We see that the convergence rate of Trans-CLIME is no worse than CLIME in Frobenius norm for any $s \geq 1$, which is a consequence of aggregation. Furthermore, Trans-CLIME has faster convergence rate when $N \gg n$ and $\delta_{h,\Omega} \ll s\sqrt{\log p/n}$. That is, if the total auxiliary sample size is dominant and the similarity is sufficiently strong (h is sufficiently small), then the learning performance can be significantly improved using Trans-CLIME. This demonstrates the gain of transfer learning in estimating the graphical models.

Remark 2.2 (*Faster convergence rate in a restricted regime*). Assume Condition 2.1. Let the Trans-CLIME estimator $\widehat{\Omega}$ be computed with

$$\lambda_\Delta = c_1\sqrt{\frac{\log p}{n}} \text{ and } \lambda_\Theta = c_1\sqrt{\frac{\log p}{N}},$$

where c_1 is a large enough constant. If $p \geq c_2 N$ and $s^2\log p \leq c_3 n$ for some positive constants c_2 and c_3, then for any true models in $\mathbb{G}_1(s,h)$, we have

$$\mathbb{E}\left[\frac{1}{p}||\widehat{\Omega} - \Omega||_F^2 \vee ||\widehat{\Omega}_j - \Omega_j||_2^2\right] \leq C\left(\frac{s\log p}{N+n} + \delta_h \wedge \frac{s\log p}{n} + \frac{1}{n}\right) \quad (15)$$

for any fixed $1 \leq j \leq p$ and some positive constant C.

In Remark 2.2, we prove another convergence rate of Trans-CLIME when $s^2\log p \lesssim n$. Its difference from (14) is that $\delta_{h,\Omega}$ is replaced by δ_h in (15). Hence, the convergence rate in (15) is sharper than the rate in (14) when $\max\{||\Omega||_{\infty,1}, h\}$ grows to infinity and is of the same order of the rate in (14) when $\max\{||\Omega||_{\infty,1}, h\}$ is finite. However, the condition on s in Remark 2.2 is stronger than the one assumed in Theorem 2.1 and it is indeed the same as the requirement in the single-task problems (Cai, Ren, and Zhou 2016; Liu and Wang 2017). In fact, we will show in Section 5.1 that the minimax lower bound in Frobenius norm is $s\log p/(N+n) + \delta_h \wedge (s\log p/n)$ for the current parameter space. The convergence rate of Trans-CLIME in (15) has one more term $1/n$, which is the cost when the relative magnitude of h and s is unknown a priori. In most nontrivial parameter spaces, Trans-CLIME enjoys minimax optimality when $s^2\log p \leq c_0 n$ or when $\max\{||\Omega||_{\infty,1}, h\}$ is finite.

3. Entry-Wise Inference

We propose in this section a debiasing procedure for inference of individual entries in the graph. The main features of our proposal are its flexibility to couple with a broad range of initial graph estimators and its computational efficiency. We first introduce a universal debiasing method and study its theoretical guarantees in Section 3.1. We will then use it for the construction of confidence intervals in the transfer learning setting in Section 3.2.

3.1. A Universal Debiasing Method

Our procedure on inference for $\Omega_{i,j}$ is inspired by the idea of debiasing quadratic forms. We begin by expressing $\Omega_{i,j}$ as a quadratic form:

$$\Omega_{i,j} = \Omega_i^\intercal \Sigma \Omega_j = \Omega_i^\intercal \mathbb{E}[\Sigma^n]\Omega_j, \quad (16)$$

where Σ^n denotes the sample covariance matrix based on the target data. In many occasions, Σ^n can be computed based on all the target data. Sometimes for a sharp theoretical analysis, sample splitting is performed and Σ^n can be computed based on a constant proportion of the target data. We call the samples involved in Σ^n the debiasing samples. Equation (16) holds for any inverse covariance matrix Ω not restricting to Gaussian random graphs.

Leveraging (16), we are able to use the idea of debiasing quadratic forms (Cai and Guo 2020) to make inference for $\Omega_{i,j}$. Specifically, $\Omega_{i,j}$ takes the same format as the coheritability if we view Ω_i and Ω_j as the regression coefficient vectors for two different outcomes and view X as the measurements of genetic variants. Motivated by this observation, we arrive at the following debiased estimator of $\Omega_{i,j}$. Let $\widehat{\Omega}^{(\mathrm{init})}$ be any initial estimator of Ω. The corresponding debiased estimator is

$$\begin{aligned}\mathring{\Omega}_{i,j}^{(db)} &= (\widehat{\Omega}_i^{(\mathrm{init})})^\intercal \Sigma^n \widehat{\Omega}_j^{(\mathrm{init})} + (\widehat{\Omega}_i^{(\mathrm{init})})^\intercal (e_j - \Sigma^n \widehat{\Omega}_j^{(\mathrm{init})}) \\ &\quad + (\widehat{\Omega}_j^{(\mathrm{init})})^\intercal (e_i - \Sigma^n \widehat{\Omega}_i^{(\mathrm{init})}) \\ &= \widehat{\Omega}_{j,i}^{(\mathrm{init})} + \widehat{\Omega}_{i,j}^{(\mathrm{init})} - (\widehat{\Omega}_j^{(\mathrm{init})})^\intercal \Sigma^n \widehat{\Omega}_i^{(\mathrm{init})}. \end{aligned} \quad (17)$$

We mention that $\widehat{\Omega}^{(\mathrm{init})}$ is not necessarily symmetric and hence, we distinguish $\widehat{\Omega}_{i,j}^{(\mathrm{init})}$ and $\widehat{\Omega}_{i,j}^{(\mathrm{init})}$. It is easy to see that the above debiasing procedure can be coupled with any $\widehat{\Omega}^{(\mathrm{init})}$ which has a sufficiently fast convergence rate in column-wise ℓ_2-norm. Hence, the candidate estimators for debiasing include, say, graphical Lasso (Friedman, Hastie, and Tibshirani 2008), CLIME (Cai, Liu, and Luo 2011), multi-task graph estimators (Guo et al. 2011; Danaher, Wang, and Witten 2014; Cai et al.

2016), and our proposed Trans-CLIME. In comparison to Liu (2013) and Ren et al. (2015), where the debiased estimators are constructed using ℓ_1-penalized node-wise regression, our proposal in (17) is more flexible in incorporating various types of initial estimators under different structural assumptions. To the best of our knowledge, $\mathring{\Omega}^{(\mathrm{db})}$ is the first universal debiasing method for graph estimators.

Besides the general applicability, our debiased estimator for the whole graph can be analytically computed in one step given the initial graph estimator $\widehat{\Omega}^{(\mathrm{init})}$. Specifically, the debiased estimator in (17) can be re-expressed as

$$\mathring{\Omega}^{(\mathrm{db})} = \widehat{\Omega}^{(\mathrm{init})} + (\widehat{\Omega}^{(\mathrm{init})})^{\intercal} - (\widehat{\Omega}^{(\mathrm{init})})^{\intercal} \Sigma^{n} \widehat{\Omega}^{(\mathrm{init})}.$$

To demonstrate the power of this universal debiasing method, we first prove the asymptotic normality for the debiased single-task CLIME estimator, which is

$$\widehat{\Omega}_{i,j}^{(\mathrm{db\text{-}CL})} = \widehat{\Omega}_{j,i}^{(\mathrm{CL})} + \widehat{\Omega}_{i,j}^{(\mathrm{CL})} - (\widehat{\Omega}_{j}^{(\mathrm{CL})})^{\intercal} \widehat{\Sigma} \widehat{\Omega}_{i}^{(\mathrm{CL})},$$

where $\widehat{\Sigma}$ is defined in (8).

Proposition 3.1 (Asymptotic normality for debiased CLIME). Assume Condition 2.1 and $s \log p \ll \sqrt{n}$. For any fixed $i \neq j$,

$$\frac{\sqrt{n}(\widehat{\Omega}_{i,j}^{(\mathrm{db\text{-}CL})} - \Omega_{i,j})}{\sqrt{V_{i,j}}} \xrightarrow{D} N(0,1)$$

for $V_{i,j} = \Omega_{i,i}\Omega_{j,j} + \Omega_{i,j}^2$.

The asymptotic normality of $\widehat{\Omega}_{i,j}^{(\mathrm{db\text{-}CL})}$ requires that $s \log p \ll \sqrt{n}$. The condition and the asymptotic distribution in Proposition 3.1 recover the results in Ren et al. (2015) and $\widehat{\Omega}_{i,j}^{(\mathrm{db\text{-}CL})}$ is indeed minimax optimal in $\mathbb{G}_q(s, \infty)$ for estimating $\Omega_{i,j}$. This demonstrates the optimality of this universal debiasing method.

3.2. Entry-Wise Confidence Intervals Based on Transfer Learning

Applying the universal debiasing scheme to the Trans-CLIME estimator $\widehat{\Omega}$, we arrive at the following debiased Trans-CLIME estimator for $\Omega_{i,j}$

$$\widehat{\Omega}_{i,j}^{(\mathrm{db})} = \widehat{\Omega}_{j,i} + \widehat{\Omega}_{i,j} - \widehat{\Omega}_{j}^{\intercal} \widetilde{\Sigma} \widehat{\Omega}_{i}. \tag{18}$$

In (18), we only use a proportion of target data, that is, those involved in $\widetilde{\Sigma}$, as debiasing samples, while the realization of $\widehat{\Omega}$ involves both target and auxiliary samples. This is because first, only the target data are known to be unbiased; second, the samples involved in $\widetilde{\Sigma}$ has mild dependence on $\widehat{\Omega}$ which is induced by the aggregation step and it allows us to prove the desirable convergence rate. Ideally, one can use some debiasing samples independent of $\widehat{\Omega}$, which can be achieved through another sample splitting. In practice, sample splitting always leads to sub-optimal empirical performance and hence, we analyze (18) and take care of the dependence through careful analysis.

Theorem 3.1 (Asymptotic normality for debiased Trans-CLIME). Assume Condition 2.1 and the sample size condition stated in (20). For any true models in $\mathbb{G}_1(s,h)$ and any fixed $1 \leq i,j \leq p$, the debiased Trans-CLIME satisfies

$$\frac{\sqrt{\tilde{n}}(\widehat{\Omega}_{i,j}^{(\mathrm{db})} - \Omega_{i,j})}{\sqrt{V_{i,j}}} \xrightarrow{D} N(0,1). \tag{19}$$

In Theorem 3.1, we establish the asymptotic normality of $\widehat{\Omega}_{i,j}^{(\mathrm{db})}$. We now discuss the improvement in the convergence rates with transfer learning. For the asymptotic normality to hold, one requires the sparsity condition that

$$\begin{cases} s \log p = o(N/\sqrt{n}) \text{ and } \delta_{h,\Omega} = o(n^{-1/2}) & \text{if } \delta_{h,\Omega} \leq s \log p/n \\ s \log p = o(\sqrt{n}) & \text{otherwise.} \end{cases} \tag{20}$$

In comparison, the sparsity condition given by the minimax rate in single-task setting is $s \log p = o(\sqrt{n})$ (Proposition 3.1). We see that the sparsity condition in (20) is weaker when $\delta_{h,\Omega} \ll s \log p/n$ and $N \gg n$. Specifically, if the similarity is sufficiently high, that is, $\delta_{h,\Omega} \leq s \log p/n$, then transfer learning relaxes the sparsity condition for asymptotic normality from $s \log p \ll \sqrt{n}$ to $s \log p \ll N/\sqrt{n}$. This relaxation is significant as the regime $s \log p \ll \sqrt{n}$ is known as the "ultra-sparse" regime and is very restrictive when n is small. When $N \gg n$, which is not hard to satisfy in many applications, the condition on s is largely relaxed for valid inference. Another way of interpreting (20) is that inference based on Trans-CLIME requires weaker sparsity conditions as long as $\widehat{\Omega}$ has a smaller estimation error than the single-task CLIME. The minimax optimality of debiased Trans-CLIME is studied in Section 5.2.

Next, we introduce an estimator for the variance of $\widehat{\Omega}_{i,j}^{(\mathrm{db})}$, which is

$$\widehat{V}_{i,j} = \widehat{\Omega}_{i,i}\widehat{\Omega}_{j,j} + \widehat{\Omega}_{i,j}\widehat{\Omega}_{j,i}. \tag{21}$$

The variance estimator $\widehat{V}_{i,j}$ is based on the limiting distribution of $\widehat{\Omega}_{i,j}^{(\mathrm{db})}$ given that the observations are Gaussian distributed.

Lemma 3.1 (Consistency of variance estimator). Under the conditions of Theorem 3.1, the variance estimator defined in (21) satisfies

$$|\widehat{V}_{i,j} - V_{i,j}| = o_P(1)$$

and hence,

$$\frac{\sqrt{\tilde{n}}(\widehat{\Omega}_{i,j}^{(\mathrm{db})} - \Omega_{i,j})}{\sqrt{\widehat{V}_{i,j}}} \xrightarrow{D} N(0,1). \tag{22}$$

Based on Lemma 3.1, a $100 \times (1-\alpha)\%$ two-sided confidence interval for $\Omega_{i,j}$ is

$$\widehat{\Omega}_{i,j}^{(\mathrm{db})} \pm z_{1-\alpha/2}(\widehat{V}_{i,j}/\tilde{n})^{1/2}.$$

We examine the empirical performance of this inference method in Section 6.

4. Edge Detection with FDR Control

An important task regarding the graphical models is edge detection with uncertainty quantification. That is, we consider testing $(H_0)_{i,j} : \Omega_{i,j} = 0 \quad 1 \le i < j \le p$. This is a multiple testing problem with $m = p(p-1)/2$ hypotheses to test in total. For the uncertainty quantification, we consider the false discovery proportion (FDP) and false discovery rate (FDR). Let $\widehat{\mathcal{R}}$ denote the set of rejected null hypotheses. The FDP and FDR are defined as, respectively,

$$\mathrm{FDP}(\widehat{\mathcal{R}}) = \frac{\sum_{(i,j)\in\widehat{\mathcal{R}}} \mathbb{1}((i,j) \in \mathcal{H}_0)}{|\widehat{\mathcal{R}}| \vee 1} \quad \text{and}$$
$$\mathrm{FDR}(\widehat{\mathcal{R}}) = \mathbb{E}[\mathrm{FDP}(\widehat{\mathcal{R}})],$$

where $\mathcal{H}_0$ is the set of true nulls. Many algorithms have been proposed and studied for FDR and FDP control in various settings. Especially, Liu (2013) proposes an FDR control algorithm for GGMs that can be easily combined with our proposed debiased estimator. The procedure is presented as Algorithm 2.

Algorithm 2: Edge detection with FDR control at level α

Input : $\{\widehat{\Omega}^{(db)}_{i,j}\}_{i<j}$, $\{\widehat{V}_{i,j}\}_{i<j}$, and FDR level α

Output: A set of selected edges $\widehat{\mathcal{R}}$

Step 1. For $1 \le i < j \le p$, let $\widehat{z}_{i,j} = \widehat{V}^{-1/2}_{i,j}\sqrt{n}\widehat{\Omega}^{(db)}_{i,j}$, where $\widehat{\Omega}^{(db)}_{i,j}$ and $\widehat{V}_{i,j}$ are defined in (18) and (21), respectively.

Step 2.

$$\hat{t} = \inf\left\{t \in [0, \sqrt{2\log m - 2\log\log m}] : \frac{2m(1-\Phi(t))}{\max\{\sum_{1\le i<j\le p}\mathbb{1}(|\widehat{z}_{i,j}| \ge t), 1\}} \le \alpha\right\}. \quad (23)$$

If (23) does not exist, we set $\hat{t} = \sqrt{2\log m}$.

Step 3. The rejected hypotheses are

$$\widehat{\mathcal{R}} = \{(i,j) : |\widehat{z}_{i,j}| \ge \hat{t}, 1 \le i < j \le p\}.$$

Let $m_0 = |\mathcal{H}_0|$ denote the cardinality of $\mathcal{H}_0$ and $m = (p^2-p)/2$ denote the total number of hypotheses to test. Define a subset of random variables "highly" correlated with the ith variable $\mathcal{C}_i(\gamma) = \{j : 1 \le j \le p, j \ne i, |\Omega_{i,j}| \ge (\log p)^{-2-\gamma}\}$.

Theorem 4.1 (FDR control). Let $p \le n^r$ for some $r > 0$ and $m_0 \ge cp^2$ for some $c > 0$. Assume that Condition 2.1 holds and the true model is in $\mathbb{G}_1(s,h)$. Suppose that

$$s(\log p)^{3/2} \ll N/\sqrt{n}, \quad \delta_{h,\Omega} \wedge s\log p/n \ll (n\log p)^{-1/2},$$

and $\max_{1\le i\le p}|\mathcal{C}_i(\gamma)| = O(p^\rho)$ for some $\rho < 1/2$ and $\gamma > 0$. We have

$$\lim_{(n,p)\to\infty} \frac{\mathrm{FDR}(\widehat{\mathcal{R}})}{\alpha m_0/m} = 1 \quad \text{and} \quad \frac{\mathrm{FDP}(\widehat{\mathcal{R}})}{\alpha m_0/m} \to 1 \text{ in probability}$$

as $(n,p) \to \infty$.

Theorem 4.1 implies that Algorithm 2 can asymptotically control FDR and FDP at nominal level under certain conditions. The sample size condition in Theorem 4.1 guarantees that the remaining bias of $\widehat{\Omega}^{(\mathrm{db})}_{i,j}$ is uniformly $o_P((n\log p)^{-1/2})$. The condition on the cardinality of $\mathcal{C}_i(\gamma)$ guarantees that the z-statistics have mild correlations such that the FDR control is asymptotically valid.

5. Minimax Optimal Rates for $q \in [0,1]$

In this section, we establish the minimax lower bounds for estimation and inference of GGMs in the parameter space $\mathbb{G}_q(s,h)$ for any fixed $q \in [0,1]$. We provide matching minimax upper bounds in the supplementary materials (Sections C.3 and C.4).

5.1. Optimal Rates Under Frobenius Norm

Theorem 5.1 (Minimax lower bounds under Frobenius norm). Assume Condition 2.1 and $3 < s\log p < c_1 N$ for some small constant c_1. (i) For $q = 0$, if $h\log p \le c_2 n$ for some small enough constant c_2, then for some positive constant C_1,

$$\inf_{\widehat{\Omega}} \sup_{\mathbb{G}_0(s,h)} \mathbb{E}\left[\frac{1}{p}||\widehat{\Omega} - \Omega||_F^2\right] \ge C_1\left\{\frac{s\log p}{N+n} + (h\wedge s)\frac{\log p}{n}\right\}.$$

(ii) For any fixed $q \in (0,1]$, if $h^q(\log p/n)^{1-q/2} \le c_3$ for some small enough constant c_3, then there are some positive constants C_2 such that

$$\inf_{\widehat{\Omega}} \sup_{\mathbb{G}_q(s,h)} \mathbb{E}\left[\frac{1}{p}||\widehat{\Omega} - \Omega||_F^2\right] \ge C_2\left\{\frac{s\log p}{N+n} + h^q\Big(\frac{\log p}{n}\Big)^{1-q/2} \wedge h^2 \wedge \frac{s\log p}{n}\right\}.$$

Theorem 5.1 establishes the minimax optimal rates under Frobenius norm. These lower bounds generalize the existing lower bound in $\mathbb{G}_q(s,\infty)$ (Cai, Liu, and Zhou 2016) to allow for arbitrarily small h. According to the discussion at the end of Section 2, the Trans-CLIME is minimax optimal with respect to Frobenius norm in most parameter spaces of interest when $s^2\log p \le c_0 n$ or when $\max\{||\Omega||_{\infty,1}, h\}$ is finite. In fact, the only difference of the rate in (15) and the minimax lower bound is the cost of aggregation, which is of order $1/n$. We mention that the estimator which achieves the minimax upper bounds depends on the relative magnitude of h and s and hence, is not adaptive. In comparison, the Trans-CLIME estimator does not depends on the unknown parameters and only has a small inflation term.

5.2. Optimal Rates for Estimating $\Omega_{i,j}$

Theorem 5.2 (Minimax lower bounds for estimating $\Omega_{i,j}$). Assume Condition 2.1 and $3 < s < c_1\min\{p^\nu, N/\log p\}$ for some small constant $c_1 > 0$ and $\nu < 1/2$. (i) For $q = 0$, if $1 \le h\log p \le c_2 n$ for some small enough constant c_2, then for some constant $C_1 > 0$,

$$\inf_{\widehat{\Omega}} \sup_{\mathbb{G}_0(s,h)} \mathbb{P}\left(|\widehat{\Omega}_{i,j} - \Omega_{i,j}| \ge C_1\left\{n^{-1/2} + \frac{s\log p}{N+n} + (h\wedge s)\frac{\log p}{n}\right\}\right) > 1/4.$$

(ii) For any fixed $q \in (0,1]$, if $h^q(\log p/n)^{1-q/2} < c_3$ for some small enough constant c_3, then

$$\inf_{\widehat{\Omega}} \sup_{\mathbb{G}_q(s,h)} \mathbb{P}\left(|\widehat{\Omega}_{i,j} - \Omega_{i,j}| \geq C_3\left\{R_q + \frac{s\log p}{N+n} + h^q\left(\frac{\log p}{n}\right)^{1-q/2} \wedge h^2 \wedge \frac{s\log p}{n}\right\}\right) > 1/4,$$

where C_3 is a positive constant and $R_q = (N+n)^{-1/2} + n^{-1/2} \wedge h$.

Theorem 5.2 establishes the minimax lower bound for estimating each entry in the graph. These lower bounds are related to the sparsity conditions in Theorem 3.1. They generalize the existing lower bound in $\mathbb{G}_q(s,\infty)$ (Ren et al. 2015) to allow for finite h. When $q = 0$, the parametric rate is $n^{-1/2}$, which is the same as in the single-task setting. When $q \in (0,1]$, the parametric rate is R_q and the rest terms are the remaining bias. We now illustrate this bound in detail with $q = 1$.

For $q = 1$, the minimax optimal rate can be achieved by $\widehat{\Omega}^{(\mathrm{db})}_{i,j}$ when $\max\{h, ||\Omega||_{\infty,1}\} \leq C$ and $h \gtrsim n^{-1/2}$. See (15) in the supplementary materials for more details. When $h \ll n^{-1/2}$, a minimax optimal estimator is debiased (single-task) CLIME using X and $X^{(k)}$, $1 \leq k \leq K$, as debiasing samples. In the scenario $h \ll n^{-1/2}$, the auxiliary studies are very similar to the target study and using $N + n$ debiasing samples can have faster parametric rate, $(n+N)^{-1/2}$, with bias no larger than h. However, the central limit theory may not hold for the rate optimal estimator when $h \ll n^{-1/2}$. This is because the parametric rate is dominated by the bias h when $(n+N)^{-1/2} \lesssim h \lesssim n^{-1/2}$. In contrast, $\widehat{\Omega}^{(\mathrm{db})}_{i,j}$ has parametric rate $n^{-1/2}$ and its asymptotic normality holds for arbitrarily small h under the conditions of Theorem 3.1. Hence, $\widehat{\Omega}^{(\mathrm{db})}_{i,j}$ is a proper choice for statistical inference.

6. Numerical Experiments

We evaluate the numerical performance of our proposal and other comparable methods. We set $n = 150$, $p = 200$, $K = 5$, and $n_k = 300$ for $k = 1, \ldots, K$. We consider three types of target graph Ω.

(i) Banded matrix with bandwidth 8. For $1 \leq i,j \leq p$, $\Omega_{i,j} = 2 \times 0.6^{|i-j|}\mathbb{1}(|i-j| \leq 7)$.
(ii) Block diagonal matrix with block size 4, where each block is Toeplitz $(1.2, 0.9, 0.6, 0.3)$.
(iii) Define $\mathring{\Omega}_{i,j} = \mathbb{1}(i=j) + u_{i,j}/\sqrt{|i-j|+1}$, where $u_{i,j}$ are independently generated from a uniform distribution with range $[0, 0.8]$. We threshold $(\mathring{\Omega} + \mathring{\Omega}^\intercal)/2$ such that only the first s largest values in each column and each row of $(\mathring{\Omega} + \mathring{\Omega}^\intercal)/2$ are kept. A diagonal matrix is added to guarantee that the minimum eigenvalue of the target graph is at least 0.1.

To accommodate the practical setting that some auxiliary studies can be very far from the target study, we define a set $\mathcal{A} \subseteq \{1, \ldots, K\}$ to be the set of informative studies. Specifically, for each setting in (i) or (ii), we generate $\Delta^{(k)}$ in two ways. For $k \in \mathcal{A}$, $\Delta^{(k)}_{i,j}$ is zero with probability 0.9 and is nonzero with probability 0.1. If an entry is nonzero, it is randomly generated from $U[-r/p, r/p]$ for $r \in \{10, 20, 30\}$. For $\Omega^{(k)}$, $k \notin \mathcal{A}$, we generate $\Delta^{(k)}_{i,j} = 0.5\mathbb{1}(i=j) + \xi_{i,j}$, where $\xi_{i,j}$ is zero with probability 0.7 and is 0.4 with probability 0.3. In fact, $\mathcal{D}_1(\Omega^{(k)}, \Omega) \approx 15$ for $k \in \mathcal{A}$ and $\mathcal{D}_1(\Omega^{(k)}, \Omega) \approx 80$ for $k \notin \mathcal{A}$ and hence, the studies not in $\mathcal{A}$ are relatively far away from the target Ω. For $k = 1, \ldots, K$, we symmetrize $\Omega^{(k)}$ and if $\Omega^{(k)}$ is not positive definite, we redefine $\Omega^{(k)}$ to be its positive definite projection.

We compare four methods in each experiment. The first one is the proposed Trans-CLIME. The second one is the single-task CLIME that only uses the data from the target study. The third one applies Trans-CLIME to the target study and informative auxiliary studies, denoted by "oracle Trans-CLIME." Here "oracle" indicates that it leverages the knowledge of oracle $\mathcal{A}$. The fourth one is multi-task graphical lasso (Guo et al. 2011), shorthanded as "MT-Glasso." For the choice of tuning parameters, we consider $\lambda_{\mathrm{CL}} = 2c_n\sqrt{\log p/n}$ for CLIME. We pick c_n to minimize the prediction error defined in (24) based on 5-fold cross-validation. For the Trans-CLIME, we set $\lambda_\Delta = 2||\widehat{\Omega}^{(\mathrm{CL})}||_1\sqrt{\log p/n}$ and $\lambda_\Theta = 2c_n\sqrt{\log p/N}$ where c_n is the same as in the CLIME optimization. For the oracle Trans-CLIME, the tuning parameters are set in the same way as in Trans-CLIME except that N is replaced by $\sum_{k\in\mathcal{A}} n_k$. For Trans-CLIME-based methods, we split the target data into two folds such that $\widehat{\Omega}^{(\mathrm{CL})}$ and $\widehat{\Theta}$ are computed based on $4n/5$ samples and the aggregation step (Step 3) is based on the rest $n/5$ samples. For the debiased Trans-CLIME, we use all the target data as debiasing samples as it has a better empirical performance. For "MT-Glasso," we implement and choose the tuning parameter based on Bayesian information criterion according to Sections 2.3 and 2.4 of Guo et al. (2011). The R code for the four methods is available at *https://github.com/saili0103/TransCLIME*.

6.1. Estimation and Prediction Results

We evaluate the estimation performance in Frobenius norm and prediction performance based on the negative log-likelihood. Specifically, we generate $x^{(\mathrm{test})}_i \sim N(0, \Sigma)$ for $i = 1, \ldots, n_{\mathrm{test}} = 100$ and $x^{(\mathrm{test})}_i$ are independent of the samples for estimation. We evaluate the out-of-sample prediction error of an arbitrary graph estimator $\widehat{\Omega}^{(\mathrm{init})}$ in the following way. We symmetrize $\widehat{\Omega}^{(\mathrm{init})}$ and compute the positive definite projection of the symmetrized $\widehat{\Omega}^{(\mathrm{init})}$, denoted by $\widehat{\Omega}^{(\mathrm{init})}_+$. The prediction error of $\widehat{\Omega}^{(\mathrm{init})}$ is evaluated via

$$\widehat{Q}(\widehat{\Omega}^{(\mathrm{init})}) = \frac{1}{p}\left\{\frac{1}{2n_{\mathrm{test}}}\sum_{i=1}^{n_{\mathrm{test}}} \mathrm{Tr}(x^{(\mathrm{test})}_i (x^{(\mathrm{test})}_i)^\intercal \widehat{\Omega}^{(\mathrm{init})}_+) - \frac{1}{2}\log\det(\widehat{\Omega}^{(\mathrm{init})}_+)\right\}. \quad (24)$$

In Figure 1, we report the estimation and prediction performance in setting (i). As the number of informative auxiliary studies increases, the estimation errors of two Trans-CLIME-based methods decrease. As r increases, the estimation errors of two Trans-CLIME-based methods increase. The oracle Trans-CLIME has a faster convergence rate than the Trans-CLIME. This is because $K - |\mathcal{A}|$ noninformative studies are used in the

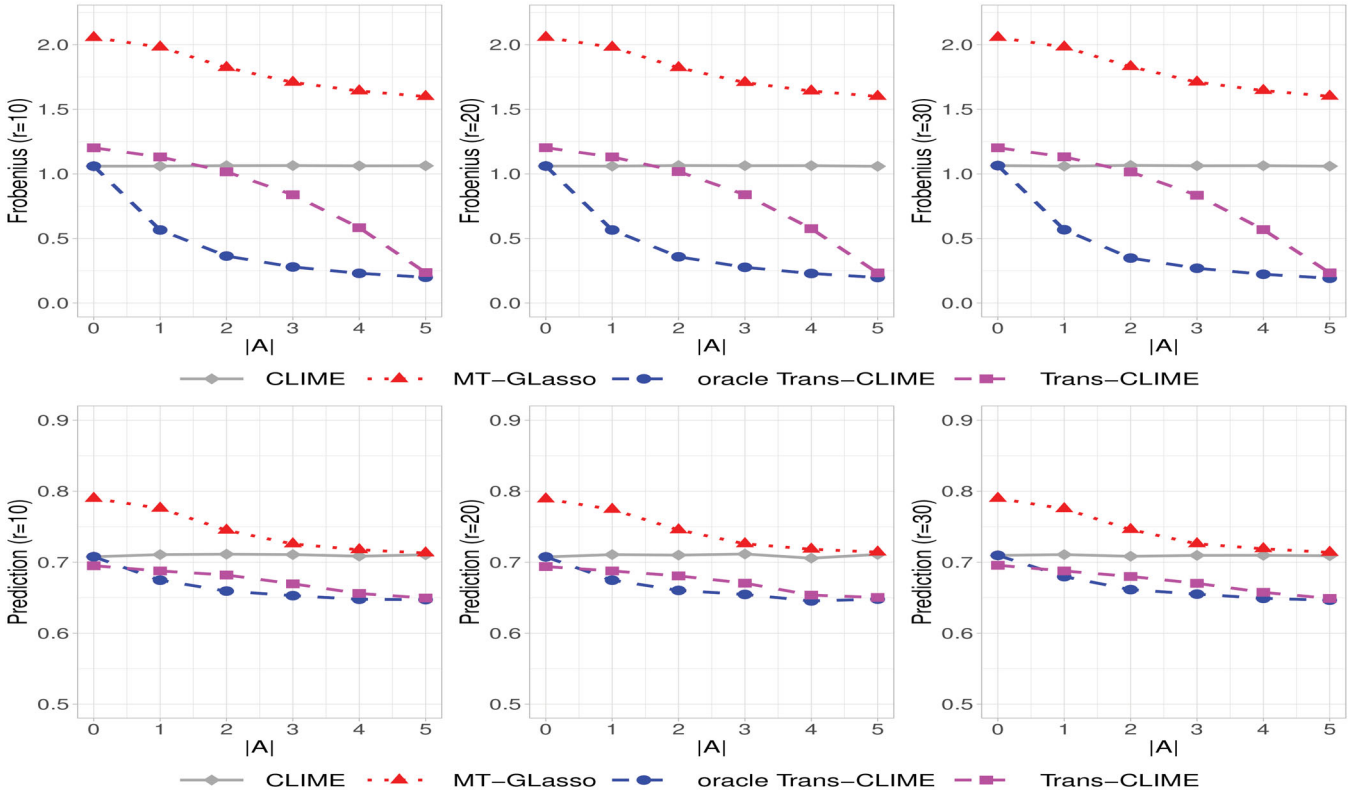

Figure 1. Estimation errors in Frobenius norm (first row) and negative log-likelihood (second row) for banded Ω (i) as a function of the number of informative studies (out of a total of $K = 5$ studies) for different values of r.

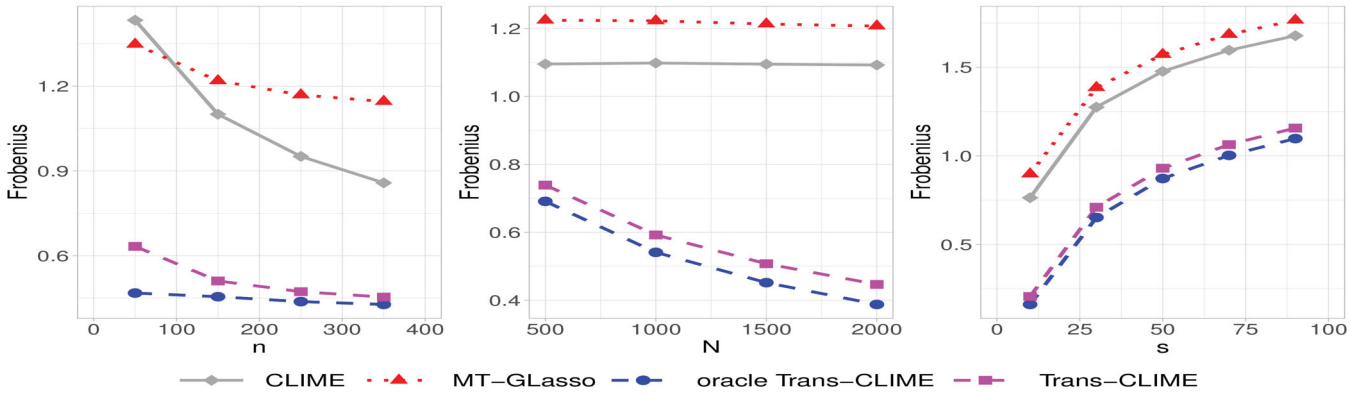

Figure 2. Estimation errors in Frobenius norm for random Ω (iii) with varying target sample size n (left), total auxiliary sample size N (middle), and the sparsity s (right). The default setting is $p = 200, n = 150, N = 1500, s = 20, r = 20$, and $\mathcal{A} = \{1, \ldots, K\}$. In each plot, all the parameters are fixed at default values except the parameter indexed by the x-axis.

Trans-CLIME, which affects the convergence rates. Nevertheless, we see that the Trans-CLIME algorithm is robust to the noninformative auxiliary studies as its performance is always not much worse than the single-task CLIME. The estimation and prediction results for settings (ii) and (iii) are reported in the supplementary materials (Section E.2). We have observed similar patterns in those settings.

6.2. Sample Complexity

We further evaluate the dependence of the estimation errors in Frobenius norm on the target sample size n, the total auxiliary sample size N, and the sparsity s. We consider the random graph (iii) that has more flexibility in varying s and present the results in Figure 2. We see that the estimation errors of single-task CLIME decrease fast as the target sample size increases. The errors of oracle Trans-CLIME, and Trans-CLIME also decrease as the target sample size increases but not in a large amount. In other words, the improvement of transfer learning becomes smaller as n increases when auxiliary samples are informative. This aligns with our theoretical results as we demonstrate the significant gain of transfer learning when $N \gg n$. As the total auxiliary sample size N increases, errors of the oracle Trans-CLIME and Trans-CLIME decrease linearly but the errors of CLIME do not change (middle plot of Figure 2). This agrees with our theoretical findings in Theorem 2.1 and Remark 2.1. Finally,

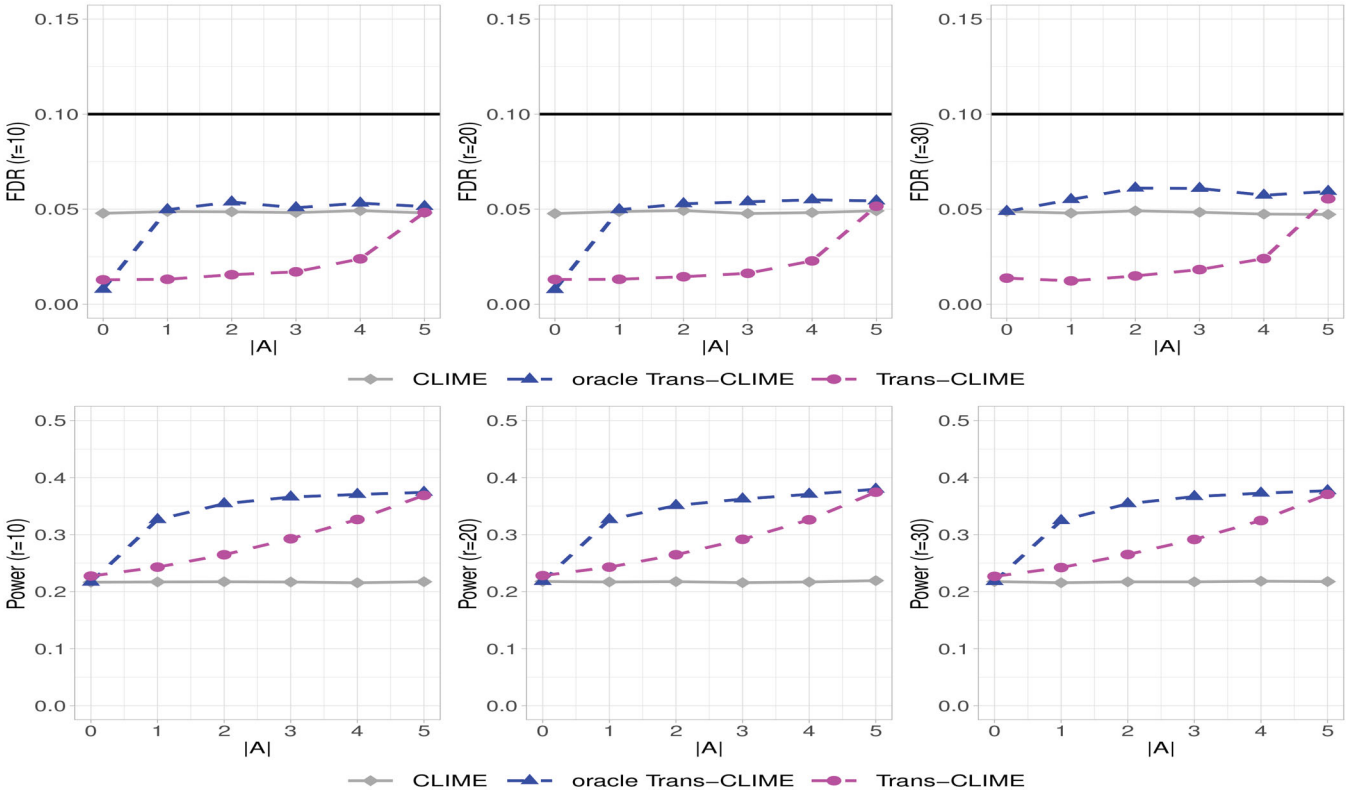

Figure 3. The FDR and power at nominal level 0.1 as a function of the number of informative studies (out of $K = 5$) and r for banded Ω (i). The methods in comparison are debiased as in Section 3.

all the methods have estimation errors increase with the sparsity parameter s with an approximately linear trend.

6.3. FDR Control

We then consider FDR control at level $\alpha = 0.1$ based on the debiasing method introduced in Section 3. From Figure 3, we see that all three debiased estimators have empirical FDR no larger than the nominal level in setting (i). In terms of power, the Trans-CLIME and oracle Trans-CLIME have higher power when $\mathcal{A}$ is nonempty. We observe the robustness of Trans-CLIME in the sense that the FDR is under control even if non-informative studies are included. The multiple testing results for settings (ii) and (iii) are reported in the supplementary materials (Section D.2). We observe that the power curves in setting (ii) are the highest in comparison to the other two settings. This is because the nonzero entries in the target graph have the largest magnitude in setting (ii).

7. Gene Network Estimation in Multiple Tissues

In this section, we apply our proposed algorithms to detect gene networks in different tissues using the Genotype-Tissue Expression (GTEx) data (*https://gtexportal.org/*). Overall, the datasets measure gene expression levels in 49 tissues from 838 human donors, comprising a total of 1,207,976 observations of 38,187 genes. We focus on genes related to central nervous system neuron differentiation, annotated as `GO:0021953`. This gene set includes a total of 184 genes. A complete list of the genes can be found at *https://www.gsea-msigdb.org/gsea/msigdb/cards/GO_CENTRAL_NERVOUS_SYSTEM_NEURON_DIFFERENTIATION*.

Our goal is to estimate and detect the gene network in a target brain tissue. Since we use 20% of the samples to compute test errors, the sample size for the target tissue should not be too small. We therefore, consider each brain tissue with at least 100 samples as the target in each experiment. We use the data from multiple other brain tissues as auxiliary samples with $K = 12$. We remove the genes that have missing values in these 13 tissues, resulting in a total of 141 genes for the graph construction. The average sample size in each tissue is 115. A complete list of tissues and their sample sizes are given in the supplementary materials.

We apply CLIME and Trans-CLIME to estimate the graph among these 141 genes in multiple target brain tissues. We first compare the prediction performance of CLIME and Trans-CLIME, where we randomly split the samples of the target tissue into five folds. We fit the model with four folds of the samples and compute the prediction error with the rest of the samples. We report the mean of the prediction errors, each based on a different fold of the samples. The prediction errors are measured by the negative log-likelihood defined in (24).

The prediction results are reported in the left panel of Figure 4. We see that the prediction errors based on Trans-CLIME are significantly lower than those based on CLIME in many cases, indicating that the brain tissues in GTEx possess relatively high similarities in gene associations. On the other hand, these brain tissues are also heterogeneous in the sense that the improvements with transfer learning are significant in some tissues (e.g., A.C. cortex and F. cortex) and they are relatively mild

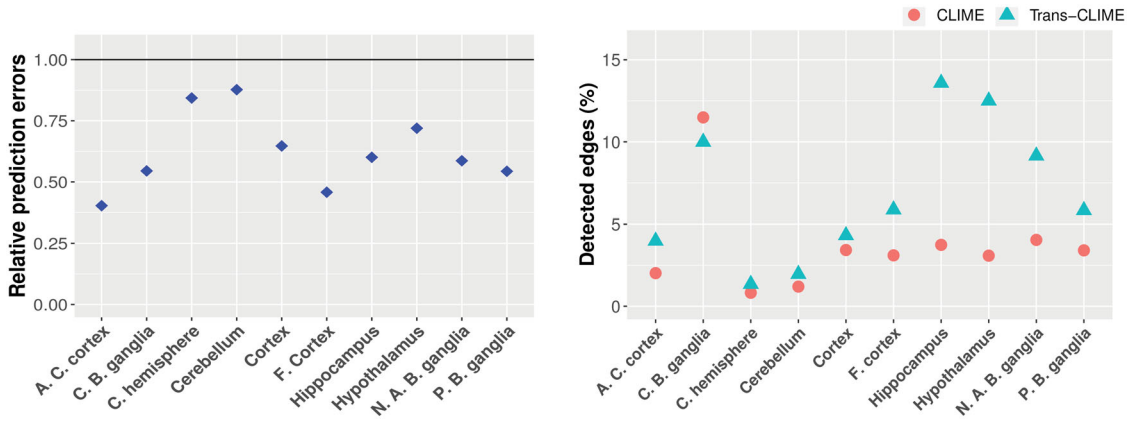

Figure 4. The left panel presents the prediction errors of Trans-CLIME relative to the prediction errors of CLIME for 10 different target tissues. The right panel presents the number of detected edges divided by $p(1-p)$ using CLIME and Trans-CLIME with FDR=0.1. The full names of the target tissues are given in the supplementary materials.

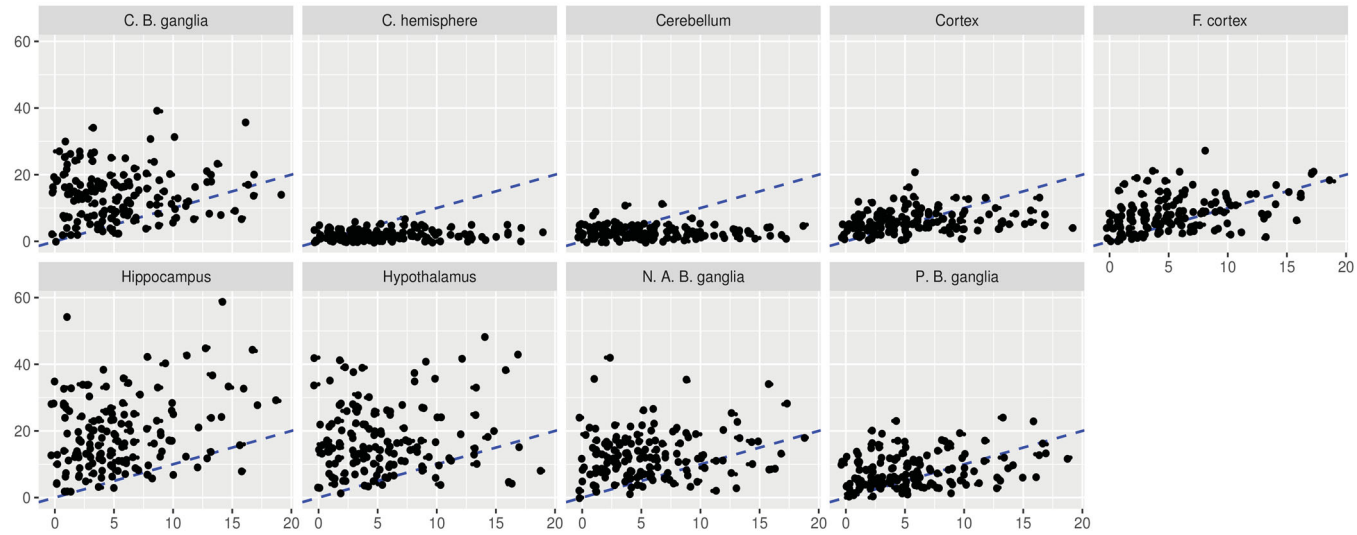

Figure 5. Comparison of the node degree distribution based on the graph estimated by Trans-CLIME for each of the tissue at FDR level of 10%. The x-axis represents the degrees of the nodes in A. C. cortex and the y-axis represents the degrees of the nodes in nine other tissues. The dashed line is diagonal.

in others (e.g., C. hemisphere and Cerebellum). We then apply Algorithm 2 with $\alpha = 0.1$ to identify the connections among these genes. The proportion of detected edges are reported in the right panel of Figure 4. We see that the proportion of detected edges are relatively low, implying that the networks are sparse. The Trans-CLIME has more discoveries than CLIME in almost all the tissues in detecting the gene-gene links, agreeing with our simulation results. In Figure 5, we evaluate the similarities among the tissues based on the constructed graphs. Specifically, we examine the degrees of nodes in A.C. cortex in comparison to the degrees of nodes in the other nine tissues, all estimated using Trans-CLIME. We see that the degree distribution in A.C. cortex is relatively similar to the degree distributions in Cortex and F. cortex.

In the supplementary materials (Section E.1), we report the hubs detected by these two methods in different tissues and observe that many hubs appear more than once in different tissues based on the results of debiased Trans-CLIME, further demonstrating a certain level of similarity in gene regulatory networks among different brain tissues. For example, for A.C Cortex with Trans-CLIME, we are able to identify the hub genes SOX1, SHANK3, ATF5, and SEMA3A. These genes are either the known transcriptional factors (SOX1, ATF5) and have been shown to be related to neurological diseases, including the leading autism gene SHANK3 (Lutz et al. 2020) and gene-related to motor neurons in ALS patients (Sema3A) (Birger et al. 2018). In comparison, the graphs estimated using CLIME in single tissue are too sparse and do not reveal any of these hub genes.

8. Discussion

In this article, we have studied the estimation and inference of Gaussian graphical models with transfer learning. Our proposed algorithm Trans-CLIME admits a faster convergence rate than the minimax rate in the single-task setting under mild conditions. The Trans-CLIME estimator can be further debiased for statistical inference.

We have seen in the numerical experiments that including some noninformative auxiliary studies can weaken the improvement of transfer learning. While our proposal is guaranteed to be no worse than the single-task minimax estimator, it may not be the most efficient way to use the auxiliary studies. A practical challenge in transfer learning is to find the best set of auxiliary studies such that the algorithm can gain the most

from the auxiliary tasks. In the high-dimensional regression problem, Li, Cai, and Li (2021) proposes to first rank all the auxiliary studies according to their similarities to the target and then perform a model selection aggregation. They prove that the aggregated estimator can be adapted to the informative set under certain conditions. In a more recent article, Hanneke and Kpotufe (2020) proves that, loosely speaking, if the ranks of the auxiliary studies can be recovered, then performing empirical risk minimization in a cross-fitting manner can achieve adaptation to the informative set to some extent in some functional classes. For the high-dimensional GMMs, heuristic rank estimators can also be derived using their connections to linear models, based on which one can perform aggregation toward an adaptive estimator. However, theoretical analysis for such rank estimators may require strong conditions, especially in the high-dimensional setting. Hence, finding the best subset of auxiliary studies is an important topic for future research.

Supplementary Materials

Supplement to "Transfer Learning in Large-Scale Gaussian Graphical Models with False Discovery Rate Control." In the supplementary materials, we provide the proofs of theorems and further results on simulations and data applications.

Funding

This research was supported by NIH grants R01GM123056 and R01GM129781.

References

Agarwal, A., Negahban, S., and Wainwright, M. J. (2012), "Noisy Matrix Decomposition via Convex Relaxation: Optimal Rates in High Dimensions," *The Annals of Statistics*, 40, 1171–1197. [1]

Bastani, H. (2021), "Predicting with Proxies: Transfer Learning in High Dimension," *Management Science*, 67, 2964–2984. [2]

Birger, A., Ottolenghi, M., Perez, L., Reubinoff, B., and Behar, O. (2018), "ALS-Related Human Cortical and Motor Neurons Survival is Differentially Affected by Sema3A," *Cell Death & Disease*, 9, 256. [11]

Cai, T. T., and Guo, Z. (2020), "Semi-Supervised Inference for Explained Variance in High-Dimensional Linear Regression and its Applications," *Journal of the Royal Statistical Society*, Series B, 82, 391–419. [5]

Cai, T. T., Li, H., Liu, W., and Xie, J. (2016), "Joint Estimation of Multiple High-Dimensional Precision Matrices," *Statistica Sinica*, 26, 445–464. [1,6]

Cai, T. T., Liu, W., and Luo, X. (2011), "A Constrained l1 Minimization Approach to Sparse Precision Matrix Estimation," *Journal of the American Statistical Association*, 106, 594–607. [1,2,3,5]

Cai, T. T., Liu, W., and Zhou, H. H. (2016), "Estimating Sparse Precision Matrix: Optimal Rates of Convergence and Adaptive Estimation," *The Annals of Statistics*, 44, 455–488. [1,7]

Cai, T. T., Ren, Z., and Zhou, H. H. (2016), "Estimating Structured High-Dimensional Covariance and Precision Matrices: Optimal Rates and Adaptive Estimation," *Electronic Journal of Statistics*, 10, 1–59. [1,3,5]

Cai, T. T., and Wei, H. (2021), "Transfer Learning for Nonparametric Classification: Minimax Rate and Adaptive Classifier," *The Annals of Statistics*, 49, 100–128. [2]

Chiquet, J., Grandvalet, Y., and Ambroise, C. (2011), "Inferring Multiple Graphical Structures," *Statistics and Computing*, 21, 537–553. [1]

Danaher, P., Wang, P., and Witten, D. M. (2014), "The Joint Graphical Lasso for Inverse Covariance Estimation Across Multiple Classes," *Journal of the Royal Statistical Society*, Series B, 76, 373–397. [1,5]

Drton, M., and Maathuis, M. H. (2017), "Structure Learning in Graphical Modeling," *Annual Review of Statistics and Its Application*, 4, 365–393. [1]

Fagny, M., Paulson, J. N., Kuijjer, M. L., Sonawane, A. R., Chen, C.-Y., Lopes-Ramos, C. M., Glass, K., Quackenbush, J., and Platig, J. (2017), "Exploring Regulation in Tissues with eQTL Networks," *Proceedings of the National Academy of Sciences of the United States of America*, 114, E7841–E7850. [1]

Friedman, J., Hastie, T., and Tibshirani, R. (2008), "Sparse Inverse Covariance Estimation with the Graphical Lasso," *Biostatistics*, 9, 432–441. [1,5]

Glymour, C., Zhang, K., and Spirtes, P. (2019), "Review of Causal Discovery Methods Based on Graphical Models," *Frontiers in Genetics*, 10, 524. [1]

Guo, J., Levina, E., Michailidis, G., Zhu, J. (2011), "Joint Estimation of Multiple Graphical Models," *Biometrika*, 98, 1–15. [1,5,8]

Hanneke, S., and Kpotufe, S. (2020), "A No-Free-Lunch Theorem for Multitask Learning," arXiv:2006.15785. [2,3,12]

Lam, C., and Fan, J. (2009), "Sparsistency and Rates of Convergence in Large Covariance Matrix Estimation," *Annals of Statistics*, 37, 4254–4278. [1]

Lecué, G., and Rigollet, P. (2014), "Optimal Learning with q-Aggregation," *The Annals of Statistics*, 42, 211–224. [3]

Li, S., Cai, T. T., and Li, H. (2021), "Transfer Learning for High-Dimensional Linear Regression: Prediction, Estimation, and Minimax Optimality," *Journal of the Royal Statistical Society*, Series B, (to appear). [2,12]

Liu, H., and Wang, L. (2017), "Tiger: A Tuning-Insensitive Approach for Optimally Estimating Gaussian Graphical Models," *Electronic Journal of Statistics*, 11, 241–294. [1,5]

Liu, W. (2013), "Gaussian Graphical Model Estimation with False Discovery Rate Control," *The Annals of Statistics*, 41, 2948–2978. [1,6,7]

Lounici, K., Pontil, M., Tsybakov, A., and Van De Geer, S. (2009), "Taking Advantage of Sparsity in Multi-Task Learning," in *COLT 2009-The 22nd Conference on Learning Theory*. [1]

Lutz, A.-K., Pfaender, S., Incearap, B., Ioannidis, V., Ottonelli, I., Föhr, K. J., Cammerer, J., Zoller, M., Higelin, J., Giona, F., Stetter, M., Stoecker, N., Alami, N. O., Schön, M., Orth, M., Liebau, S., Barbi, G., Grabrucker, A. M., Delorme, R., Fauler, M., Mayer, B., Jesse, S., Roselli, F., Ludolph, A. C., Bourgeron, T., Verpelli, C., Demestre, M., and Boeckers, T. M. (2020), "Autism-Associated SHANK3 Mutations Impair Maturation of Neuromuscular Junctions and Striated Muscles," *Science Translational Medicine*, 12, eaaz3267. [11]

Pan, S. J., and Yang, Q. (2009), "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering*, 22, 1345–1359. [2]

Pierson, E., Koller, D., Battle, A., and Mostafavi, S. (2015), "Sharing and Specificity of Co-expression Networks Across 35 Human Tissues," *PLOS Computational Biology*, 11, e1004220. [1]

Ravikumar, P., Raskutti, G., Wainwright, M. J., and Yu, B. (2008), "Model Selection in Gaussian Graphical Models: High-Dimensional Consistency of l_1-Regularized MLE," in *Advances in Neural Information Processing Systems*, pp. 1329–1336. [1]

Ravikumar, P., Wainwright, M. J., Raskutti, G., Yu, B. (2011), "High-Dimensional Covariance Estimation by Minimizing l_1-Penalized Log-Determinant Divergence," *Electronic Journal of Statistics*, 5, 935–980. [1]

Ren, Z., Sun, T., Zhang, C.-H., and Zhou, H. H. (2015), "Asymptotic Normality and Optimalities in Estimation of Large Gaussian Graphical Models," *The Annals of Statistics*, 43, 991–1026. [1,4,6,8]

Rigollet, P., and Tsybakov, A. (2011), "Exponential Screening and Optimal Rates of Sparse Estimation," *The Annals of Statistics*, 39, 731–771. [3]

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008), "Sparse Permutation Invariant Covariance Estimation," *Electronic Journal of Statistics*, 2, 494–515. [1]

Tripuraneni, N., Jin, C., and Jordan, M. (2021), "Provable Meta-Learning of Linear Representations," in *International Conference on Machine Learning*, pp. 10434–10443. PMLR. [2]

Tripuraneni, N., Jordan, M., and Jin, C. (2020), "On the Theory of Transfer Learning: The Importance of Task Diversity," *Advances in Neural Information Processing Systems*, 33, 7852–7862. [2]

Tsybakov, A. B. (2014), "Aggregation and Minimax Optimality in High-Dimensional Estimation," in *Proceedings of the International Congress of Mathematicians* (vol. 3), pp. 225–246. [3]

Turki, T., Wei, Z., and Wang, J. T. (2017), "Transfer Learning Approaches to Improve Drug Sensitivity Prediction in Multiple Myeloma Patients," *IEEE Access*, 5, 7381–7393. [2]

Varoquaux, G., Gramfort, A., Poline, J.-B. and Thirion, B. (2010), "Brain Covariance Selection: Better Individual Functional Connectivity Models Using Population Prior," in *Advances in Neural Information Processing Systems*, pp. 2334–2342. [1]

Xia, Y., Cai, T., and Cai, T. T. (2015), "Testing Differential Networks with Applications to Detecting Gene-by-Gene Interactions," *Biometrika*, 102, 247–266. [1]

Yuan, M., and Lin, Y. (2007), "Model Selection and Estimation in the Gaussian Graphical Model," *Biometrika*, 94, 19–35. [1]

Zhao, S. D., Cai, T. T., and Li, H. (2014), "Direct Estimation of Differential Networks," *Biometrika*, 101, 253–268. [1]