Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://amstat.tandfonline.com/loi/uasa20

Tuning-Free Heterogeneous Inference in Massive Networks

Zhao Ren, Yongjian Kang, Yingying Fan & Jinchi Lv

To cite this article: Zhao Ren, Yongjian Kang, Yingying Fan & Jinchi Lv (2019) Tuning-Free Heterogeneous Inference in Massive Networks, Journal of the American Statistical Association, 114:528, 1908-1925, DOI: 10.1080/01621459.2018.1537920

To link to this article: https://doi.org/10.1080/01621459.2018.1537920

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2019, VOL. 114, NO. 528, 1908–1925: Theory and Methods
https://doi.org/10.1080/01621459.2018.1537920

Tuning-Free Heterogeneous Inference in Massive Networks

Zhao Ren[a], Yongjian Kang[b], Yingying Fan[b], and Jinchi Lv[b]

[a]Department of Statistics, University of Pittsburgh, Pittsburgh, PA; [b]Data Sciences and Operations Department, Marshall School of Business, University of Southern California, Los Angeles, CA

ABSTRACT

Heterogeneity is often natural in many contemporary applications involving massive data. While posing new challenges to effective learning, it can play a crucial role in powering meaningful scientific discoveries through the integration of information among subpopulations of interest. In this article, we exploit multiple networks with Gaussian graphs to encode the connectivity patterns of a large number of features on the subpopulations. To uncover the underlying sparsity structures across subpopulations, we suggest a framework of large-scale tuning-free heterogeneous inference, where the number of networks is allowed to diverge. In particular, two new tests, the chi-based and the linear functional-based tests, are introduced and their asymptotic null distributions are established. Under mild regularity conditions, we establish that both tests are optimal in achieving the testable region boundary and the sample size requirement for the latter test is minimal. Both theoretical guarantees and the tuning-free property stem from efficient multiple-network estimation by our newly suggested heterogeneous group square-root Lasso for high-dimensional multi-response regression with heterogeneous noises. To solve this convex program, we further introduce a scalable algorithm that enjoys provable convergence to the global optimum. Both computational and theoretical advantages are elucidated through simulation and real data examples. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received July 2017
Revised October 2018

KEYWORDS
Efficiency; Heterogeneous group square-root Lasso; Heterogeneous learning; High dimensionality; Large-scale inference; Multiple networks; Scalability; Sparsity.

1. Introduction

In the era of data deluge one can easily collect a massive amount of data from multiple sources, each of which may come from a certain subpopulation of a larger population of interest. For example, these subpopulations can be different studies on the same subjects or represent different cancer types, brain disorders, or product choices. A large number of features are often associated with each subject. Allowing and understanding the heterogeneity in the association structures of these features across subpopulations can be important in empowering meaningful scientific discoveries or effective personalized choices in our lives. Meanwhile allowing heterogeneity in the data also poses new challenges to effective learning and calls for new developments of methods, theory, and algorithms with scalability and statistical efficiency.

Heterogeneity can take different forms in various applications such as the differences among link strengths or the sparsity patterns over multiple networks, and the differences in noise levels or distributions over multiple subpopulations. To avoid potential ambiguity, we would like to make it explicit that throughout this article, we allow two particular types of heterogeneity which are the heterogeneity in link strengths and the heterogeneity in noise levels. The former means the connectivity strengths between nodes can change across subpopulations, and the latter means the variability of nodes can change across subpopulations. The latter also has to do with a very important property of our underlying estimation procedure and will be made clear in later sections. To approach the problem of heterogeneous learning in these contexts, we exploit the model setting of multiple networks with Gaussian graphs each of which encodes the connectivity pattern among features for each subpopulation. The edges of these networks are characterized by the inverse covariances for each pair of nodes from a subpopulation. The focus on this particular type of network models enables us to present our main idea with technical brevity. See, for example, Teng (2016) for an account of more general network models beyond graphical models. In fact, as a popular choice of network models Gaussian graphical models involving the inverse covariances have been used widely in applications to characterize the conditional dependency structure among features. In such models, the joint distribution of p features $X_1, \ldots, X_p$ is modeled by a multivariate Gaussian distribution $N(0, \Omega^{-1})$, where the $p \times p$ matrix Ω is called the precision matrix or inverse covariance matrix of these p features. A basic fact is that each pair of features, X_a and X_b, are conditionally independent given all other features if and only if the (a, b)th entry of the precision matrix Ω is zero. The conditional dependency structure in a Gaussian graph is, therefore, determined completely by the associated precision matrix Ω. See, for instance, Lauritzen (1996) and Wainwright and Jordan (2008) for more detailed accounts and applications of these models.

There is a growing literature on Gaussian graphical models. Much recent attention has been given to the problem of

CONTACT Zhao Ren zren@pitt.edu Department of Statistics, University of Pittsburgh, Pittsburgh, PA 15260.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

support recovery and link strength estimation, which focuses on identifying the nonzero entries of the precision matrix and estimating their strengths. Among those endeavors, a majority of the work has focused primarily on the case of a single Gaussian graphical model; see, for example, Meinshausen and Bühlmann (2006), Yuan and Lin (2007), Friedman, Hastie, and Tibshirani (2008), Fan, Feng, and Wu (2009), Yuan (2010), Cai, Liu, and Luo (2011), Ravikumar et al. (2011), Liu (2013), Zhang and Zou (2014), Ren et al. (2015), Fan and Lv (2016), and among many others. A common assumption in this line of work is that the data is assumed to be homogeneous with all observations coming from a single population. More detailed discussions and comparisons of these methods can be found in, for instance, Ren et al. (2015) and Fan and Lv (2016). Yet as mentioned before heterogeneity in the data can be prevalent in many contemporary applications involving massive data. The existing methods for analyzing data from each individual source become insufficient due to the assumption of homogeneity. Naively combining the results from these individual analyses may also yield suboptimal performance of statistical estimation and inference.

The setting of multiple networks with Gaussian graphical models has gained more recent attention. A lot of work assumes a time-varying graphical structure across different graphs. In particular, one assumes that there is a natural ordering of the graphs and the parameters of interest vary smoothly according to this order. For these developments, some smoothing techniques such as the kernel smoothing are key to the construction of the estimators as well as the analysis of their theoretical properties. While the time-varying graphical model is not the focus of our current article, one may refer to, for example, Zhou, Lafferty, and Wasserman (2010), Kolar et al. (2010), Chen, Xu, and Wu (2013), Qiu et al. (2016), and Lu, Kolar, and Liu (2015) for more details on this line of work.

In contrast, our setting of multiple networks with Gaussian graphical models is along another line that makes no assumption on the ordering of the graphs. In this line of work, the main assumption is a common sparsity structure across different graphs. In particular, the estimators proposed in Guo et al. (2011), Danaher, Wang, and Witten (2014), and Zhu, Shen, and Pan (2014) employ the approach of penalized likelihood with different choices of the penalty function, while the MPE method introduced in Cai et al. (2016) takes a weighted constrained ℓ_∞ and ℓ_1 minimization approach, which can be seen as an extension of the CLIME estimator for a single graph (Cai, Liu, and Luo 2011). The focus of such existing work is placed on the problem of support recovery and link strength estimation. Moreover, by the nature of these methods their computational cost increases drastically with both the dimensionality and the number of graphs, which can limit their practical use in analyzing massive datasets. How to develop a scalable procedure for large-scale inference in the setting of multiple Gaussian graphical models still remains largely open.

To uncover the underlying connectivity patterns among features across subpopulations and address the aforementioned challenges, in this article, we suggest a new testing framework of large-scale tuning-free heterogeneous inference (THI), where the number of networks is allowed to diverge and the number of features can grow exponentially with the number of observations. Distinct from the existing methods, our procedure identifies the sparsity patterns among a diverging number of graphs by testing the following null hypothesis

$$H_{0,ab}: \omega^0_{a,b} = (\omega^{(1)}_{a,b}, \ldots, \omega^{(k)}_{a,b})' = \mathbf{0} \tag{1}$$

associated with the joint link strength vector for each pair of features $1 \le a, b \le p$ with $a \ne b$, where $\Omega^{(t)} = (\omega^{(t)}_{a,b})$ with $1 \le t \le k$ denotes the precision matrix associated with the tth graph. To approach the statistical inference problem in (1), we propose two new tests, named the chi-based test and the linear functional-based test, for two different scenarios. The former test which is for the general scenario requires no extra information from the graphs and is shown to perform well as long as the ℓ_2 norm of the joint link strength vector $\omega^0_{a,b}$ is large. The chi-based test is named after the property that the null distribution of this test statistic is shown to converge to the chi-distribution. The latter one relies on some extra information on the signs of $\omega^{(t)}_{a,b}$. Such extra information is indeed available in some applications. For example, in some genome-wide association studies (GWAS) it was discovered that the association structures can be portable between certain subpopulations (Marigorta and Navarro 2013). In such a scenario, the linear functional-based test can be constructed and shown to perform well when the ℓ_1 norm of the vector $\omega^0_{a,b}$ becomes large.

An interesting property of both tests is that each of them is established under mild regularity conditions to be optimal in the sense of achieving the testable region boundary, where the testable region boundary is defined as the smallest signal strength below which no test is able to detect if the observations are from the null hypothesis against the alternative hypothesis and above which some test can distinguish successfully between the two hypotheses. We further show that for the linear functional-based test, the sample size requirement is in fact minimal. A natural question is whether naively combining the tests constructed from k individual graphs might suffice. Our theoretical results provide insights into this question and demonstrate the advantages of our tests in terms of weaker sample size requirement than the naive combination approach. We also would like to mention that although the main focus of our article is on hypothesis testing, our procedure can be modified easily by introducing an additional thresholding step for support recovery; see Section 2.5 for detailed discussions and comparisons with existing approaches. In particular, our modified procedure achieves support recovery under milder sample size assumption than many existing methods. To the best of our knowledge, the testing of multiple networks with graphs and the optimality study are both new to the literature.

The challenges of heterogeneous learning in the setting of multiple networks are rooted in the statistical inference with efficiency, the scalability, and the selection of tuning parameters which is often an implicit bottleneck of existing methods. Our THI framework addresses all these challenges in a harmonious fashion. Both theoretical guarantees and the tuning-free property are enabled through efficient multiple-network estimation by our newly suggested approach of heterogeneous group square-root Lasso (HGSL) in the setting of high-dimensional multi-response regression with heterogeneous noises. More specifically, we reduce the problem of estimating k graphs simultaneously to that of running p multi-response regressions

with heterogeneous noises. This new formulation allows us to borrow information across graphs when estimating their structures, which results in improved rates of convergence. To solve the convex programs from these multi-response regressions, we introduce a new tuning-free algorithm, that is, scalable and admits provable convergence to the global optimum. Compared to existing methods in the literature, our new procedure enjoys four main advantages. First, it is justified theoretically that our HGSL estimators have faster rates of convergence. Second, the HGSL method is capable of handling heterogeneous noises, the presence of which causes intrinsic difficulty for developing a tuning-free procedure. Third, our new algorithm is simple and truly tuning free, and scales up easily. Fourth, we provide theoretical justification on the convergence of the tuning-free algorithm. An R package HGSL implementing our suggested THI framework is available on CRAN (*https://cran.r-project.org/web/packages/HGSL/index.html*).

The rest of the article is organized as follows. Section 2 introduces the THI framework with the chi-based test and the linear functional-based test in multiple networks, and establishes their optimality properties. We present the HGSL approach for fitting high-dimensional multi-response regression with heterogeneous noises, and provide the estimation and prediction bounds for the estimator in Section 3. The newly proposed tuning-free algorithm for HGSL as well as a convergence analysis for the algorithm is relegated to the supplementary material. Section 4 details several numerical examples of simulation studies and real data analysis. We discuss some extensions of the suggested method to a few settings, including the setting with sub-Gaussian features in Section 5. The proofs of all the results and technical details are provided in the supplementary material.

2. Tuning-Free Heterogeneous Inference in Multiple Networks

2.1. Model Setting

As mentioned in Section 1, we adopt the setting of multiple networks with Gaussian graphical models to encode the connectivity patterns among p features $X_1, \ldots, X_p$ measured on k subpopulations of a general population, which yields k datasets. In this model, for each class $1 \le t \le k$ the p features jointly follow a multivariate Gaussian distribution

$$X^{(t)} = (X_1^{(t)}, \ldots, X_p^{(t)})' \sim N(0, (\Omega^{(t)})^{-1}), \tag{2}$$

where the superscript (t) means that these p features are measured on the tth subpopulation and $\Omega^{(t)}$ is the $p \times p$ precision matrix of the tth class. In addition, the distributions of $X^{(1)}, \ldots, X^{(k)}$ are assumed to be independent. Each of the k precision matrices $\Omega^{(t)} = (\omega_{a,b}^{(t)})_{p\times p}$ reflects the conditional dependency structure among the p features $X_1^{(t)}, \ldots, X_p^{(t)}$. In the high-dimensional setting, where the dimensionality p can be very large compared to the sample size, it is common in many applications such as genomic studies to assume that each precision matrix $\Omega^{(t)}$ has certain sparsity structure. The goals in these studies include the estimation of precision matrices $\Omega^{(t)}$ and the statistical inference, such as p-value or confidence interval on their entries $\omega_{a,b}^{(t)}$.

When there is only one class of data, that is, $k = 1$, our setting coincides with that of single Gaussian graphical model. For the general case of multiple graphs with $k \ge 2$, it can be beneficial to borrow the strength across all k classes of data to achieve more accurate estimation of the k precision matrices if the k classes are related to each other. With this spirit, we assume that the k classes share some similar sparsity structure, and the heterogeneity captures the differences among these graphs and variations of connectivities between nodes as (a, b) varies. In particular, we are interested in the scenario where for each pair of nodes (a, b) with $1 \le a \ne b \le p$, either $\omega_{a,b}^{(t)} = 0$ simultaneously for all $1 \le t \le k$ or alternatively the joint link strength vector $\omega_{a,b}^0 = (\omega_{a,b}^{(1)}, \ldots, \omega_{a,b}^{(k)})'$ is significantly different from the zero vector. Throughout the article, we denote by

$$\mathcal{E} = \left\{(a, b) : 1 \le a \ne b \le p \text{ and } \omega_{a,b}^0 \ne \mathbf{0}\right\} \tag{3}$$

the edge set corresponding to the k graphs given in model (2).

The main goal of our article is to develop an effective and efficient procedure for testing the null hypothesis $H_{0,ab}$ defined in (1) for multiple networks, which provides an inferential approach to uncovering the feature association structures across the k subpopulations. Depending on the type of the alternative hypothesis, we will introduce two different fully data-driven test statistics and establish their advantages over those obtained by naively combining the tests constructed from each individual graph.

2.2. Chi-Based Test

We begin with introducing the first test for our THI framework in multiple networks. To ease the presentation, we introduce some compact notation. Denote by a_{-j} the subvector of a vector $a = (a_1, \ldots, a_p)'$ with the jth component removed, and for any matrix $A = (a_{i,j})$ denote by $A_{*,j}$ its jth column, $A_{-j,j}$ the subvector of $A_{*,j}$ with the jth component removed, and $A_{*,-j}$ the submatrix of A with the jth column removed. Our testing idea is based on a simple observation that for each $1 \le j \le p$, the conditional distribution of $X_j^{(t)}$ given all remaining features $X_{-j}^{(t)}$ in class t follows the Gaussian distribution

$$X_j^{(t)} | X_{-j}^{(t)} \sim N(X_{-j}^{(t)\prime} C_j^{(t)}, 1/\omega_{j,j}^{(t)}) \tag{4}$$

with the $(p-1)$-dimensional coefficient vector $C_j^{(t)} = -\Omega_{-j,j}^{(t)}/\omega_{j,j}^{(t)}$. Based on the distributional representation in (4), one can see that the error random variables $\epsilon_j^{(t)} = X_j^{(t)} - X_{-j}^{(t)\prime} C_j^{(t)}$ are independent across t and follow the distribution $N(0, 1/\omega_{j,j}^{(t)})$. Moreover, it holds for each pair of nodes (a, b) with $1 \le a, b \le p$ that

$$\operatorname{cov}(\epsilon_a^{(t)}, \epsilon_b^{(t)}) = \frac{\omega_{a,b}^{(t)}}{\omega_{a,a}^{(t)} \omega_{b,b}^{(t)}}. \tag{5}$$

The key representation in (5) entails that accurate estimators of $\omega_{a,b}^{(t)}$ with $a \ne b$ can be constructed if one can estimate $\omega_{a,a}^{(t)}$, $\omega_{b,b}^{(t)}$, and $\operatorname{cov}(\epsilon_a^{(t)}, \epsilon_b^{(t)})$ well.

Another important observation is that under the null hypothesis $H_{0,ab}$ in (1), the conditional distributions of the k classes $X_j^{(t)}|X_{-j}^{(t)}$ with $1 \le t \le k$ indeed share similar sparsity structure on the coefficient vectors $C_j^{(t)}$ thanks to the representation $C_j^{(t)} = -\Omega_{-j,j}^{(t)}/\omega_{j,j}^{(t)}$. In fact, it is clear that $C_{a,b}^{(t)} = 0$ for all $1 \le t \le k$ under $H_{0,ab}$, where $C_{a,b}^{(t)} = -\omega_{a,b}^{(t)}/\omega_{a,a}^{(t)}$ is the component of vector $C_a^{(t)}$ corresponding to feature $X_b^{(t)}$. This observation suggests that we can borrow information from different graphs when testing the joint sparsity structure of multiple graphs. Motivated by such observation, we turn the problem of multiple-network estimation into that of high-dimensional multi-response linear regression

$$\begin{pmatrix} X_j^{(1)} \\ X_j^{(2)} \\ \vdots \\ X_j^{(k)} \end{pmatrix} = \begin{pmatrix} X_{-j}^{(1)} & & & \\ & X_{-j}^{(2)} & & \\ & & \ddots & \\ & & & X_{-j}^{(k)} \end{pmatrix} \begin{pmatrix} C_j^{(1)} \\ C_j^{(2)} \\ \vdots \\ C_j^{(k)} \end{pmatrix} + \begin{pmatrix} \epsilon_j^{(1)} \\ \epsilon_j^{(2)} \\ \vdots \\ \epsilon_j^{(k)} \end{pmatrix} \tag{6}$$

for $1 \le j \le p$. A distinct characteristic of the above multi-response regression model (6) is that it has heterogeneous noises since $\omega_{j,j}^{(t)}$ generally varies over $1 \le t \le k$.

As mentioned before, we also have the group sparsity structure of the regression coefficient vector $C_j^0 = \left(C_j^{(1)\prime}, \ldots, C_j^{(k)\prime}\right)' \in \mathbb{R}^{(p-1)k}$ in model (6). More specifically, denote the k-dimensional subvector of C_j^0 corresponding to the lth group by

$$C_{j(l)}^0 = \left(C_{j,l}^{(1)}, \ldots, C_{j,l}^{(k)}\right)'. \tag{7}$$

Then, we see that $C_{j(l)}^0 = \mathbf{0}$ for all pairs $(j,l) \in \mathcal{E}^c$, the complement of $\mathcal{E}$ defined in (3). We will suggest in Section 3 an efficient estimation procedure that utilizes the group sparsity structure in the regression coefficients and also accounts for the heterogeneity in the noises in model (6).

From now on we work with a sample from model (2), that is, comprised of $n^{(t)}$ independent and identically distributed (iid) observations $X_{1,*}^{(t)}, \ldots, X_{n^{(t)},*}^{(t)}$ for each class $1 \le t \le k$, where $X_{i,*}^{(t)} = (X_{i,1}^{(t)}, \ldots, X_{i,p}^{(t)})' \sim N(0, (\Omega^{(t)})^{-1})$ and the observations across different classes are independent. Suppose that we have some initial estimator $\hat{C}_j^0 = (\hat{C}_j^{(1)\prime}, \ldots, \hat{C}_j^{(k)\prime})'$ for the $(p-1)k$-dimensional regression coefficient vector C_j^0, where we will provide details on one such construction in Section 3. Then the random errors for each $1 \le t \le k$ can be estimated by the residuals

$$\hat{E}_{i,j}^{(t)} = X_{i,j}^{(t)} - X_{i,-j}^{(t)\prime}\hat{C}_j^{(t)} \tag{8}$$

with $1 \le i \le n^{(t)}$ and $1 \le j \le p$. In view of the representation in (5), we can estimate $\omega_{j,j}^{(t)}$ associated with the noise level of class t as $\hat{\omega}_{j,j}^{(t)} = n^{(t)}/(\sum_{i=1}^{n^{(t)}} \hat{E}_{i,j}^{(t)}\hat{E}_{i,j}^{(t)})$. In contrast, the estimation of $\omega_{a,b}^{(t)}$ with $a \ne b$ is slightly more complicated. To estimate the negative covariance $-\mathrm{cov}(\epsilon_a^{(t)}, \epsilon_b^{(t)}) = -\omega_{a,b}^{(t)}/(\omega_{a,a}^{(t)}\omega_{b,b}^{(t)})$, we exploit the following bias corrected statistic

$$T_{n,k,a,b}^{(t)} = \frac{1}{n^{(t)}}\left[\sum_{i=1}^{n^{(t)}} \hat{E}_{i,a}^{(t)}\hat{E}_{i,b}^{(t)} + \sum_{i=1}^{n^{(t)}} \left(\hat{E}_{i,a}^{(t)}\right)^2 \hat{C}_{b,a}^{(t)} + \sum_{i=1}^{n^{(t)}} \left(\hat{E}_{i,b}^{(t)}\right)^2 \hat{C}_{a,b}^{(t)}\right]. \tag{9}$$

Observe that the first term on the right-hand side of (9) corresponds to the sample covariance of the residuals from features $X_a^{(t)}$ and $X_b^{(t)}$. When $a = b$, this sample covariance is asymptotically unbiased in estimating $\mathrm{var}(\epsilon_a^{(t)}) = 1/\omega_{a,a}^{(t)}$. Such sample covariance is, however, biased in the case of $a \ne b$ and thus two additional terms are introduced for $T_{n,k,a,b}^{(t)}$ in (9) to correct the bias. Indeed, we can show that after the bias correction the statistic $T_{n,k,a,b}^{(t)}$ is asymptotically close to the statistic

$$J_{n,k,a,b}^{(t)} = \left[1 - \omega_{a,a}^{(t)}(\hat{\omega}_{a,a}^{(t)})^{-1} - \omega_{b,b}^{(t)}(\hat{\omega}_{b,b}^{(t)})^{-1}\right]\frac{\omega_{a,b}^{(t)}}{\omega_{a,a}^{(t)}\omega_{b,b}^{(t)}}, \tag{10}$$

which is in turn asymptotically close to the negative covariance $-\mathrm{cov}(\epsilon_a^{(t)}, \epsilon_b^{(t)})$.

When there is only a single graph, that is, $k = 1$, the above statistic $T_{n,k,a,b}^{(t)}$ in (9) reduces to the one introduced by Liu (2013) to address the bias issue in the testing for a single Gaussian graph. In the scenario of multiple graphs, we observe a similar phenomenon and provide in Theorem 2.1 later a formal theoretical justification. It is worth mentioning that the key estimators $\hat{\omega}_{j,j}^{(t)}$ and $T_{n,k,a,b}^{(t)}$ introduced above are constructed using the residuals $\hat{E}_{i,j}^{(t)}$ instead of the estimated regression coefficients $\hat{C}_{a,b}^{(t)}$, though the regression coefficients $C_{a,b}^{(t)}$ are also closely related to the entries of the precision matrix $\Omega^{(t)}$. The main advantage of using residuals $\hat{E}_{i,j}^{(t)}$ over coefficients $\hat{C}_{a,b}^{(t)}$ is rooted in the fact that obtaining asymptotically unbiased estimates of the latter is much more challenging in high dimensions, largely due to the well-known bias issue associated with the regularization methods, than accurately estimating the former, which is closely related to the prediction problem.

The new formulation in (6) not only allows us to solve the problem of multiple-graph estimation efficiently through p multi-response regressions as detailed in Section 3, but also enables us to construct new tests that are more powerful than existing methods by borrowing information from different graphs. We are now ready to present the first such test. Due to the group sparsity structure and the target of our null hypothesis $H_{0,ab} : \omega_{a,b}^0 = \mathbf{0}$ in (1), we naturally construct our test statistics using certain functions of all statistics $T_{n,k,a,b}^{(t)}$ in (9) with $1 \le t \le k$. Thanks to the joint estimation accuracy for the $(p-1)k$-dimensional regression coefficient vector C_j^0, we define our first test statistic, the chi-based test statistic $U_{n,k,a,b}$, as

$$U_{n,k,a,b}^2 = \sum_{t=1}^{k} n^{(t)}\hat{\omega}_{b,b}^{(t)}\hat{\omega}_{a,a}^{(t)}\left(T_{n,k,a,b}^{(t)}\right)^2 \tag{11}$$

for testing the null hypothesis $H_{0,ab}$ against the alternative hypothesis for which the condition is imposed on the ℓ_2 norm

$\|\omega^0_{a,b}\|$. In other words, our test statistic is powerful whenever the signal strength $\|\omega^0_{a,b}\|$ is larger than some testable region boundary, which will be characterized later in Section 2.4.

To characterize the limiting distribution of the chi-based test statistic $U_{n,k,a,b}$ in (11) under the null, we introduce two additional statistics $V^{*(t)}_{n,k,a,b}$ and $U^*_{n,k,a,b}$ as

$$V^{*(t)}_{n,k,a,b} = \sqrt{\frac{\omega^{(t)}_{b,b}\tilde{\omega}^{(t)}_{a,a}}{n^{(t)}}} \sum_{i=1}^{n^{(t)}} \left(E^{(t)}_{i,a}E^{(t)}_{i,b} - \mathbb{E}E^{(t)}_{i,a}E^{(t)}_{i,b}\right), \tag{12}$$

$$\begin{aligned} U^{*2}_{n,k,a,b} &= \sum_{t=1}^{k} \left(V^{*(t)}_{n,k,a,b}\right)^2 \\ &= \sum_{t=1}^{k} \frac{\omega^{(t)}_{b,b}\tilde{\omega}^{(t)}_{a,a}}{n^{(t)}} \left[\sum_{i=1}^{n^{(t)}} \left(E^{(t)}_{i,a}E^{(t)}_{i,b} - \mathbb{E}E^{(t)}_{i,a}E^{(t)}_{i,b}\right)\right]^2, \end{aligned} \tag{13}$$

where $E^{(t)}_{i,j} = X^{(t)}_{i,j} - X^{(t)\prime}_{i,-j}C^{(t)}_j$ is the random error and $\tilde{\omega}^{(t)}_{j,j} = n^{(t)}/(E^{(t)\prime}_{*,j}E^{(t)}_{*,j})$ is the oracle estimator of $\omega^{(t)}_{jj}$ since the random error vector $E^{(t)}_{*,j} = (E^{(t)}_{1,j}, \ldots, E^{(t)}_{n^{(t)},j})'$ is unobservable in practice. Hereafter, the expectation sign is denoted as $\mathbb{E}$ to distinguish from the random error $E^{(t)}_{(i,j)}$. It is interesting to observe that under the null, the Gaussian vector $E^{(t)}_{*,b} \sim N(0, (\omega^{(t)}_{b,b})^{-1}I)$ is independent of $E^{(t)}_{*,a}$, which entails that $V^{*(t)}_{n,k,a,b} \sim N(0,1)$ and they are independent of each other over $1 \le t \le k$. Consequently, under the null hypothesis $H_{0,ab}$ in (1) it holds that $U^{*2}_{n,k,a,b} \sim \chi^2(k)$.

Before formally presenting our first main result, we introduce the following two regularity conditions on our model (2).

Condition 2.1. There exists some constant $M > 0$ such that $1/M \le \lambda_{\min}(\Omega^{(t)}) \le \lambda_{\max}(\Omega^{(t)}) \le M$ for each $1 \le t \le k$, where $\lambda_{\min}$ and $\lambda_{\max}$ denote the smallest and largest eigenvalues of a matrix.

Condition 2.2. It holds that $n^{(1)} \asymp \cdots \asymp n^{(k)}$ with $\max_{1\le t\le k}\{n^{(t)}\}/n^{(0)} \le M_0$, where $\asymp$ means the same order, $n^{(0)} = \min_{1\le t\le k}\{n^{(t)}\}$, and M_0 is some positive constant.

The well-conditionedness of the precision matrices $\Omega^{(t)}$ assumed in Condition 2.1 simplifies our technical presentation. For simplicity, we also assume in Condition 2.2 that our sample is balanced with the sample sizes of each of the k classes comparable to each other. With slight abuse of notation, we denote by $n^{(0)}$ this common level whenever the rate is involved. We proceed with introducing additional notation and technical conditions. Denote by $\Delta_j = \hat{C}^0_j - C^0_j$ and $\Delta_{j(l)} = \hat{C}^0_{j(l)} - C^0_{j(l)}$ the estimation errors of $\hat{C}^0_j$ and $\hat{C}^0_{j(l)}$, respectively, with the k-dimensional subvector $\hat{C}^0_{j(l)}$ defined in a similar way to $C^0_{j(l)}$ in (7). To characterize the sparsity level, we define the joint sparsity of the k networks as the maximum node degree corresponding to the edge set $\mathcal{E}$ in (3),

$$s \equiv \max_{1\le a\le p} \sum_{1\le b\ne a\le p} 1\{\omega^0_{a,b} \ne \mathbf{0}\}. \tag{14}$$

We further assume that with high probability the initial estimator $\hat{C}^0_j$ satisfies

$$\frac{1}{\sqrt{k}}\|\Delta_j\| = \frac{1}{\sqrt{k}}\left\|\hat{C}^0_j - C^0_j\right\| \le C_1\left[s\frac{1+(\log p)/k}{n^{(0)}}\right]^{1/2}, \tag{15}$$

$$\sum_{l\ne j}\frac{1}{\sqrt{k}}\|\Delta_{j(l)}\| = \sum_{l\ne j}\frac{1}{\sqrt{k}}\left\|\hat{C}^0_{j(l)} - C^0_{j(l)}\right\| \le C_2 s\left[\frac{1+(\log p)/k}{n^{(0)}}\right]^{1/2}, \tag{16}$$

$$\frac{1}{k}\sum_{t=1}^{k}\frac{\left\|X^{(t)}_{*,-j}\left(\hat{C}^{(t)}_j - C^{(t)}_j\right)\right\|^2}{n^{(t)}} \le C_3 s\frac{1+(\log p)/k}{n^{(0)}}, \tag{17}$$

where C_1, C_2, and C_3 are some positive constants and $\|\cdot\|$ denotes the ℓ_2 norm. The properties (15)–(17) are crucial working assumptions in our testing for k networks.

In Section 3, we will provide one valid and tuning-free construction of initial estimators with the above desired properties. A distinct characteristic is that the analysis of our tuning-free estimator is new due to the heterogeneity of noises across different classes, which makes typical tuning-free procedures such as the scaled Lasso (Sun and Zhang 2012) and the square-root Lasso (Belloni, Chernozhukov, and Wang 2011) no longer work in the current setting; see Section 3 for more detailed discussions.

Theorem 2.1. Assume that Conditions 2.1–2.2 hold, the initial estimators $\hat{C}^0_j$ each satisfy properties (15)–(17) with probability at least $1 - C_0p^{1-\delta}$, $s\left(k + \log p\right)/n^{(0)} = o(1)$, and $\log(k/\delta_1) = O\{s[1 + (\log p)/k]\}$ for some constants $C_0 > 0, \delta > 1$ and $\delta_1 = o(1)$. Then for each pair (a, b) with $1 \le a \ne b \le p$, it holds with probability at least $1 - (12 + C_0)p^{1-\delta} - 4\delta_1$ that

$$\left|\left[\sum_{t=1}^{k} n^{(t)}\hat{\omega}^{(t)}_{b,b}\hat{\omega}^{(t)}_{a,a}\left(T^{(t)}_{n,k,a,b} - J^{(t)}_{n,k,a,b}\right)^2\right]^{1/2} - U^*_{n,k,a,b}\right| \le C\left(s\frac{k+\log p}{\sqrt{n^{(0)}}}\right),$$

where $C > 0$ is some constant. Moreover, under null hypothesis $H_{0,ab}$ in (1) we have $U^{*2}_{n,k,a,b} \sim \chi^2(k)$ and with the same probability bound that $\left|U_{n,k,a,b} - U^*_{n,k,a,b}\right| \le C\left(s\frac{k+\log p}{\sqrt{n^{(0)}}}\right)$.

The coupling result in Theorem 2.1 motivates us to propose the chi-based test ϕ_2 defined as

$$\phi_2 = 1\left\{U_{n,k,a,b} > z^{l2}_k(1-\alpha)\right\} \tag{18}$$

for our THI framework in multiple networks which tests the null hypothesis $H_{0,ab}$ in (1) using the test statistic $U_{n,k,a,b}$ given in (11), where $\alpha \in (0,1)$ is a fixed significance level and $z^{l2}_k(1-\alpha)$ denotes the $100(1-\alpha)$th percentile of the chi distribution with degrees of freedom k. The name of this test is from the property that the null distribution of the test statistic is asymptotically close to the chi distribution.

Proposition 2.1. Assume that all the conditions of Theorem 2.1 hold and $s^2(k+\log p)^2 = o(n^{(0)})$. Then the chi-based test ϕ_2 in (18) has asymptotic significance level α.

As formally justified in Proposition 2.1, the chi-based test ϕ_2 introduced in (18) is indeed an asymptotic test with significance level α under the sample size requirement of $n^{(0)} \gg s^2(k+\log p)^2$, in the asymptotic setting in which the number of nodes p, the number of networks k, and the joint sparsity of the networks s can diverge simultaneously as the common level of sample sizes $n^{(0)} \to \infty$.

2.3. Linear Functional-Based Test

The chi-based test ϕ_2 introduced in Section 2.2 serves as a general procedure to test whether the joint link strength vector $\omega^0_{a,b}$ is zero when there is no additional information assumed on the k networks. In some scenarios when certain extra knowledge is available, it is possible to design more powerful testing procedures. In this spirit, we now present an alternative test for our THI framework in multiple networks based on a linear functional of $\omega^0_{a,b}$, which is closely related to its ℓ_1 norm. The main motivation is that in some applications such as the GWAS example mentioned in Section 1 (Marigorta and Navarro 2013), the sign relationship of some target edge across k graphs is provided implicitly or explicitly. For example, one may expect that all the $\omega^{(t)}_{a,b}$ with $1 \le t \le k$ share the same sign, that is, they are either all nonpositive or all nonnegative. In such a scenario, testing the null hypothesis $H_{0,ab} : \omega^0_{a,b} = \mathbf{0}$ is equivalent to testing $\|\omega^0_{a,b}\|_1 = |\sum_{t=1}^k \omega^{(t)}_{a,b}| = 0$. In a more general setting, the sign relationship can be represented by a unique sign vector $\xi = (\xi_1, \ldots, \xi_k)' \in \{1, -1\}^k$, up to a single sign, such that $\|\omega^0_{a,b}\|_1 = \sum_{t=1}^k \xi_t \omega^{(t)}_{a,b}$ or $\|\omega^0_{a,b}\|_1 = |\sum_{t=1}^k \xi_t \omega^{(t)}_{a,b}|$, and thus the null hypothesis $H_{0,ab} : \omega^0_{a,b} = \mathbf{0}$ takes an equivalent form of $\|\omega^0_{a,b}\|_1 = |\sum_{t=1}^k \xi_t \omega^{(t)}_{a,b}| = 0$.

Given the above sign vector ξ, we define our second test statistic, the linear functional-based test statistic $V_{n,k,a,b}(\xi)$, as

$$V_{n,k,a,b}(\xi) = \sum_{t=1}^k \xi_t \sqrt{n^{(t)} \hat{\omega}^{(t)}_{a,a} \hat{\omega}^{(t)}_{b,b}}\, T^{(t)}_{n,k,a,b} \tag{19}$$

with the bias corrected statistic $T^{(t)}_{n,k,a,b}$ given in (9). Intuitively, with the sign information, the proposed test statistic $V_{n,k,a,b}(\xi)$ has the same order of magnitude as $\sqrt{n}\|\omega^0_{a,b}\|_1$, noting that $T^{(t)}_{n,k,a,b}$ is close to $-\mathrm{cov}(\epsilon^{(t)}_a, \epsilon^{(t)}_b)$. To characterize the limiting distribution of the linear functional-based test statistic $V_{n,k,a,b}$ under the null, we introduce another statistic $V^*_{n,k,a,b}(\xi)$ as

$$\begin{aligned} V^*_{n,k,a,b}(\xi) &= \sum_{t=1}^k \xi_t V^{*(t)}_{n,k,a,b} \\ &= \sum_{t=1}^k \xi_t \sqrt{\frac{\omega^{(t)}_{b,b} \tilde{\omega}^{(t)}_{a,a}}{n^{(t)}}} \sum_{i=1}^{n^{(t)}} \left(E^{(t)}_{i,a} E^{(t)}_{i,b} - \mathbb{E} E^{(t)}_{i,a} E^{(t)}_{i,b} \right), \end{aligned}$$

where the statistic $V^{*(t)}_{n,k,a,b}$ is given in (12). With the extra sign information, our new test statistic is powerful whenever the signal strength $\|\omega^0_{a,b}\|_1$ becomes large; see Section 2.4 for the characterization of the testable region boundary under the alternative hypothesis for which the condition is imposed on the ℓ_1 norm $\|\omega^0_{a,b}\|_1$. It is easy to see that under the null, $V^{*(t)}_{n,k,a,b} \sim N(0,1)$ are independent of each other over $1 \le t \le k$, and consequently $V^*_{n,k,a,b}(\xi) \sim N(0,k)$ for any given sign vector ξ.

Theorem 2.2. Assume that all the conditions of Theorem 2.1 hold. Then for each pair (a,b) with $1 \le a \ne b \le p$, it holds with probability at least $1 - (12 + C_0)p^{1-\delta} - 4\delta_1$ that

$$\begin{aligned} &\left| \sum_{t=1}^k \xi_t \left[\sqrt{n^{(t)} \hat{\omega}^{(t)}_{b,b} \hat{\omega}^{(t)}_{a,a}} \left(T^{(t)}_{n,k,a,b} - J^{(t)}_{n,k,a,b} \right) - V^{*(t)}_{n,k,a,b} \right] \right| \\ &\quad \le C \left(s \frac{k+\log p}{\sqrt{n^{(0)}}} \right), \end{aligned} \tag{20}$$

where $C > 0$ is some constant. Moreover, under null hypothesis $H_{0,ab}$ in (1) we have $J^{(t)}_{n,k,a,b} = 0$, $V^*_{n,k,a,b}(\xi) \sim N(0,k)$ and with the same probability bound, $\left| V_{n,k,a,b}(\xi) - V^*_{n,k,a,b}(\xi) \right| \le C \left(s \frac{k+\log p}{\sqrt{n^{(0)}}} \right)$.

Theorem 2.2 quantifies the asymptotic behavior of the linear functional-based test statistic $V_{n,k,a,b}(\xi)$ under the null hypothesis $H_{0,ab}$ in (1). Assume further that the sign vector ξ is given uniquely such that $\|\omega^0_{a,b}\|_1 = \sum_{t=1}^k \xi_t \omega^{(t)}_{a,b}$ under the alternative hypothesis. Then Theorem 2.2 and the definition of the statistic $J^{(t)}_{n,k,a,b}$ in (10) motivate us to propose a one-sided test, the linear functional-based test ϕ_1, defined as

$$\phi_1 = 1 \left\{ \frac{V_{n,k,a,b}(\xi)}{\sqrt{k}} < z(\alpha) \right\} \tag{21}$$

for our THI framework in multiple networks, where $\alpha \in (0,1)$ is a fixed significance level and $z(\alpha)$ stands for the 100αth percentile of the standard Gaussian distribution. When the sign vector ξ is given up to a single sign, for example, when we know only that all the signs ξ_t with $1 \le t \le k$ are identical, it is more natural to define a two-sided test. We omit the details of such two-sided test for simplicity.

Proposition 2.2. Assume that all the conditions of Theorem 2.2 hold and $s^2 k^{-1}(k+\log p)^2 = o(n^{(0)})$. Then the linear functional-based test ϕ_1 in (21) has asymptotic significance level α.

Remark 2.1. The sample size requirements in Propositions 2.1–2.2 can be weakened if one only cares about the statistical inference of the joint link strength vector $\omega^0_{(a,b)}$ for some fixed pair (a,b). Indeed, Propositions 2.1–2.2 are still valid as long as the joint degrees of k networks for nodes a and b satisfy the corresponding sample size requirements. That is, one can replace s by $\max_{i \in \{a,b\}} \sum_{1 \le j \ne i \le p} 1\{\omega^0_{i,j} \ne \mathbf{0}\}$. This is a weaker assumption because the degrees for nodes other than a and b can be much larger.

Proposition 2.2 which is based on Theorem 2.2 shows that the linear functional-based test ϕ_1 introduced in (21) is indeed an asymptotic test with significance level α under the sample size requirement of $n^{(0)} \gg s^2 k^{-1}(k+\log p)^2$. It is worth mentioning

that most existing results in the literature either focus on testing procedures for a single graph or develop estimation procedures for multiple graphs without statistical inference in high dimensions. In contrast, our developments in Theorems 2.1–2.2 and Propositions 2.1–2.2 provide testing procedures in multiple graphs for the first time. For the case of a single graph with $k = 1$, our test statistics essentially reduce to the one introduced in Liu (2013). This suggests an alternative way of constructing test statistics, which is to construct a test statistic for each individual graph $1 \le t \le k$ as in Liu (2013) and then naively pool them together in the same way as for our ϕ_2 and ϕ_1.

Let us gain some insights into our tests with a comparison to the above naive combination procedure. The advantage of our linear functional-based test ϕ_1 is reflected on the sample size requirement of $s^2k^{-1}(k+\log p)^2 = o(n^{(0)})$ established in Proposition 2.2, thanks to the information of structural similarity across the k graphs which makes the working assumptions (15)–(17) possible. In comparison, to test the null hypothesis $H_{0,ab} : \omega^0_{a,b} = \mathbf{0}$ one can also apply the procedure in Liu (2013) to each of the k graphs and then construct a similar linear functional-based test as in (21). For such naive combination procedure, it can be shown that a stronger sample size assumption $s^2k\left(\log p\right)^2 = o(n^{(0)})$ is required. In fact, we further establish in Section 2.4 that the sample size requirement $s^2k^{-1}(k+\log p)^2 = o(n^{(0)})$ for our linear functional-based test ϕ_1 is minimal in a decision theoretic framework.

Similarly the advantage of our chi-based test ϕ_2 is rooted in the sample size requirement of $s^2(k+\log p)^2 = o(n^{(0)})$ obtained in Proposition 2.1. In contrast, one can also construct a similar chi-based test as in (18) based on the residuals $\hat{E}^{(t)}_{i,j}$ which are obtained through an application of the procedure in Liu (2013) to each individual graph. For such naive combination testing procedure, it can be shown that the sample size assumption $s^2k\left(\log p\right)^2 = o(n^{(0)})$ is required. This demonstrates that in a range of typical scenarios when the number of networks does not grow excessively fast with $k = o\{\left(\log p\right)^2\}$, our chi-based test ϕ_2 indeed has a weaker sample size requirement.

2.4. Optimality of Tests and Minimum Sample Size Requirement

So far we have introduced our THI framework in multiple networks with two different types of tests for testing the null hypothesis $H_{0,ab} : \omega^0_{a,b} = \mathbf{0}$ in (1). The constructions of our test statistics are motivated by the possible alternative hypothesis. In particular, the chi-based test ϕ_2 should be powerful as long as the joint link strength $\|\omega^0_{a,b}\|$ is away from zero, while the linear functional-based test ϕ_1 will be powerful when the signs of $\omega^0_{a,b}$ are known and $\|\omega^0_{a,b}\|_1$ becomes large. Along this direction, we now further investigate two types of composite alternative hypotheses. We define the set of all s-sparse multiple networks as

$$\mathcal{F}(s) = \mathcal{F}(s, M) = \left\{ \Omega^0 : \max_{1\le a\le p} \sum_{1\le b\ne a\le p} 1\{\omega^0_{a,b} \ne \mathbf{0}\} \le s \right. \\ \left. \text{and Condition 2.1 holds} \right\}, \tag{22}$$

where $\Omega^0 = \{\Omega^{(t)}\}_{t=1}^k$ stands for the set of k precision matrices with slight abuse of notation and s is some positive integer. Then the null hypothesis $H_{0,ab}$ in (1) can be rewritten as

$$H_{0,ab} = H_{0,ab}(s) : \Omega^0 \in \mathcal{N}(s) \equiv \left\{\Omega^0 : \Omega^0 \in \mathcal{F}(s),\ \omega^0_{a,b} = \mathbf{0}\right\}. \tag{23}$$

In particular, we consider the following two alternative hypotheses

$$H^{l2}_{1,ab}(s,\epsilon) : \Omega^0 \in \mathcal{A}^{l2}(s,\epsilon) \equiv \left\{\Omega^0 : \Omega^0 \in \mathcal{F}(s),\ \left\|\omega^0_{a,b}\right\| \ge \epsilon\right\}, \tag{24}$$

$$H^{l1}_{1,ab}(s,\epsilon,\xi) : \Omega^0 \in \mathcal{A}^{l1}(s,\epsilon,\xi) \\ \equiv \left\{\Omega^0 : \Omega^0 \in \mathcal{F}(s),\ \xi'\omega^0_{a,b} = \left\|\omega^0_{a,b}\right\|_1 \ge \epsilon\right\}, \tag{25}$$

where the former is introduced to investigate the chi-based test ϕ_2, the latter is for the linear functional-based test ϕ_1, and $\epsilon > 0$.

It is clear that the difficulty of testing the null $H_{0,ab}$ in (23) against the alternative $H^{l2}_{1,ab}(s,\epsilon)$ in (24) or against the alternative $H^{l1}_{1,ab}(s,\epsilon,\xi)$ in (25) depends critically on the quantity ϵ. The smaller ϵ is, the more difficult to distinguish between the null and alternative hypotheses. A natural and fundamental question is what the boundary of the testable region is. Such a boundary means that it is impossible to detect whether the observations are from the null against the alternative as long as ϵ is smaller than it, while there exists some test which can distinguish between the two hypotheses whenever ϵ is far larger than it.

To characterize the testable region boundary, we introduce the separating rate ϵ_n of null $H_{0,ab}$ against alternative $H^{l2}_{1,ab}(s,\epsilon)$ or $H^{l1}_{1,ab}(s,\epsilon,\xi)$. For any fixed significance level $\alpha \in (0,1)$ and power $\alpha < \beta < 1$, the *separating rate* for alternative $H_1 = H^{l2}_{1,ab}(s,\epsilon)$ or $H^{l1}_{1,ab}(s,\epsilon,\xi)$ is said to be ϵ_n if there exist some test ψ_0 of asymptotic significance level α and some absolute large constant $c > 0$ such that

$$\lim_{n^{(0)}\to\infty} \inf_{\nu\in\mathcal{A}(c)} \mathbb{P}_\nu(\psi_0 \text{ rejects } H_{0,ab}) \ge \beta, \tag{26}$$

while there exists some absolute small constant $c' > 0$ such that for any test ψ of asymptotic significance level α, it holds that

$$\lim_{n^{(0)}\to\infty} \inf_{\nu\in\mathcal{A}(c')} \mathbb{P}_\nu(\psi \text{ rejects } H_{0,ab}) < \beta, \tag{27}$$

where $\mathcal{A}(c)$ represents $\mathcal{A}^{l2}(s, c\epsilon_n)$ or $\mathcal{A}^{l1}(s, c\epsilon_n, \xi)$. By symmetry, it is easy to see that the separating rate ϵ_n for alternative $H^{l1}_{1,ab}(s,\epsilon,\xi)$ defined above is free of the sign vector ξ.

Our major goals in this section are 2-fold. First, we identify the separating rates ϵ_n for alternative $H^{l2}_{1,ab}(s,\epsilon)$ under the sample size assumption $s^2(k+\log p)^2 = o(n^{(0)})$ and for alternative $H^{l1}_{1,ab}(s,\epsilon,\xi)$ under the sample size assumption $s^2k^{-1}(k+\log p)^2 = o(n^{(0)})$. In particular, we show later in Theorem 2.3 that $\epsilon_n \asymp \sqrt{k^{1/2}/n^{(0)}}$ for alternative $H^{l2}_{1,ab}(s,\epsilon)$ and $\epsilon_n \asymp \sqrt{k/n^{(0)}}$ for alternative $H^{l1}_{1,ab}(s,\epsilon,\xi)$. Moreover, our newly suggested chi-based test ϕ_2 and linear functional-based test ϕ_1 achieve these two separating rates, respectively, and hence are optimal in this sense. Second, we investigate the optimality of the sample size assumption $s^2k^{-1}(k+\log p)^2 = o(n^{(0)})$

for the ℓ_1 type alternative $H^{l1}_{1,ab}(s,\epsilon,\xi)$ in (25). Specifically, we establish later in Theorem 2.4 that in order to have separating rate $\epsilon_n \asymp \sqrt{k/n^{(0)}}$, this sample size requirement is necessary under the setting of $k = O(\log p)$. Therefore, we conclude that the linear functional-based test ϕ_1 is optimal to test null $H_{0,ab}$ from alternative $H^{l1}_{1,ab}(s,\epsilon,\xi)$ under the minimum sample size requirement. It is worth mentioning that major contributions of our second goal lie in a new construction of a related minimax lower bound argument.

Theorem 2.3. (1) Under the conditions of Proposition 2.1, the separating rate for testing $H_{0,ab}$ against $H^{l2}_{1,ab}(s,\epsilon)$ is $\epsilon_n = \sqrt{k^{1/2}/n^{(0)}}$ and the chi-based test ϕ_2 in (18) achieves this rate, that is, for any given $\beta > \alpha$, (26) is valid with $\psi_0 = \phi_2$ and $\mathcal{A}(c) = \mathcal{A}^{l2}(s, c\epsilon_n)$ for some sufficiently large constant $c > 0$.

(2) Under the conditions of Proposition 2.2, the separating rate for testing $H_{0,ab}$ against $H^{l1}_{1,ab}(s,\epsilon,\xi)$ is $\epsilon_n = \sqrt{k/n^{(0)}}$ and the linear functional-based test ϕ_1 in (21) achieves this rate.

In fact, the detection problems of the separating rates for $H^{l2}_{1,ab}(s,\epsilon)$ and $H^{l1}_{1,ab}(s,\epsilon,\xi)$ investigated in Theorem 2.3 are closely related to those of optimal quadratic functional and linear functional estimation for Gaussian sequence models, respectively. See, for example, Baraud (2002), Ingster and Suslina (2012), and Collier, Comminges, and Tsybakov (2017) for more details. Yet Gaussian graphical models are much more complicated than Gaussian sequence models. Even for the simple setting of $k = 1$, it was shown in Ren et al. (2015) that minimax estimation of each single edge $\omega_{a,b}$ can be different from the parametric rate $\sqrt{n}$. This subtle difference is reflected in the sample size requirements stated in Theorem 2.3 for the setting of multiple networks.

Theorem 2.4. Assume that $k \le M_1 \log p$, $s > 2$, $s^2k^{-1}(k + \log p)^2 > Cn^{(0)}$, $p > s^\mu$, and $s[1 + (\log p)/k]/n^{(0)} = o(1)$ for some large constants $M_1, C > 0$ and some $\mu > 2$. Then given any $\alpha < \beta < 1$ and some constant $c > 0$, there exists no test of asymptotic significance level α satisfying (26) with $\mathcal{A}(c) = \mathcal{A}^{l1}(s, c\epsilon_n, \xi)$ and $\epsilon_n = \sqrt{k/n^{(0)}}$.

Theorem 2.4 further justifies that the sample size requirement of $s^2k^{-1}(k + \log p)^2 = o(n^{(0)})$ for the ℓ_1 type alternative $H^{l1}_{1,ab}(s,\epsilon,\xi)$ in (25) is indeed sharp. To obtain such result, one needs to construct a lower bound involving the sample size requirement and the separating rate. For the single graph setting of $k = 1$, this is related to the minimax lower bound of estimating each single edge $\omega_{a,b}$, which was explored in Ren et al. (2015). The lower bound argument in Ren et al. (2015) is, however, not applicable to the current setting even for the case of $k = 1$, since the construction of the least favorable subset of the parameter space in Ren et al. (2015) does not allow $\omega_{a,b}$ to be close to zero, which is in fact the focus of the testing problem. To overcome such difficulty, we propose a very different least favorable subset in our analysis of Theorem 2.4.

2.5. Comparisons With Existing Methods

As mentioned in the Section 1, there is a rich and growing line of research on multiple networks in the setting of Gaussian graphical models. Due to the space constraint, we compare our procedure with some most relevant ones in the literature. Our work makes no assumption on the ordering for the k networks. Existing work along this line includes, for instance, Guo et al. (2011), Danaher, Wang, and Witten (2014), Zhu, Shen, and Pan (2014), and Cai et al. (2016). The main advantages of our proposed THI method over these existing approaches are 3-fold. First, our THI framework with the two specific testing procedures provides statistical inference for each joint link strength vector $\omega^0_{a,b}$ over k networks to reflect its statistical significance. This is of crucial importance for model interpretation, false discovery rate control, and global multiple precision matrices estimation in applications. In contrast, none of these previous attempts along this line goes beyond point estimation to investigate statistical inference.

Second, our theoretically optimal procedure is tuning free and data driven. This is mainly due to a novel approach of HGSL as a convex program as well as a computationally fast algorithm with convergence guarantees suggested in Section 3 for the setting of high-dimensional multi-response regression with heterogeneous noises, which may be of independent interest. Different from ours, all existing methods typically involve one or more tuning parameters. Moreover, some of these methods rely on nonconvex optimization problems whose global solutions cannot always be guaranteed to be computable. In contrast, our procedure not only enjoys the computational efficiency but also avoids the additional practical and theoretical issues caused by the use of the cross-validation; see the simulation studies in Section 4.1 for a detailed comparison on the computational cost of our algorithm with competitors which demonstrates the computational advantage of our procedure. Third, our procedure admits the optimality properties established for two different types of tests in terms of the separating rates, which follow from three new lower bound arguments introduced in Sections C.3 and C.4 of the supplementary material. To the best of our knowledge, there are no such immediate results available in the literature of multiple Gaussian graphical models. The obtained optimality results ensure that our testing procedures are optimal.

More thorough theoretical comparisons of our method with competitors are possible but involved, particularly given that no results of hypothesis testing are provided for these existing methods. For a fair comparison, we now focus on the requirements for support recovery results of different methods under the assumption that all k graphs share a common sparsity structure. To this end, we need to go a little further based on our chi-based test ϕ_2 by replacing α in (18) by $p^{-2-\rho}$ with some $\rho > 0$. Specifically, for any given $\rho > 0$ we define the THI estimator $\hat{\mathcal{E}}$ for the support or edge set $\mathcal{E}$ corresponding to the k graphs in (3) as

$$(a,b) \in \hat{\mathcal{E}} \quad \text{when} \quad U_{n,k,a,b} > z^{l2}_k(1 - p^{-2-\rho}), \tag{28}$$

where all the notation is the same as in (18). The following proposition establishes that the THI estimator $\hat{\mathcal{E}}$ introduced in (28) is indeed capable of recovering the network structure exactly with large probability as long as the minimum signal strength is above a certain threshold.

Proposition 2.3. Assume that all the conditions of Proposition 2.1 hold and $\min_{(a,b)\in\mathcal{E}} \|\omega^0_{a,b}\| > C\sqrt{[(k\log p)^{1/2} + \log p]/n^{(0)}}$ for some sufficiently large constant $C > 0$. Then the THI estimator $\hat{\mathcal{E}}$ given in (28) satisfies $\hat{\mathcal{E}} = \mathcal{E}$ with probability at least $1 - O(p^{-\rho})$.

In view of the separating rate $C\sqrt{k^{1/2}/n^{(0)}}$ obtained in Theorem 2.3 (1) for a single joint link strength vector, we see that the lower bound on the minimum signal strength $\min_{(a,b)\in\mathcal{E}} \|\omega^0_{a,b}\|$ in Proposition 2.3 for support recovery comes with an extra factor of $(\log p)^{1/4}$ for the case of $\log p = O(k)$, or with the factor $k^{1/4}$ replaced by $(\log p)^{1/2}$ for the case of $k = O(\log p)$. We would like to point out that such increased minimum signal strength generally cannot be avoided and stems from the union bound argument taken over all pairs of nodes (a, b) in the edge set $\mathcal{E}$.

Let us gain some insights into the advantage of our THI procedure on support recovery in comparison to some existing approaches. To recover the support successfully, at least the minimum signal strength requirement of $\min_{(a,b)\in\mathcal{E}} \|\omega^0_{a,b}\| \geq C\sqrt{k}$ is needed in Guo et al. (2011), and the assumption of $\min_{(a,b)\in\mathcal{E}} \|\omega^0_{a,b}\| \geq CM_n\sqrt{(k\log p)/n}$ is needed in Cai et al. (2016), where $M_n \equiv \max_{1\leq t\leq k} \max_{1\leq b\leq p} \Sigma^p_{a=1} |\omega^{(t)}_{a,b}|$ denoting the largest matrix 1-norm among k graphs can diverge with $n^{(0)}$ under our setting, and $C > 0$ is some constant. In addition, no theoretical justification is provided in Danaher, Wang, and Witten (2014), and the support recovery result in Zhu, Shen, and Pan (2014) cannot be easily compared due to an extra clustering structural assumption. In summary, compared with existing methods our optimal THI approach yields a sharper minimum signal strength requirement for recovering the support of the networks with common structure, thanks to our optimal testing procedures.

3. Tuning-Free Heterogeneous Group Square-Root Lasso

Our THI framework suggested in Section 2 for uncovering the heterogeneity in sparsity patterns among multiple networks via large-scale inference relies critically on an efficient procedure for fitting the high-dimensional multi-response linear regression model (6) for each node $1 \leq j \leq p$. We now introduce such an approach HGSL that can be of independent interest when one is in need of a tuning-free method for the general setting of high-dimensional multi-response regression with heterogeneous noises. Specifically, we need to construct some initial estimators $\hat{C}^0_j = (\hat{C}^{(1)\prime}_j, \ldots, \hat{C}^{(k)\prime}_j)'$ for the $(p-1)k$-dimensional regression coefficient vectors $C^0_j = \left(C^{(1)\prime}_j, \ldots, C^{(k)\prime}_j\right)'$ in model (6) with $1 \leq j \leq p$ that each satisfy properties (15)–(17) with significant probability, say, at least $1 - C_0 p^{1-\delta}$ for some positive constants C_0 and $\delta > 1$.

By symmetry, we can focus only on the case of $j = 1$, hereafter without loss of generality. Recall that in our model (2), for each graph $1 \leq t \leq k$ we have an $n^{(t)} \times p$ data matrix $\mathbf{X}^{(t)} = (X^{(t)}_{1,*}, \ldots, X^{(t)}_{n^{(t)},*})'$ with iid rows $X^{(t)}_{i,*} = (X^{(t)}_{i,1}, \ldots, X^{(t)}_{i,p})' \sim N(0, (\Omega^{(t)})^{-1})$ for $1 \leq i \leq n^{(t)}$. Using the matrix notation, the multi-response linear regression model (6) can be rewritten as

$$\begin{pmatrix} X^{(1)}_{*,1} \\ X^{(2)}_{*,1} \\ \vdots \\ X^{(k)}_{*,1} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^{(1)}_{*,-1} & & & \\ & \mathbf{X}^{(2)}_{*,-1} & & \\ & & \ddots & \\ & & & \mathbf{X}^{(k)}_{*,-1} \end{pmatrix} \begin{pmatrix} C^{(1)}_1 \\ C^{(2)}_1 \\ \vdots \\ C^{(k)}_1 \end{pmatrix} + \begin{pmatrix} E^{(1)}_{*,1} \\ E^{(2)}_{*,1} \\ \vdots \\ E^{(k)}_{*,1} \end{pmatrix} \equiv \mathbf{X}^0_{*,-1} C^0_1 + E^0_{*,1} \tag{29}$$

lying in the N-dimensional Euclidean space, where $X^{(t)}_{*,1} = (X^{(t)}_{1,1}, \ldots, X^{(t)}_{n^{(t)},1})'$, $N = \sum^k_{t=1} n^{(t)}$ denotes the total sample size, $E^{(t)}_{*,1} = (E^{(t)}_{1,1}, \ldots, E^{(t)}_{n^{(t)},1})'$ is the same as in (12) with iid components from distribution $N(0, (\omega^{(t)}_{1,1})^{-1})$, and we adopt the compact notation introduced in Section 2.2. In addition, we have the group sparsity structure for the regression coefficient vector C^0_1, which means that all but at most s subvectors $C^0_{1(l)} \in \mathbb{R}^k$ are zero with $C^0_{1(l)}$ and s defined in (7) and (14), respectively.

The joint group structure and sparsity structure in the multi-response linear regression model (29) naturally motivate us to exploit some variant of the group Lasso method (Yuan and Lin 2006) to estimate the coefficient vector C^0_1. The asymptotic properties of the standard group Lasso are well understood and imply faster rates of convergence in estimating C^0_1 and $\mathbf{X}^0_{*,-1} C^0_1$, compared to the standard Lasso approach (Tibshirani 1996). See, for instance, Huang and Zhang (2010) and Lounici et al. (2011) for more details as well as Uematsu et al. (2017) for more flexible high-dimensional multi-response regression. The optimal choice of an important tuning parameter, the regularization parameter $\lambda \geq 0$, in these methods, however, depends critically on the common noise level σ and is thus typically unknown in practice. Hence, one needs a practical and data-driven choice of λ that can lead to optimal estimation. This important issue has been investigated recently in Bunea, Lederer, and She (2014) and Mitra and Zhang (2016) by extending the tuning-free methods of the square-root Lasso (Belloni, Chernozhukov, and Wang 2011) and the scaled Lasso (Sun and Zhang 2012) to the group setting, respectively.

Yet the aforementioned existing tuning-free approaches in the standard group Lasso setting are not applicable to the model setting (29), which is due to the heterogeneity of the noise level in our model. Indeed, instead of a common noise level for all components of the error vector $E^0_{*,1} = (E^{(1)\prime}_{*,1}, \ldots, E^{(k)\prime}_{*,1})'$, we allow each class to have its own noise level, say, $(\omega^{(t)}_{1,1})^{-1}$ for $1 \leq t \leq k$. The strategy used in the square-root Lasso and the scaled Lasso, which essentially includes an additional parameter for the noise level, can handle only the homogeneous noises. To deal with such heterogeneity, we extend the group square-root Lasso one step further to allow for heterogeneous noises. We would like to point out that such extension for achieving the tuning-free property is generally never trivial, and the novelty of our analysis is due to an intrinsic constant level upper bound obtained on the fitted residual level for each class; see Lemma D.7 in Section D.7 of the supplementary material for more details. Liu, Wang, and Zhao (2015) also considered heterogeneous noises but the proposed method cannot be readily applied in our model.

To ease the presentation, we first introduce some notation. Define a function $Q_t(\beta^{(t)}) = \|X^{(t)}_{*,1} - \mathbf{X}^{(t)}_{*,-1}\beta^{(t)}\|^2/n^{(0)}$ with $\beta^{(t)} = (\beta^{(t)}_2, \ldots, \beta^{(t)}_p)' \in \mathbb{R}^{p-1}$ matching the index set of $C^{(t)}_1$ and $1 \le t \le k$. Denote by $\beta^0 = (\beta^{(1)\prime}, \ldots, \beta^{(k)\prime})'$ a $(p-1)k$-dimensional vector and $\beta^0_{(l)} = (\beta^{(1)}_l, \ldots, \beta^{(k)}_l)' \in \mathbb{R}^k$ the lth group of β^0 with $1 \le l \le p$ in the same way as we defined $C^0_{1(l)}$ in (7). We further introduce a diagonal matrix $\bar{D}^{(t)}_1 = \mathrm{diag}(\mathbf{X}^{(t)\prime}_{*,-1}\mathbf{X}^{(t)}_{*,-1}/n^{(t)})$ of order $p-1$ and then put them together to form a new diagonal scaling matrix $\bar{D}_1$ of order $(p-1)k$, with the submatrix of $\bar{D}_1$ corresponding to the lth group denoted by $\bar{D}_{1(l)}$ and the tth entry on the diagonal of $\bar{D}_{1(l)}$ given by $\mathbf{X}^{(t)\prime}_{*,l}\mathbf{X}^{(t)}_{*,l}/n^{(t)}$.

Our new approach of the heterogeneous group square-root Lasso (HGSL) is defined as the one given by the following optimization problem

$$\hat{C}^0_1 = \arg\min_{\beta^0 \in \mathbb{R}^{(p-1)k}} \left\{ \sum_{t=1}^{k} Q^{1/2}_t(\beta^{(t)}) + \lambda \sum_{l=2}^{p} \left\| \bar{D}^{1/2}_{1(l)} \beta^0_{(l)} \right\| \right\}, \tag{30}$$

where the regularization parameter $\lambda > 0$ which is chosen to be independent of the noise levels $(\omega^{(t)}_{1,1})^{-1}$ for $1 \le t \le k$ will be provided explicitly later. Clearly, our HGSL procedure defined in (30) is a convex program and yields an estimator for the $(p-1)k$-dimensional regression coefficient vectors C^0_1. For the estimation of general C^0_j with $1 \le j \le p$, one can simply replace the corresponding subscript 1 by j in the above method (30). The optimization problem in (30) coincides with the standard square-root Lasso in Belloni, Chernozhukov, and Wang (2011) for the case of $k = 1$, and differs from the standard group square-root Lasso in Bunea, Lederer, and She (2014) which is defined with the loss function $(\sum_{t=1}^k Q_t(\beta^{(t)}))^{1/2}$ in place of ours $\sum_{t=1}^k Q^{1/2}_t(\beta^{(t)})$ when $k \ge 2$. Without such new formulation, the standard group square-root Lasso, however, cannot carry over to take into account the heterogeneity issue when the noise level varies across different classes.

As revealed in the analysis of Theorem 3.1 to be presented, a key ingredient for the success of our HGSL estimators is an event $\mathcal{B}_1$ defined as

$$\mathcal{B}_1 = \left\{ \frac{\max_{2 \le l \le p} \left\| \bar{D}^{-1/2}_{E1} \bar{D}^{-1/2}_{1(l)} \mathbf{X}^{0\prime}_{*,(l)} E^0_{*,1} \right\|}{\sqrt{n^{(0)}}} \le \lambda \frac{\xi - 1}{\xi + 1} \right\} \tag{31}$$

for any fixed scalar $\xi > 1$, where $\mathbf{X}^0_{*,(l)}$ is an $N \times k$ submatrix of $\mathbf{X}^0_{*,-1}$ given by columns corresponding to the lth group and $\bar{D}_{E1}$ is a $k \times k$ diagonal matrix with tth diagonal entry the squared ℓ_2 norm of the error vector $E^{(t)}_{*,1}$, that is, $(\bar{D}_{E1})_{t,t} = \|E^{(t)}_{*,1}\|^2$ for $1 \le t \le k$. Similarly, we can define the event $\mathcal{B}_j$ as in (31) for each node $1 \le j \le p$. Each event $\mathcal{B}_j$ represents the one that the pure noise incurred is dominated by the penalty level. To ensure that event $\mathcal{B}_j$ holds with high probability, we need to carefully pick a sharp choice of the regularization parameter λ, that is, free of the heterogeneous noise levels.

Theorem 3.1. Assume that Conditions 2.1–2.2 hold, $s \le C_\xi n^{(0)}/\log p$ for some constant $C_\xi > 0$, and let $\hat{C}^0_j$ be the solution as in (30) for $1 \le j \le p$ with $\lambda = \frac{\xi+1}{\xi-1}\left[\frac{k + 2\delta \log p + 2\sqrt{\delta k \log p}}{n^{(0)}(1-\tau)}\right]^{1/2}$, $\tau^2 = 8(\delta \log p + \log k)/n^{(0)} = o(1)$, and $\delta > 1$ some constant. Then the event $\mathcal{B}_j$ holds with probability at least $1 - 3p^{1-\delta}$, and it holds with probability at least $1 - 4p^{1-\delta}$ that

$$\sum_{1 \le l \le p, l \ne j} \frac{1}{\sqrt{k}} \left\| \hat{C}^0_{j(l)} - C^0_{j(l)} \right\| \le Cs \left[\frac{1 + (\log p)/k}{n^{(0)}} \right]^{1/2}, \tag{32}$$

$$\frac{1}{\sqrt{k}} \left\| \hat{C}^0_j - C^0_j \right\| \le C \left[s \frac{1 + (\log p)/k}{n^{(0)}} \right]^{1/2}, \tag{33}$$

$$\frac{1}{k} \sum_{t=1}^{k} \frac{\left\| \mathbf{X}^{(t)}_{*,-1} \left(\hat{C}^{(t)}_j - C^{(t)}_j \right) \right\|^2}{n^{(0)}} \le Cs \frac{1 + (\log p)/k}{n^{(0)}}, \tag{34}$$

where $C > 0$ is some constant.

Theorem 3.1 establishes the estimation and prediction bounds for our HGSL estimators. The novelty of our technical analysis comes from an intrinsic upper bound on the fitted residual level for each class. It is worth mentioning that with the knowledge of such quantity, we can also apply the regular group Lasso with a tuning parameter depending on this quantity and obtain a corresponding justifiable theorem. The intrinsic upper bound in our analysis, however, does not appear in the HGSL optimization problem in (30) and provides only theoretical support, while the regular group Lasso implemented in the above way has to apply it in the tuning parameter explicitly. Consequently, this possibly loose intrinsic upper bound can yield large bias for the regular group Lasso, but still sharp results for our HGSL method; see the proofs of Theorem 3.1 and Lemma D.7 in Sections C.5 and D.7 of the supplementary material, respectively, for more details.

Let us gain some further insights into our tuning-free HGSL method by comparing the sharpness of our regularization parameter λ specified in Theorem 3.1 with the one used by Bunea, Lederer, and She (2014) for the setting of homogeneous noises. One advantage of our choice of λ comes from the use of the scaling matrix $\bar{D}_1$, which makes the noise per column of $\mathbf{X}^0_{*,(l)}$ homogeneous and sharpens λ by a factor given by the ratio of the largest and the smallest ℓ_2 norms among all columns. Moreover, thanks to the simple block diagonal structure of matrices $\mathbf{X}^0_{*,(l)}$ a direct and sharp chi-square tail probability (Laurent and Massart 2000) provides us sharper constant factors for both k and $\log p$.

As demonstrated in Theorem 3.1, the tuning parameter λ can be calculated theoretically by applying the formula therein with some small and fixed constants ξ and δ, for example, $\xi = \delta = 1.001$. Our empirical studies show that HFSL estimators are not sensitive to ξ and δ as long as they are not chosen too large. Theorem 3.1 guarantees that with such pre-calculated λ, HGSL estimator enjoys the nice properties as described in (32)– (34). Therefore, in this sense, our method is truly tuning-free. This is in fact also the major distinction from many existing methods in the literature, which depend on some tuning parameters that need to be chosen adaptively using training data by, for example, cross-validation.

The theoretical choice of parameter λ for HGSL established in Theorem 3.1 has been justified to yield optimal conver-

gence rates as sample size goes to infinity. With finite sample, however, such λ may not yield the best results. Next, we introduce a simulation strategy to choose λ which can adapt automatically with sample size. We first simulate the value of $\|\bar{D}_{E1}^{-1/2}\bar{D}_{1(2)}^{-1/2}\mathbf{X}_{*,(2)}^{0\prime}E_{*,1}^{0}\|/(n^{(0)})^{1/2}$ for 10,000 times and pick the $100(1-1/p^{\delta})$th percentile of its empirical distribution as our choice of $\lambda(\xi-1)/(\xi+1)$ with some constant $\delta > 1$. Here, we take $\delta > 1$ because of the union bound argument given that only the setting of $l = 2$ is simulated. It is important to note that the components of $\bar{D}_{E1}^{-1/2}\bar{D}_{1(2)}^{-1/2}\mathbf{X}_{*,(2)}^{0\prime}E_{*,1}^{0}$ are independent and their distributions can be characterized easily since they do not depend on the variances of $\mathbf{X}_{*,(2)}^{0\prime}$ and $E_{*,1}^{0}$. More specifically, for each replication $1 \le T \le 10{,}000$ we simulate the tth component of $\bar{D}_{E1}^{-1/2}\bar{D}_{1(2)}^{-1/2}\mathbf{X}_{*,(2)}^{0\prime}E_{*,1}^{0}$ independently by first generating $Z_{1,t,T}, Z_{2,t,T} \sim N(0,I) \in \mathbb{R}^{n^{(t)}}$ independently and then calculating $Z_{t,T} = (n^{(t)})^{1/2}Z_{1,t,T}^{\prime}Z_{2,t,T}/(\|Z_{1,t,T}\|\|Z_{1,t,T}\|)^{1/2}$. The simulated value of $\|\bar{D}_{E1}^{-1/2}\bar{D}_{1(2)}^{-1/2}\mathbf{X}_{*,(2)}^{0\prime}E_{*,1}^{0}\|$ can then be written as $(\sum_{t=1}^{k}Z_{t,T}^{2})^{1/2}$. Thus, our simulation strategy provides a specific choice of the parameter λ given by

$$\lambda_{sim} = \frac{1}{\sqrt{n^{(0)}}}\frac{\xi+1}{\xi-1}\inf\left\{\nu : \sum_{T=1}^{10{,}000} 1\left\{\Big(\sum_{t=1}^{k}Z_{t,T}^{2}\Big)^{1/2} < \nu\right\} /10{,}000 \ge 1-1/p^{\delta}\right\}. \tag{35}$$

We will further discuss the choices of δ and ξ in Section 4.1 when implementing our proposed procedure THI with the HGSL.

In our simulation studies, we compared the aforementioned two methods for choosing λ—the theoretical-based and the simulation-based ones. The results are similar with the latter slightly outperforms the former because of the finite sample effects. For this reason, all our numerical analyses use the simulation-based choice of λ.

4. Numerical Studies

4.1. Simulation Studies

We now proceed with investigating the finite-sample performance of THI with the chi-based test ϕ_2 and the linear functional-based test ϕ_1, which are referred to as procedures THI-ϕ_2 and THI-ϕ_1, respectively. In particular, Section 4.1.1 presents the hypothesis testing results of our methods. As discussed in Sections 1 and 2.5, the existing methods on multiple graphs have focused on the estimation problem instead of statistical inference such as hypothesis testing. As such, we modify our procedures correspondingly to obtain estimates for the precision matrix and then compare them with some popularly used approaches such as the MPE (Cai et al. 2016) and the GGL and FGL (Danaher, Wang, and Witten 2014) in Section 4.1.2. Section 4.1.3 further examines the robustness of our methods in the presence of heavy-tailed distributions.

We consider two different model settings, Models I and II, for generating the k networks with Gaussian graphical models given by precision matrices $\Omega^{(t)} = (\omega_{a,b}^{(t)})$ with $1 \le t \le k$. In both models, the block diagonal structure is used to introduce sparsity in the precision matrices in the sense that all the entries outside the diagonal blocks are equal to zero. More specifically, our Model I assumes that all k precision matrices share the same block diagonal structure and all diagonal blocks have the same size. For each pair (a,b) with $1 \le a \ne b \le p$, if the (a,b)th entry belongs to a diagonal block, then we draw the values for $\omega_{a,b}^{(1)},\ldots,\omega_{a,b}^{(k)}$ independently from the uniform distribution $U[0.2,0.4]$ or $U[0.6,1.2]$, depending on whether it belongs to the upper half diagonal blocks or the lower half diagonal blocks, respectively. All the off-diagonal entries within the diagonal blocks are generated independently. Finally we set the diagonal entries as 1 for the upper half diagonal blocks and 3 for the lower half ones. Observe that in Model I, each joint link strength vector $\omega_{a,b}^{0} = (\omega_{a,b}^{(1)},\ldots,\omega_{a,b}^{(k)})^{\prime}$ with $a \ne b$ is either a zero vector or of k nonzero components.

To make the sparsity pattern more flexible compared to Model I, our Model II employs a different data generating scheme for entries inside the diagonal blocks with the rest of the setting the same as in Model I. Specifically, for each entry (a,b) with $a \ne b$ inside a diagonal block we first flip a fair coin. If it is heads, then the joint link strength vector $\omega_{a,b}^{0}$ is generated in the same way as in Model I. If it is tail, we randomly draw an integer k_0 from the uniform distribution over $\{1,\ldots,k\}$, and then set $\omega_{a,b}^{(t)} = 0$ for each $1 \le t \ne k_0 \le k$ and generate $\omega_{a,b}^{(k_0)}$ from the uniform distribution $U[0.2,0.4]$ or $U[0.6,1.2]$, depending on whether the pair (a,b) falls in the upper half diagonal blocks or the lower half diagonal blocks, respectively. Clearly, Model II is sparser than Model I.

For each of the two models introduced above, we further consider three different settings of parameters by varying the number of networks k and the number of nodes p, while fixing the sample sizes $n^{(t)} = n^{(0)}$ at 100 for Model I and at 200 for Model II with $1 \le t \le k$. We also fix the block size to be 8 and set the number of repetitions as 100 in each setting. The tuning-free regularization parameter λ is chosen as λ_{sim} in (35) using our simulation strategy with $\delta = 1$ and $\xi = \infty$. Alternatively one can also use the choice of parameter λ given in Theorem 3.1, which results in similar but slightly worse performance compared to the use of λ_{sim}.

4.1.1. Testing Results

To see how our proposed methods THI-ϕ_2 and THI-ϕ_1 perform in finite samples, let us start with the hypothesis testing results in Models I and II. For each simulated dataset, we apply the THI procedure with the chi-based test ϕ_2 and the linear functional-based test ϕ_1 with sign vector $\xi = (1,\ldots,1)^{\prime}$ to each pair of nodes (a,b) with $a \ne b$ to detect whether some edges exist between nodes a and b for any of the k networks. We set the significance level α to be 0.05 and employ two different methods to calculate the critical values. The first method computes the critical values using the asymptotic null distributions established in Theorems 2.1 and 2.2, with the corresponding critical values named as "Theoretical" in Tables 1 and 2. The second method, called "Empirical" in Tables 1 and 2, computes the critical values empirically based on the values of the test statistic $U_{n,k,a,b}$ for the chi-based test ϕ_2, or the test statistic $V_{n,k,a,b}(\xi)$ for the linear functional-based test ϕ_1, for the entries outside the diagonal blocks. Since the entries outside the diagonal blocks are all equal to zero across the k networks, the 5% critical value can

Table 1. Means and SD (in parentheses) of testing results for THI methods in Model I with $\alpha = 0.05$.

Method		k	p	FNR ($\times 10^{-2}$) Empirical	FNR ($\times 10^{-2}$) Theoretical	FPR ($\times 10^{-2}$)	ROC area ($\times 10^{-2}$)
THI-ϕ_1	Setting 1	5	50	0.375 (0.484)	0.369 (0.454)	5.044 (0.656)	99.90 (0.078)
	Setting 2	10	50	0 (0)	0 (0)	4.945 (0.752)	1 (0)
	Setting 3	10	200	0.001 (0.014)	0.001 (0.014)	5.005 (0.170)	1 (0)
THI-ϕ_2	Setting 1	5	50	3.268 (1.568)	3.161 (1.422)	5.123 (0.722)	99.26 (0.319)
	Setting 2	10	50	0.006 (0.060)	0.006 (0.060)	5.352 (0.751)	1 (0.010)
	Setting 3	10	200	0.077 (0.100)	0.077 (0.098)	4.896 (0.177)	99.97 (0.019)

Table 2. Means and SD (in parentheses) of testing results for THI methods in Model II with $\alpha = 0.05$.

Method		k	p	FNR ($\times 10^{0}$) Empirical	FNR ($\times 10^{0}$) Theoretical	FPR ($\times 10^{-2}$)	ROC area ($\times 10^{-2}$)
THI-ϕ_1	Setting 1	5	50	0.226 (0.043)	0.224 (0.038)	5.151 (0.821)	94.54 (1.346)
	Setting 2	10	50	0.327 (0.041)	0.327 (0.038)	5.046 (0.932)	90.26 (2.07)
	Setting 3	10	200	0.306 (0.017)	0.305 (0.016)	5.04 (0.233)	91.12 (0.771)
THI-ϕ_2	Setting 1	5	50	0.066 (0.019)	0.064 (0.017)	5.125 (0.747)	98.42 (0.520)
	Setting 2	10	50	0.099 (0.021)	0.094 (0.020)	5.416 (0.750)	97.66 (0.560)
	Setting 3	10	200	0.090 (0.010)	0.090 (0.010)	5.017 (0.149)	97.79 (0.302)

be calculated as the 95th percentile of the pooled test statistics for all such null entries.

It is worth pointing out that the "Empirical" critical value mentioned above relies on the knowledge of true nulls and thus can only be calculated in simulation studies. The main purpose of using both methods for determining the critical values is to compare the "Theoretical" values with the "Empirical" ones to justify our findings on the null distributions of our tests ϕ_2 and ϕ_1 in Theorems 2.1 and 2.2, respectively. With these critical values, we can calculate the false positive rate (FPR) and the false negative rate (FNR). Clearly, with the "Empirical" critical value the FPR should be exactly 5%, and thus we omit its values and include only the FPR based on the "Theoretical" critical value in Tables 1 and 2, which present the means and SD of testing results in Models I and II, respectively. The FNRs based on both critical values are reported. In fact, we see from Tables 1 and 2 that the "Theoretical" values for both FPR and FNR are very close to the "Empirical" ones, indicating that the asymptotic null distributions obtained in Theorems 2.1 and 2.2 indeed match the empirical distributions very closely. To better evaluate these methods, we also vary the critical value and generate a full receiver operating characteristic (ROC) curve. The areas under the ROC curves are summarized in Tables 1 and 2. It is seen that both methods THI-ϕ_2 and THI-ϕ_1 have areas under the ROC curve close to 1 across all settings.

In particular, we see from Table 1 that the linear functional-based test ϕ_1 is significantly better than the chi-based test ϕ_2 over all settings of Model I. From setting 1 to setting 2, both testing procedures become better, while both procedures perform worse from setting 2 to setting 3. These are consistent with our theoretical results. To understand this, let us take the entry $(1, 2)$ as an example. In view of Theorem 2.3, the separating rate for alternative $H^{l1}_{1,12}(s, \epsilon, \xi)$ with the corresponding optimal test ϕ_1 is $\|\omega^0_{1,2}\|_1 \geq \epsilon_n \asymp \sqrt{k/n^{(0)}}$. Since the components of the joint link strength vector $\omega^0_{1,2}$ are iid from the uniform distribution $U[0.2, 0.4]$, as the number of networks k increases the separating rate condition becomes weaker because $\|\omega^0_{1,2}\|_1$ grows linearly with k, while the right-hand side $\epsilon_n \asymp \sqrt{k/n^{(0)}}$ grows at a slower rate of $\sqrt{k}$. Thus, the growth of k makes the separating rate condition easier to be satisfied. The results for the chi-based test ϕ_2 can be understood similarly.

Comparing Table 2 with Table 1, we see that the performance of both testing procedures ϕ_2 and ϕ_1 becomes worse. This is reasonable since Model II is sparser than Model I and thus the separating rate conditions indicated in Theorem 2.3 are harder to be satisfied for these sparser entries with only one nonzero component across k networks, because this nonzero entry needs to have magnitude much larger than $\epsilon_n \asymp \sqrt{k/n^{(0)}}$ for test ϕ_1 or $\epsilon_n \asymp \sqrt{k^{1/2}/n^{(0)}}$ for test ϕ_2. As a consequence, different from Table 1 in which the separating rate conditions become easier for denser entries with all k nonzero components as k increases, these conditions become more stringent for sparser entries with only one nonzero component as k increases. Such increased difficulty for sparser entries is more severe for the linear functional-based test ϕ_1 than for the chi-based test ϕ_2 in light of the separating rates ϵ_n in Theorem 2.3.

4.1.2. *Precision Matrix Estimation*

As mentioned before, almost all existing methods on multiple graphs focus on the estimation part. To compare with these existing methods, we modify our THI procedure to generate sparse estimates of the precision matrices. Specifically, we suggest a two-step procedure. In the first step, for each entry (a, b) with $a \neq b$, we conduct hypothesis testing at significance level α to see whether the null hypothesis $H_{0,ab}$ in (1) is rejected or not. The critical values at significance level α are calculated using the asymptotic distributions established in Theorems 2.1 and 2.2. In the second step, for each $1 \leq a \leq p$ we estimate the (a, a)th entry of the tth graph as $\hat{\omega}^{(t)}_{a,a}$, and for each rejected null hypothesis $H_{0,ab}$ we estimate the (a, b)th entry of the tth graph as $-\hat{\omega}^{(t)}_{a,a}\hat{\omega}^{(t)}_{b,b}T^{(t)}_{n,k,a,b}$ in view of (10), where all the notation is the same as in Section 2.2.

In our two-step procedure suggested above, there is one tuning parameter which is the significance level α. To tune such parameter, we generate an independent validation set with the

same sample sizes $n^{(t)} = n^{(0)} = 100$ for Model I and 200 for Model II with $1 \le t \le k$. Then for each given value of α, we obtain a set of sparse precision matrix estimates $\hat{\Omega}^0 = (\hat{\Omega}^{(1)}, \ldots, \hat{\Omega}^{(k)})$ for the k graphs using the training data, and calculate the value of the loss function

$$L(\hat{\Omega}^0) = \sum_{t=1}^{k} \left\{ \log[\det(\hat{\Omega}^{(t)})] - \mathrm{tr}(\hat{\Sigma}^{(t)}\hat{\Omega}^{(t)}) \right\}, \qquad (36)$$

where $\hat{\Sigma}^{(1)}, \ldots, \hat{\Sigma}^{(k)}$ are the sample covariance matrix estimators for the k graphs constructed based on the validation data. The parameter α is then chosen by minimizing the loss function in (36) over a grid of 10 values for α. We compare our THI approach with three commonly used competitor methods MPE, GGL, and FGL, each with one regularization parameter to tune. For a fair comparison, for each method we use the same validation set to tune the regularization parameter and choose the one minimizing the loss function in (36) over a grid of 10 values.

To evaluate the performance of different methods, we calculate three loss functions of the matrix 1-norm, the spectral norm, and the Frobenius norm for the estimation errors, which are denoted as ℓ_1, ℓ_2, and ℓ_F, respectively. The precision matrix estimation results for different methods in Models I and II are summarized in Tables 3 and 4, respectively. In particular, for setting 3 of both models the results of MPE and FGL are not reported because the results cannot be obtained within a reasonable amount of time due to their excessively high computational costs. To gain insights into the computational costs of various methods, we record in Table 7 in the supplementary material the average computational cost measured as the CPU time in seconds for each method.

We see from Table 3 that across all three settings, both methods THI-ϕ_2 and THI-ϕ_1 outperform the MPE, FGL, and GGL significantly. Similar phenomenon can be observed from Table 4. In light of the computational cost presented in Table 7 in supplementary material, the overall performance of our methods is superior to that of all three competing methods. Observe that setting 1 differs from setting 2 only in the number of networks k. Therefore, it is fair to conclude that compared to other approaches, our methods have greater advantages in estimating a large number of graphs simultaneously, which is in line with our theoretical findings that our methods allow the number of networks k to diverge with the sample size $n^{(0)}$ at a faster rate.

Table 3. Means and SD (in parentheses) of precision matrix estimation results for different methods in Model I.

	k	p	Method	ℓ_1	ℓ_2	ℓ_F
Setting 1	5	50	THI-ϕ_1	4.968 (0.041)	3.417 (0.036)	6.657 (0.036)
			THI-ϕ_2	5.68 (0.070)	3.894 (0.081)	7.578 (0.131)
			MPE	7.556 (0.024)	6.347 (0.056)	11.53 (0.083)
			GGL	8.331 (0.009)	7.289 (0.005)	13.05 (0.005)
			FGL	7.989 (0.046)	7.247 (0.044)	13.13 (0.069)
Setting 2	10	50	THI-ϕ_1	5.117 (0.102)	3.281 (0.103)	6.416 (0.194)
			THI-ϕ_2	5.191 (0.104)	3.333 (0.108)	6.542 (0.202)
			MPE	7.075 (0.022)	5.618 (0.048)	10.44 (0.070)
			GGL	8.193 (0.006)	7.241 (0.005)	12.98 (0.010)
			FGL	8.132 (0.003)	7.461 (0.003)	13.36 (0.004)
Setting 3	10	200	THI-ϕ_1	5.84 (0.096)	3.997 (0.116)	14.3 (0.474)
			THI-ϕ_2	6.466 (0.111)	4.674 (0.142)	16.79 (0.594)
			MPE	–	–	–
			GGL	8.467 (0.006)	7.489 (0.003)	27.01 (0.003)
			FGL	–	–	–

Table 4. Means and SD (in parentheses) of precision matrix estimation results for different methods in Model II.

	k	p	Method	ℓ_1	ℓ_2	ℓ_F
Setting 1	5	50	THI-ϕ_1	3.651 (0.035)	2.091 (0.018)	4.723 (0.023)
			THI-ϕ_2	3.368 (0.045)	2.042 (0.023)	4.392 (0.043)
			MPE	4.909 (0.020)	3.289 (0.015)	6.668 (0.018)
			GGL	7.087 (0.009)	5.155 (0.004)	9.653 (0.005)
			FGL	6.748 (0.007)	4.942 (0.004)	9.563 (0.006)
Setting 2	10	50	THI-ϕ_1	3.095 (0.018)	1.898 (0.009)	4.213 (0.011)
			THI-ϕ_2	3.019 (0.020)	1.878 (0.011)	4.099 (0.013)
			MPE	3.613 (0.013)	2.264 (0.010)	4.325 (0.014)
			GGL	5.708 (0.006)	4.325 (0.003)	8.238 (0.004)
			FGL	5.606 (0.005)	4.27 (0.003)	8.228 (0.004)
Setting 3	10	200	THI-ϕ_1	6.035 (0.077)	2.7 (0.018)	11.18 (0.078)
			THI-ϕ_2	5.595 (0.085)	3.448 (0.061)	15.19 (0.306)
			MPE	–	–	–
			GGL	6.976 (0.005)	5.195 (0.004)	18.23 (0.004)
			FGL	–	–	–

4.1.3. Heavy-Tailed Distributions

Model misspecification (Cule, Samworth, and Stewart 2010) can often occur in applications. Thus, it is important to examine the robustness of proposed methods. With this in mind, we now investigate the finite-sample performance of our THI procedure in the presence of heavy-tailed distributions such as the Laplace distribution, as opposed to the Gaussianity assumed in our theoretical developments. For each previous setting in Models I and II, after generating the precision matrix $\Omega^{(t)}$, instead of sampling the data matrix $\mathbf{X}^{(t)}$ from the Gaussian distribution with mean zero and covariance matrix $(\Omega^{(t)})^{-1}$ we draw $\mathbf{X}^{(t)}$ from the multivariate Laplace distribution with covariance matrix $(\Omega^{(t)})^{-1}$. More specifically, we first generate a random vector whose components are iid Laplace random variables with location parameter zero and scale parameter $1/\sqrt{2}$, and then multiply this vector by $(\Omega^{(t)})^{-1/2}$ to obtain the desired Laplace random vector. All the rest of the settings are the same as before.

Table 5 presents the testing results of our methods THI-ϕ_2 and THI-ϕ_1 in the setting of heavy-tailedness. Compared to the results in Tables 1 and 2, we observe that across all settings of Models I and II, the performance of our methods stays almost the same when the Gaussian distribution is replaced by the Laplace distribution, demonstrating the robustness of our methods to the heavy-tailed distributions. We have also explored other heavy-tailed distributions such as the t-distribution with 5 degrees of freedom and the results are very similar. To save the space, these additional results are not presented here but are available upon request.

4.2. Real Data Analysis

In this section, we demonstrate the performance of our methods using three microarray datasets on triple-negative breast cancer. The three datasets come from separate studies on the same set of genes for three different groups of cancer patients. Direct merging of the datasets is usually less favored due to the inherent discrepancy among the studies. Thus, we expect that the underlying sparsity structure is the same but the nonzero

Table 5. Means and SD (in parentheses) of testing results for THI methods in Models I and II with the Laplace distribution and $\alpha = 0.05$.

Model I							
				FNR($\times 10^{-2}$)		FPR	ROC Area
Method		k	p	Empirical	Theoretical	($\times 10^{-2}$)	($\times 10^{-2}$)
THI-ϕ_1	Setting 1	5	50	0.345 (0.480)	0.357 (0.440)	4.986 (0.723)	99.91 (0.068)
	Setting 2	10	50	0 (0)	0 (0)	5.089 (0.991)	100 (0)
	Setting 3	10	200	0 (0)	0 (0)	5.03 (0.172)	100 (0)
THI-ϕ_2	Setting 1	5	50	3.012 (1.555)	2.810 (1.438)	5.293 (0.669)	99.32 (0.287)
	Setting 2	10	50	0 (0)	0 (0)	5.701 (0.824)	100 (0.004)
	Setting 3	10	200	0.066 (0.094)	0.063 (0.094)	5.073 (0.171)	99.98 (0.016)
Model II							
				FNR ($\times 10^{0}$)		FPR	ROC Area
Method		k	p	Empirical	Theoretical	($\times 10^{-2}$)	($\times 10^{-2}$)
THI-ϕ_1	Setting 1	5	50	0.226 (3.594)	0.226 (3.414)	5.046 (0.973)	94.49 (1.311)
	Setting 2	10	50	0.317 (3.765)	0.319 (3.497)	5.011 (0.908)	90.88 (1.806)
	Setting 3	10	200	0.309 (1.574)	0.308 (1.567)	5.048 (0.219)	91.03 (0.766)
THI-ϕ_2	Setting 1	5	50	0.069 (0.020)	0.066 (0.019)	5.388 (0.854)	98.43 (0.512)
	Setting 2	10	50	0.093 (0.020)	0.090 (0.019)	5.375 (0.725)	97.66 (0.629)
	Setting 3	10	200	0.089 (0.010)	0.088 (0.010)	5.083 (0.177)	97.83 (0.320)

link strengths may vary across these groups. The same dataset has been analyzed in Ma, Ren, and Tseng (2018). Following the procedure therein, we pre-process the data and after the initial filtering, we have 275, 178, and 165 samples in each dataset, respectively, where for each observation 3377 genes are retained. To further reduce the dimensionality, we rank the genes according to the sum of variances across the three datasets and keep only the top $p = 150$ ones with largest variations for our analysis.

Since our procedure is designed to analyze a large number of graphs, to demonstrate the full advantage of our method, we further randomly split each of the three datasets into two subsets of approximately equal sample size. Thus, we end up with $k = 6$ subgroups of sample size 137, 138, 89, 89, 83, and 82, respectively, where each observation is a vector with dimension $p = 150$. It is seen that the first two subgroups should have identical graphical structure, so are the middle two subgroups and the last two subgroups.

Since for this particular dataset, we have no additional information on the signs of $\omega_{a,b}^{(k)}$ across subgroups, we only apply our proposed Chi-based test ϕ_2 to these $k = 6$ subgroups. As discussed in Section 2.5, our method differs from most existing ones in that it can produce p-values for testing the significance of the connectivity of nodes. Since there are $p(p-1)/2$ pair of nodes, we indeed face the problem of large-scale multiple comparisons. Thanks to the availability of p-values, we adopt the procedure in Benjamini and Hochberg (1995) to achieve FDR control at some target level q. This gives us a sparse graphical model where two nodes are connected if they are connected in any of the k graphs at the prespecified level q of FDR. Note that for each split of the data, we can produce one such graph. To account for the randomness caused by data split, we repeat the entire procedure 100 times, and at the end, we aggregate the results by retaining edges that only appear more than 70% of time.

For comparison, we also apply our chi-based test ϕ_2 to the $k = 3$ original subgroups and produce a graph at the same target level of FDR. This is done only once because there is no random data split involved. It is intuitive that the results from $k = 3$ original groups should be more accurate because it relies on larger sample sizes. In addition, we also report the graph returned by correlation network using the original data from three subgroups. That is, for each of these $k = 3$ subgroups, we calculate the sample correlation matrix ($\rho_{a,b}^{(k)}$), and apply the Fisher transformation to each correlation coefficient $\rho_{a,b}^{(k)}$ to make it close to normal distribution. Then for each pair of nodes (a, b), under the null hypothesis $\tilde{H}_{0,ab} : \rho_{a,b}^{(1)} = \cdots = \rho_{a,b}^{(k)} = 0$, the squared summation of the transformed correlation coefficients across three groups should be approximately χ_3^2 distributed. Thus, the p-value for testing each $\tilde{H}_{0,ab}$ can be calculated and the same FDR control procedure can be applied to obtain a sparse correlation matrix.

We will compare the aforementioned three graphs: (1) the aggregated graph produced by ϕ_2 from $k = 6$ subgroups and 100 random splits, (2) the one produced by ϕ_2 from $k = 3$ original subgroups, and (3) the one produced by correlation network from $k = 3$ original subgroups. Note that for any pair of genes (a, b) with different connectivities across the three disease subgroups, the corresponding null hypothesis $H_{0,ab}$ (or $\tilde{H}_{0,ab}$) should be rejected. Thus, the three identified graphs discussed above should be able to tell us some information on which pair of genes exhibit different connectivity pattern for triple-negative breast cancer.

For performance measure, motivated by the definition of central nodes introduced in Cai et al. (2016), we calculate the degree for each node and define important nodes as the ones with largest degrees in the graphs. Table 6 lists the top 20 nodes produced by the aforementioned three graphs by setting the target FDR level at $q = 0.001$. It is seen that for the two graphs produced by ϕ_2, there are 9 overlaps out of the top 20, showing a good level of consistency. To visually compare the two graphs by ϕ_2, we also plot their connectivities. Since there are large number of nodes, to make the graphs easier to read, we reduce the FDR level to 5×10^{-5} and exclude all degree 0 nodes. The resulting graphs are summarized in Figure 1. There

Table 6. Top 20 nodes with highest degrees identified by THI-ϕ_2 in descending order.

	Corr	THI-ϕ_2	THI-ϕ_2
k	3	3	6
1	EGFR	**SOX10**	FOXC1
2	CCND1	**COL3A1**	**COL3A1**
3	BAMBI	**LUM**	SYNM
4	MT1E	**SELL**	**SPARCL1**
5	NDRG1	IGFBP7	**SRGN**
6	NQO1	**COL5A2**	**LUM**
7	SFN	CLDN3	EFHD1
8	KIT	AKR1C2	TMEM158
9	MYB	KRT14	**SOX10**
10	SCD	COL1A1	**IFI27**
11	MMP9	**IFI27**	ALDH3B2
12	LBP	**DCN**	GBP1
13	PI3	SPDEF	F13A1
14	GSTT1	CYP1B1	**SELL**
15	RBP1	ALCAM	CKS2
16	PON3	**SPARCL1**	**COL5A2**
17	MT1G	**SRGN**	CXCR4
18	KLK6	COL6A3	ELF5
19	HIST1H2BK	MLPH	SERHL2
20	SLPI	KRT7	**DCN**

NOTE: Nodes in bold highlight 9 overlaps between the two graphs produced by THI-ϕ_2.

are a few interesting findings. First, the common 9 nodes in the top 20 list shares similar connectivity pattern in the two graphs. For instance, for 'SOX10', it is connected with "ELF5" and "ALCAM" in both graphs. The cluster of genes, "COL3A1," "COL1A1," "COL5A2," "COL6A3," "GJA1," are connected in both graphs. The cluster of genes, "CXCL13," "SRGN," "RGS2," "CXCR4," are connected in both graphs. The cluster of genes "IFI27," "IFI16," "IFI44L" are connected in both graphs. And the same is true for genes "COL10A1," "CTSK," and "F13A1." Besides those common clusters corresponding to top ranked genes, we also observe the common cluster "HLA-DRA," "HLA-DQA1," "HLA-DPA1," "CD74" in both subgraphs.

We also plot the graph produced by correlation network when setting $q = 5 \times 10^{-5}$ in Figure 2. It is seen that the graph is very dense, suggesting that the correlation network is possibly not as sparse as the graphical network. Due to the very dense natural of the correlation network, interpretation is impossible.

5. Discussion

In this article, we have introduced the THI framework with the chi-based test and the linear functional-based test to detect the sparsity patterns of multiple networks in the setting of Gaussian graphical models. Such a framework is not only scalable to large scales, but also enjoys optimality properties in the scenario where the number of networks is allowed to diverge and the number of features can be much larger than the sample size. Our theoretical justifications show that under mild regularity conditions, the linear functional-based test has the minimum requirement on the sample size.

Two testing procedures in our THI framework can be extended to the sub-Gaussian distribution setting. When the p-dimensional feature vector $X^{(t)}$ in (2) jointly follows a multivariate sub-Gaussian distribution with bounded sub-Gaussian norm (see, e.g., Definition 5.22 in Vershynin 2010) in each class, the value of interest $\omega_{a,b}^{(t)}$ has a natural interpretation of partial correlation between $X_a^{(t)}$ and $X_b^{(t)}$. Under such settings, one can show that two tests are still valid asymptotically by

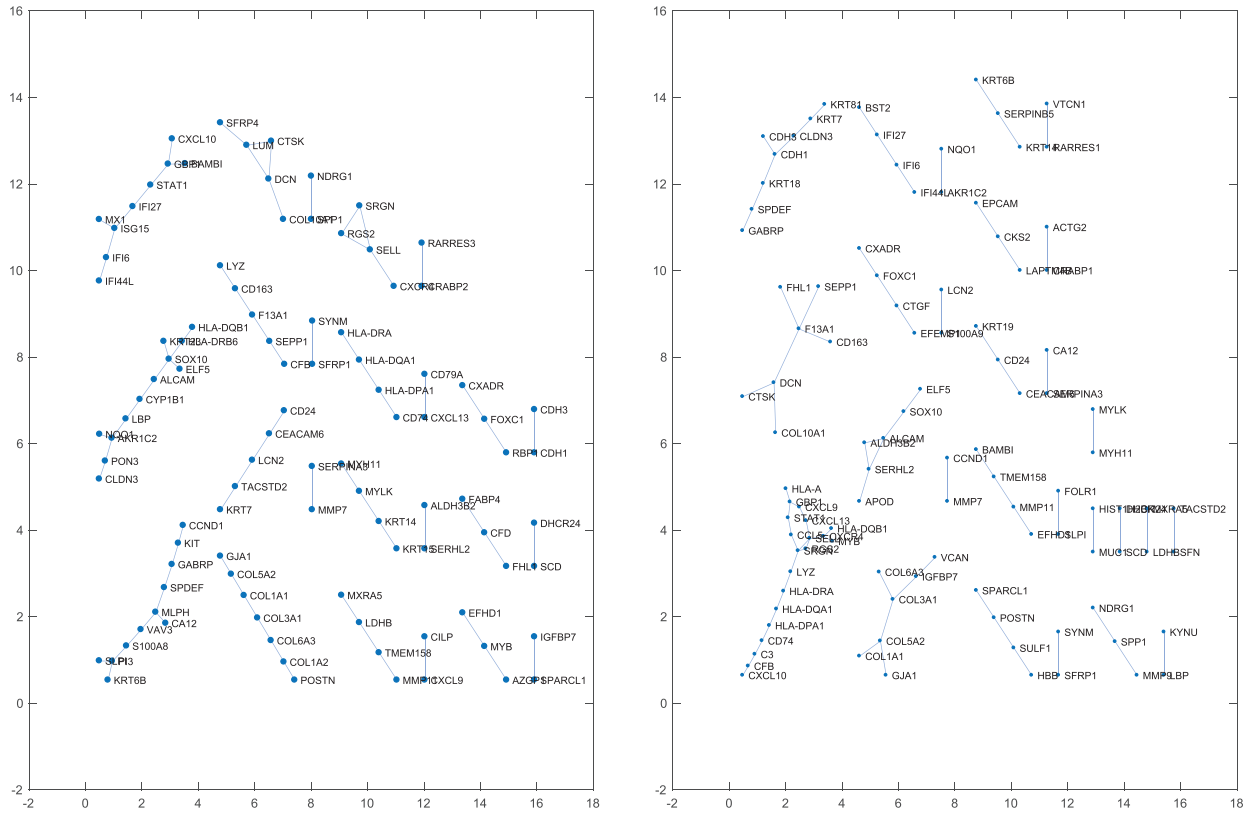

Figure 1. Common edges identified by methods THI-ϕ_2 using original 3 subgroups (left panel) and 6 subgroups with random splits (right panel).

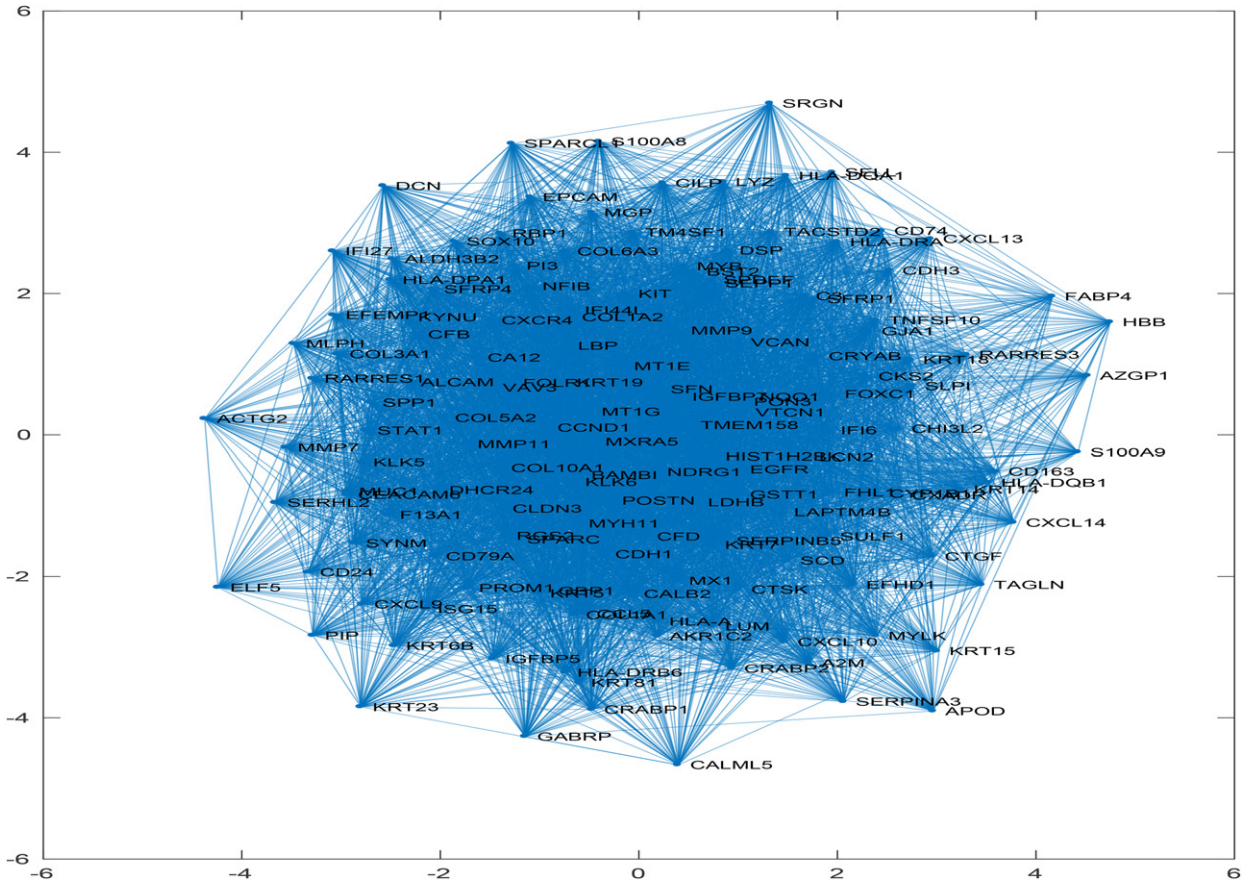

Figure 2. Graph produced by correlation network learning based on original three subgroups.

scrutinizing the proof details with a modification. Indeed, to test the null hypothesis $H_{0,ab}$, one only need to replace the quantity $\hat{\omega}_{b,b}^{(t)}\hat{\omega}_{a,a}^{(t)}$ in the chi-based test statistic $U_{n,k,a,b}$ of (11) and in the linear functional-based test statistic $V_{n,k,a,b}(\xi)$ of (19) by $1/\hat{\iota}_{a,b}^{(t)}$, where $\hat{\iota}_{a,b}^{(t)} = (\sum_{i=1}^{n^{(t)}}(\hat{E}_{i,a}^{(t)}\hat{E}_{i,b}^{(t)})^2)/n^{(t)}$ is an estimator of the variance of $E_{i,a}^{(t)}E_{i,b}^{(t)}$ under null (see (12)). For Gaussian distribution, this variance coincides with $1/(\omega_{b,b}^{(t)}\omega_{a,a}^{(t)})$ under null.

Yet the optimality of the sample size requirement for the chi-based test, that is, the minimum sample size requirement with the optimal separating rate $\epsilon_n = \sqrt{k^{1/2}/n^{(0)}}$ for testing null $H_{0,ab}$ against alternative $H_{1,ab}^{l2}$, still remains as an open problem for future investigation. The main challenges lie in the need of constructing a new lower bound as in Theorem 2.4 for alternative $H_{1,ab}^{l1}$, which involves both the sample size requirement and the separating rate. Moreover, the technical analysis in the proof of Theorem 2.1 contains a relatively loose bound between the ℓ_1 and ℓ_2 norms, which implies that the sample size requirement imposed in Proposition 2.1 may not be sharp, though sharper than that for the naive combination testing procedure discussed in Section 2.3.

As mentioned in Section 1, our article assumes common sparsity and allows two aspects of heterogeneity which are the heterogeneity in link strengths over multiple networks and the heterogeneity in noise levels over multiple subpopulations. The appealing characteristics of our THI framework for addressing these issues are empowered by our newly suggested convex approach of heterogeneous group square-root Lasso (HGSL) for the setting of high-dimensional multi-response regression with heterogeneous noises. Other aspects of heterogeneous learning and inference can certainly be interesting as well. For example, in practice one might be interested in studying whether the link strengths across different graphs are identical or not. This is a more general yet more challenging problem that deserves further study. Some efforts along this direction have been made in the literature. For instance, Danaher, Wang, and Witten (2014) proposed a penalized likelihood method using the fused Lasso to estimate the common link strength among multiple Gaussian graphs. This method, however, focuses only on the estimation of common link strength and lacks theoretical justification for its performance. Moreover, their proposed algorithm is not scalable due to the complicated form of the likelihood function. Thus, it would be interesting to extend the methods developed in our article to the problem of testing for heterogeneity in link strengths.

Our studies are only among the first attempts to address the challenging issues of heterogeneity in multiple networks inference in the setting of Gaussian graphical models. It would be interesting to extend our inferential approach to the settings of multiple matrix graphical models, multiple tensor graphical models, and multiple non-Gaussian graphical models, as well as other network models beyond graphical models. Furthermore, in some applications, it is possible that a fraction of the class labels for the subpopulations or even all the class labels can be unavailable, in which clustering techniques can play a crucial

role. In addition, there can exist some latent features which would require a broader class of network structures. The developments on heterogeneous inference in multiple networks can also motivate new approaches for regression and classification problems that have networks as an input. The possible extensions addressing these issues are beyond the scope of the current article and will be interesting topics for future research.

Supplementary Material

The online supplementary materials contain a scalable HGSL algorithm with provable convergence, the proofs of Theorems 2.1-3.1 and Propositions 2.1-2.3, as well as the proofs of key lemmas and additional technical details. Additional computational cost comparison with existing methods is also provided.

Acknowledgments

Part of this work was completed while the last two authors visited the Departments of Statistics at University of California, Berkeley and Stanford University. These authors sincerely thank both departments for their hospitality.

Funding

This work was supported by NSF Grant DMS-1812030, NIH funding: NIH Grant 1R01GM131407-01, NSF CAREER Awards DMS-0955316, and DMS-1150318, a grant from the Simons Foundation, and Adobe Data Science Research Award.

References

Baraud, Y. (2002), "Non-Asymptotic Minimax Rates of Testing in Signal Detection," *Bernoulli*, 8, 577–606. [1915]

Belloni, A., Chernozhukov, V., and Wang, L. (2011), "Square-Root Lasso: Pivotal Recovery of Sparse Signals via Conic Programming," *Biometrika*, 98, 791–806. [1912,1916,1917]

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society*, Series B, 57, 289–300. [1921]

Bunea, F., Lederer, J., and She, Y. (2014), "The Group Square-Root Lasso: Theoretical Properties and Fast Algorithms," *IEEE Transactions on Information Theory*, 60, 1313–1325. [1916,1917]

Cai, T., Liu, W., and Luo, X. (2011), "A Constrained ℓ_1 Minimization Approach to Sparse Precision Matrix Estimation," *Journal of the American Statistical Association*, 106, 594–607. [1909]

Cai, T. T., Li, H., Liu, W., and Xie, J. (2016), "Joint Estimation of Multiple High-Dimensional Precision Matrices," *Statistica Sinica*, 26, 445–464. [1909,1915,1916,1918,1921]

Chen, X., Xu, M., and Wu, W. B. (2013), "Covariance and Precision Matrix Estimation for High-Dimensional Time Series," *The Annals of Statistics*, 41, 2994–3021. [1909]

Collier, O., Comminges, L., and Tsybakov, A. B. (2017), "Minimax Estimation of Linear and Quadratic Functionals on Sparsity Classes," *The Annals of Statistics*, 45, 923–958. [1915]

Cule, M., Samworth, R., and Stewart, M. (2010), "Maximum Likelihood Estimation of a Multi-Dimensional Log-Concave Density," *Journal of the Royal Statistical Society*, Series B, 72, 545–607. [1920]

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. [1909,1915,1916,1918,1923]

Fan, J., Feng, Y., and Wu, Y. (2009), "Network Exploration via the Adaptive LASSO and SCAD Penalties," *The Annals of Applied Statistics*, 3, 521–541. [1909]

Fan, Y., and Lv, J. (2016), "Innovated Scalable Efficient Estimation in Ultra-Large Gaussian Graphical Models," *The Annals of Statistics*, 44, 2098–2126. [1909]

Friedman, J., Hastie, T., and Tibshirani, R. (2008), "Sparse Inverse Covariance Estimation With the Graphical Lasso," *Biostatistics*, 9, 432–441. [1909]

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2011), "Joint Estimation of Multiple Graphical Models," *Biometrika*, 98, 1–15. [1909,1915,1916]

Huang, J., and Zhang, T. (2010), "The Benefit of Group Sparsity," *The Annals of Statistics*, 38, 1978–2004. [1916]

Ingster, Y., and Suslina, I. A. (2012), *Nonparametric Goodness-of-Fit Testing Under Gaussian Models* (Vol. 169), Berlin: Springer Science & Business Media. [1915]

Kolar, M., Song, L., Ahmed, A., and Xing, E. P. (2010), "Estimating Time-Varying Networks," *The Annals of Applied Statistics*, 4, 94–123. [1909]

Laurent, B., and Massart, P. (2000), "Adaptive Estimation of a Quadratic Functional by Model Selection," *The Annals of Statistics*, 28, 1302–1338. [1917]

Lauritzen, S. L. (1996), *Graphical Models*, Oxford: Oxford University Press. [1908]

Liu, H., Wang, L., and Zhao, T. (2015), "Calibrated Multivariate Regression With Application to Neural Semantic Basis Discovery," *Journal of Machine Learning Research*, 16, 1579–1606. [1916]

Liu, W. (2013), "Gaussian Graphical Model Estimation With False Discovery Rate Control," *The Annals of Statistics*, 41, 2948–2978. [1909,1911,1914]

Lounici, K., Pontil, M., Van De Geer, S., and Tsybakov, A. B. (2011), "Oracle Inequalities and Optimal Inference Under Group Sparsity," *The Annals of Statistics*, 39, 2164–2204. [1916]

Lu, J., Kolar, M., and Liu, H. (2015), "Post-Regularization Inference for Dynamic Nonparanormal Graphical Models," arXiv preprint arXiv:1512.08298. [1909]

Ma, T., Ren, Z., and Tseng, G. C. (2018), "Variable Screening With Multiple Studies," *Statistica Sinica*. [1921]

Marigorta, U., and Navarro, A. (2013), "High Trans-Ethnic Replicability of GWAS Results Implies Common Causal Variants," *PLoS Genet*, 9, e1003566. [1909,1913]

Meinshausen, N., and Bühlmann, P. (2006), "High Dimensional Graphs and Variable Selection With the Lasso," *The Annals of Statistics*, 34, 1436–1462. [1909]

Mitra, R., and Zhang, C.-H. (2016), "The Benefit of Group Sparsity in Group Inference With De-Biased Scaled Group Lasso," *Electronic Journal of Statistics*, 10, 1829–1873. [1916]

Qiu, H., Han, F., Liu, H., and Caffo, B. (2016), "Joint Estimation of Multiple Graphical Models From High Dimensional Time Series," *Journal of the Royal Statistical Society*, Series B, 78, 487–504. [1909]

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011), "High-Dimensional Covariance Estimation by Minimizing ℓ_1 Penalized Log-Determinant Divergence," *Electronic Journal of Statistics*, 5, 935–980. [1909]

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. [1909,1915]

Sun, T., and Zhang, C.-H. (2012), "Scaled Sparse Linear Regression," *Biometrika*, 99, 879–898. [1912,1916]

Teng, S.-H. (2016), "Scalable Algorithms for Data and Network Analysis," *Foundations and Trends® in Theoretical Computer Science*, 12, 1–274. [1908]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [1916]

Uematsu, Y., Fan, Y., Chen, K., Lv, J., and Lin, W. (2017), "SOFAR: Large-Scale Association Network Learning," arXiv:1704.08349. [1916]

Vershynin, R. (2010), "Introduction to the Non-Asymptotic Analysis of Random Matrices," arXiv:1011.3027. [1922]

Wainwright, M. J., and Jordan, M. I. (2008), "Graphical Models, Exponential Families, and Variational Inference," *Foundations and Trends in Machine Learning*, 1, 1–305. [1908]

Yuan, M. (2010), "High Dimensional Inverse Covariance Matrix Estimation via Linear Programming," *Journal of Machine Learning Research*, 11, 2261–2286. [1909]

Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression With Grouped Variables," *Journal of the Royal Statistical Society*, Series B, 68, 49–67. [1916]

Yuan, M., and Lin, Y. (2007), "Model Selection and Estimation in the Gaussian Graphical Model," *Biometrika*, 94, 19–35. [1909]

Zhang, T., and Zou, H. (2014), "Sparse Precision Matrix Estimation via Lasso Penalized D-trace Loss," *Biometrika*, 101, 103–120. [1909]

Zhou, S., Lafferty, J., and Wasserman, L. (2010), "Time Varying Undirected Graphs," *Machine Learning*, 80, 295–319. [1909]

Zhu, Y., Shen, X., and Pan, W. (2014), "Structural Pursuit Over Multiple Undirected Graphs," *Journal of the American Statistical Association*, 109, 1683–1696. [1909,1915,1916]