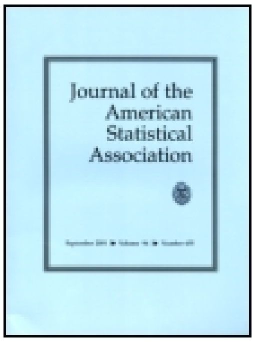

Journal of the
American
Statistical
Association

Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

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

Likelihood Ratio Tests for a Large Directed Acyclic Graph

Chunlin Li, Xiaotong Shen & Wei Pan

To cite this article: Chunlin Li, Xiaotong Shen & Wei Pan (2019): Likelihood Ratio Tests for a Large Directed Acyclic Graph, Journal of the American Statistical Association, DOI: 10.1080/01621459.2019.1623042

To link to this article: https://doi.org/10.1080/01621459.2019.1623042

Accepted author version posted online: 24 May 2019.
Published online: 25 Jun 2019.

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2019, VOL. 00, NO. 0, 1–16: Theory and Methods
https://doi.org/10.1080/01621459.2019.1623042

Taylor & Francis
Taylor & Francis Group

Likelihood Ratio Tests for a Large Directed Acyclic Graph

Chunlin Li[a], Xiaotong Shen[a], and Wei Pan[b]

[a]School of Statistics, University of Minnesota, Minneapolis, MN; [b]Division of Biostatistics, University of Minnesota, Minneapolis, MN

ABSTRACT
Inference of directional pairwise relations between interacting units in a directed acyclic graph (DAG), such as a regulatory gene network, is common in practice, imposing challenges because of lack of inferential tools. For example, inferring a specific gene pathway of a regulatory gene network is biologically important. Yet, frequentist inference of directionality of connections remains largely unexplored for regulatory models. In this article, we propose constrained likelihood ratio tests for inference of the connectivity as well as directionality subject to nonconvex acyclicity constraints in a Gaussian directed graphical model. Particularly, we derive the asymptotic distributions of the constrained likelihood ratios in a high-dimensional situation. For testing of connectivity, the asymptotic distribution is either chi-squared or normal depending on if the number of testable links in a DAG model is small. For testing of directionality, the asymptotic distribution is the minimum of d independent chi-squared variables with one-degree of freedom or a generalized Gamma distribution depending on if d is small, where d is number of breakpoints in a hypothesized pathway. Moreover, we develop a computational method to perform the proposed tests, which integrates an alternating direction method of multipliers and difference convex programming. Finally, the power analysis and simulations suggest that the tests achieve the desired objectives of inference. An analysis of an Alzheimer's disease gene expression dataset illustrates the utility of the proposed method to infer a directed pathway in a gene network.

ARTICLE HISTORY
Received October 2018
Accepted May 2019

KEYWORDS
Directed acyclic graph; Gene network; High-dimensional inference; L_0-regularization; Nonconvex minimization.

1. Introduction

Directional pairwise relations have been essential to represent local Markov property (Edwards 2012) between interacting units in a directed acyclic graph (DAG), as in regulatory gene network analysis (Sachs et al. 2005) and in human brain network analysis (Liu et al. 2017). In brain network analysis, effective connectivity or directional function connectivity becomes informative to infer the interactions among brain regions of interest from a neuroscience perspective. In this article, we develop constrained likelihood ratio tests to make a formal inference of directional pairwise relations with respect to the strength of connectivity as well as directionality.

In the statistics literature, in spite of its critical importance, significance testing concerning directional pairwise relations in DAG models has received much less attention than discovery or point estimation of the DAG structures (Bühlmann, Peters, and Ernest 2014; Fu and Zhou 2013; Gu, Fu, and Zhou 2017), including Bayesian network inference (Cheng et al. 2002; Liu 2001; Luo and Zhao 2011). To our knowledge, frequentist inference of directional pairwise relations remains largely unexplored for DAG models, especially so with an unknown DAG partial ordering and a growing dimension of model parameters, although a confidence set for relations to a single target node has recently become available in a different context for causal discovery with interventions (Peters, Bühlmann, and Meinshausen 2016; Rothenhäusler, Bühlmann, and Meinshausen 2019). In our situation, several challenges emerge. First, any local approach based on multiple tests of the node connectivity of a graph, followed by enumeration over all candidate nodes, is likely to lose power when a multiplicity adjustment is made for a large number of individual tests to be performed. Second, a series of directional pairwise relations defined by a pathway requires certain acyclicity constraints of directional connections to ensure validity of the local Markov property (Yuan et al. 2019). Consequently, the corresponding inference is constrained, making it difficult to derive the distribution of a test statistic, especially with a growing dimension of constraints. Note that even the asymptotic distribution of a likelihood ratio test subject to low-dimensional constraints can be nonstandard (Geyer 1994). In other words, new treatments are necessary.

In this article, we propose constrained likelihood ratio tests for the strength of directed connectivity and for a given directed pathway, of a large directed graph. In particular, a constrained likelihood ratio test is subject to nonconvex constraints regularizing nuisance parameters that are not being tested, in addition to an acyclicity constraint to ensure a DAG. The parameters of interest are not regularized to alleviate the impact of regularization on inference. For testing the strength of directed connectivity, the asymptotic distribution of the constrained likelihood ratio is a chi-squared or a normal distribution, depending on if the number of testable links in a DAG model is fixed or increases with the sample size, particularly when $\frac{|D^0|^{1/2}|E^0 \cup D^0| \log p}{n} \to 0$ (Assumption 4), where D^0 is the largest testable subindex set (Definition 1) defined by the null hypothesis and $D^0 \cup E^0$ is the DAG index set under its alternative, and p is the number of nodes and $| \cdot |$ denotes a set's size. For testing a given

CONTACT Xiaotong Shen ✉ shenx002@umn.edu 🖃 School of Statistics, University of Minnesota, Minneapolis, MN 55455.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.

directed pathway, the asymptotic distribution of the constrained likelihood ratio is the minimum of d independent chi-squared random variables with one-degree of freedom when d is small, but is a generalized Gamma distribution when d is large, where d is the number of breakpoints in a hypothesized pathway when $\frac{|E^0| \log p}{n} \rightarrow 0$ (Assumption 5), where $|E^0|$ is the number of nonzero links of the truth DAG. In either situation, p may exceed the sample size n. Moreover, our power analysis suggests that the proposed tests achieve the desired objective of inference. To our knowledge, our result is the first of the kind, providing constrained likelihood ratio tests to infer multiple directional pairwise relations in a high-dimensional situation. Finally, we design an algorithm based on difference convex programming and alternating direction method of multipliers (Gabay and Mercier 1975) to compute constrained likelihood ratios, guaranteeing a resulting graph to be DAG under the null and alternative hypotheses.

The rest of the article is organized as follows. Section 2 briefly introduces the DAG model and its associated acyclicity constraint. Section 3 proposes two constrained likelihood ratio tests for inference of graph linkages and a directed pathway, and provides specific conditions under which asymptotic approximations of the sampling distributions of the test statistics are valid. Section 4 develops computational methods to solve constrained optimization. Section 5 performs simulation studies, followed by an application of the tests to infer some gene pathways possibly related to Alzheimer's disease with a late-onset Alzheimer's disease (LOAD) dataset (Webster et al. 2009).

2. DAG Models and Acyclicity

To infer pairwise relations imposed by the local Markov dependence (Edwards 2012) for a vector of Gaussian random variables $(Y_1, \ldots, Y_p)^T$, consider a DAG model induced by structure equations

$$Y_j = \sum_{k:k \neq j} U_{kj} Y_k + e_j, \quad e_j \sim N(0, \sigma^2), \tag{1}$$

where $\boldsymbol{U} = (U_{kj})_{p \times p}$ is an adjacency matrix, ^{T} denotes the transpose, and $e_1, \ldots, e_p$ are independently identically distributed random errors. Note that (1) is identifiable with respect to model parameters $\boldsymbol{U}$ and σ^2 (Peters and Bühlmann 2014).

Model (1), when satisfying acyclicity constraint (3), induces a DAG model (Yuan et al. 2019), factorizing the joint probability distribution of $(Y_1, \ldots, Y_p)^T$, $P(y_1, \ldots, y_p) = \prod_{j=1}^p P(y_j \mid \mathrm{pa}_j)$, into a product of conditional distributions of each variable given its parent, where pa_j denotes the parent set of Y_j and is allowed to be empty when Y_j has no parent. This factorization is known as the local Markov property (Edwards 2012) and is related to generalized Markov dependence (Núñez-Antón and Zimmerman 2009). Importantly, $U_{jk} \neq 0$ encodes a directed edge from Y_j to Y_k under (3), whereas $U_{jk} = 0$ implies conditional independence of Y_j and Y_k given the parent of Y_i, occurring when Y_k is not a parent of Y_j.

In (1), given a random sample $\{(Y_{i1}, \ldots, Y_{ip})^T\}_{i=1}^n$ of size n, the log-likelihood is proportional to

$$l(\boldsymbol{U}, \sigma^2) = -\sum_{j=1}^p \Big(\frac{1}{2\sigma^2} \sum_{i=1}^n \big(y_{ij} - \sum_{k \neq j} U_{jk} y_{ik}\big)^2 + \frac{n}{2} \log \sigma^2 \Big), \tag{2}$$

where $\boldsymbol{U}$ and σ^2 are estimated from (2) subject to the requirement that $\boldsymbol{U}$ defines a DAG, or a directed graph without directed cycles. This enables us to identify all directional pairwise relations simultaneously by identifying nonzero off-diagonals of $\boldsymbol{U}$.

According to Yuan et al. (2019), $\boldsymbol{U}$ renders a DAG if the following acyclicity constraints are satisfied:

$$\sum_{j_1 = j_{L+1} : 1 \leq k \leq L} \mathbb{I}(U_{j_{k-1} j_k} \neq 0) \leq L - 1; \\ \text{any } (j_1, \ldots, j_L),\ L = 2, \ldots, p, \tag{3}$$

which ensures the local Markov property defining directional pairwise relations (Edwards 2012). It follows from Lemma B1 of Yuan et al. (2019) that (3) is recast into a constraint on the maximum of p linear programs; $k = 1, \ldots, p$,

$$(p-1) \geq \max_{q_{ij}} \sum_{1 \leq i,j \leq p} c_{ij}^k q_{ij}, \\ \text{subj to } (q_{ij})_{p \times p} \text{ is a doubly stochastic matrix,} \tag{4}$$

where $c_{ij}^k = \mathbb{I}(U_{ij} \neq 0)$ if $i \neq j$, $c_{ij}^k = 0$ if $i = j = k$, and $c_{ii}^k = 1$ otherwise. After introducing nonnegative dual variables $\boldsymbol{\Lambda} = (\lambda_{ik})_{p \times p}$ for (4), we obtain an equivalent form of (3),

$$\lambda_{ik} + \mathbb{I}(j \neq k) - \lambda_{jk} \geq \mathbb{I}(U_{ij} \neq 0);\ i, j, k = 1, \ldots, p,\ i \neq j. \tag{5}$$

See Yuan et al. (2019) for technical details.

To treat nonconvex constraints in (5), we replace the corresponding indicator functions in (5) by its computational surrogate $\mathrm{J}_\tau(z) = \min(\frac{|z|}{\tau}, 1)$ (Shen, Pan, and Zhu 2012) to circumvent the discontinuity in optimization so that difference convex programming (Horst and Tuy 2013) is applicable, where τ is a small tuning parameter controlling the degree of approximation in that $\mathrm{J}_\tau(z)$ approximates the indicator function as $\tau \rightarrow 0^+$. This yields that

$$\lambda_{ik} + \mathbb{I}(j \neq k) - \lambda_{jk} \geq \mathrm{J}_\tau(U_{ij}); \quad i, j, k = 1, \ldots, p,\ i \neq j. \tag{6}$$

3. Constrained Likelihood Ratio Tests

In the framework of (1), we develop two types of tests concerning directional pairwise relations, encoded by adjacent matrix $\boldsymbol{U} = (U_{ij})_{p \times p}$ in Sections 3.1 and 3.2.

3.1. Test of Graph Linkages

Before specifying hypotheses, we define an index set $F \subseteq \{1, \ldots, p\}^2$, where an index (j, k) represents a directed connection from variables j and k, or the corresponding node j is a parent node of k for a DAG under (1). Given the index set F, the null H_0 and alternative H_a hypotheses for testing linkages are

$$H_0 : U_{ij} = 0;\ \text{all } (i, j) \in F \ \text{ versus } \ H_a : \text{ not } H_0,$$

with unspecified nuisance parameters U_{ij}; $(i,j) \in F^c$ and $\sigma^2 > 0$, where ^{c} denotes complement. Rejection of H_0 indicates the strength of directed connectivity specified by F.

The constrained likelihood ratio for H_0 versus H_a is $Lr = l(\hat{\boldsymbol{U}}^{H_a}, \hat{\sigma}^2_{H_a}) - l(\hat{\boldsymbol{U}}^{H_0}, \hat{\sigma}^2_{H_0})$, where

$$\begin{aligned}
(\hat{\boldsymbol{U}}^{H_0}, \hat{\sigma}^2_{H_0}) &= \underset{(\boldsymbol{U},\sigma^2,\boldsymbol{\Lambda})}{\arg\max}\ l(\boldsymbol{U},\sigma^2), \text{subj to (6)},\ \boldsymbol{U}_F = 0,\\
&\quad \sum_{(i,j)\in F^c} J_\tau(|U_{ij}|) \le \kappa;\\
(\hat{\boldsymbol{U}}^{H_a}, \hat{\sigma}^2_{H_a}) &= \underset{(\boldsymbol{U},\sigma^2,\boldsymbol{\Lambda})}{\arg\max}\ l(\boldsymbol{U},\sigma^2),\\
&\quad \text{subj to (6)},\ \sum_{(i,j)\in F^c} J_\tau(|U_{ij}|) \le \kappa, \qquad (7)
\end{aligned}$$

where $\kappa \ge 0$ is an integer-valued tuning parameter and $\boldsymbol{\Lambda} = (\lambda_{ij})_{p\times p}$ is defined after (6).

To derive the asymptotic distribution of Lr under H_0, we make some technical assumptions. Denote by $\boldsymbol{\Omega} = (\boldsymbol{I}-\boldsymbol{U})^T(\boldsymbol{I}-\boldsymbol{U})/\sigma^2$, and $\boldsymbol{\Omega}_S = (\boldsymbol{I}-\boldsymbol{U}_S)^T(\boldsymbol{I}-\boldsymbol{U}_S)/\sigma^2$, where $\boldsymbol{U}_S$ is an adjacency matrix with its support index set S, that is, $\boldsymbol{U}_{S^c} = 0$. In what follows, let $\boldsymbol{\Omega}^0, \boldsymbol{U}^0$ be the truth parameters, and E^0 represents the true support index of $\boldsymbol{U}^0$.

Assumption 1 (Boundedness of the parameter space). For some constants $c_1 > 0$ and $c_2 > 0$, $c_1 \le \lambda_{\min}(\boldsymbol{\Omega}) \le \lambda_{\max}(\boldsymbol{\Omega}) \le c_2$, where $\lambda_{\min}(\boldsymbol{\Omega})$ and $\lambda_{\max}(\boldsymbol{\Omega})$ denote the smallest and largest eigenvalues of $\boldsymbol{\Omega}$, respectively.

Assumption 2 (Degree of separation). For some positive constant c_3,

$$\begin{aligned}
C_{\min} \equiv \inf_{\{\boldsymbol{\Omega}_{S_1\cup S_2}: S_1\cup S_2 \not\supseteq E^0, |S_1| \le |E^0\setminus F|, S_2 \subseteq F, S_1\cup S_2 \text{ forms DAG}\}}\\
\frac{-\log(1-h^2(\boldsymbol{\Omega}_{S_1\cup S_2},\boldsymbol{\Omega}^0))}{|E^0\setminus(S_1\cup S_2)|}\\
\ge 4c_3 n^{-1}\max(|E^0|+|F|, \log p),
\end{aligned}$$

where $h^2(\boldsymbol{\Omega},\boldsymbol{\Omega}^0) \equiv 1 - \left(\frac{[\det(\boldsymbol{\Omega})\det(\boldsymbol{\Omega}^0)]^{1/2}}{\det(\frac{\boldsymbol{\Omega}+\boldsymbol{\Omega}^0}{2})}\right)^{1/2}$.

Assumption 3 (Approximation). For some positive constants c_4–c_6,

$$h^2(\boldsymbol{\Omega},\boldsymbol{\Omega}^0) \ge c_4 h^2(\boldsymbol{\Omega}_\tau,\boldsymbol{\Omega}^0) - c_6 p\tau^{c_5},$$

where $\boldsymbol{\Omega}_\tau = (\boldsymbol{I}-\boldsymbol{U}_\tau)^T(\boldsymbol{I}-\boldsymbol{U}_\tau)/\sigma^2$ and $\boldsymbol{U}_\tau$ is the truncated version of $\boldsymbol{U}$ on F^c, defined by $(U_\tau)_{ij} = U_{ij}\mathbb{I}(|U_{ij}| > \tau)$ for $(i,j) \in F^c$, and $(U_\tau)_{ij} = U_{ij}$ for $(i,j) \in F$.

Assumptions 1–3 are regularity conditions to quantify the parameter space and the degree of separation of candidate models.

Before proceeding, we note that not every link in F may contribute one degree of freedom to Lr due to the acyclicity constraint. For example, in a two-node graph, if $E^0 = \{(1,2)\}$ and $F = \{(2,1)\}$, then $E^0 \cup F$ does not render a DAG, and $Lr = 0$ is degenerate asymptotically. This motivates us to introduce the notion of "testability" and "nonsingularity."

Definition 1 (Testability and nonsingularity). A link $(j,k) \in F$ is called testable if $\{(j,k)\} \cup E^0$ does not contain a directed cycle. An index set F is called nonsingular if $D^0 \cup E^0$ forms a DAG, where D^0 consists of all testable links. Note that D^0 can be empty when F is trivially nonsingular.

Interestingly, only the links in D^0 contribute degrees of freedom to Lr. Next we impose a condition that restricts the size of the index set F under H_0.

Assumption 4 (Restriction of the index set F). Assume that F is nonsingular under H_0 and

$$\frac{|D^0|^{1/2}|E^0 \cup D^0|\log p}{n} \to 0, \quad \text{as } n \to \infty,$$

where $|\cdot|$ denotes the size of the set.

Assumption 4 restricts $|D^0|$, $|E^0|$, p, and n in that $\frac{|D^0|^{1/2}|E^0\cup D^0|\log p}{n} \to 0$, permitting p exceeding n provided that the number of testable links in D^0 and nonzero nuisance links under H_0 are not too large.

Theorem 1 (Sampling distribution of Lr for testing linkage under H_0). Assume that Assumptions 1–4 are met. Then there exist tuning parameters $\kappa = |E^0 \setminus F|$ and $\tau \le C_{\min}c_1/4p$ such that under H_0, as $n \to \infty$,

$$\begin{aligned}
&(i) && P(Lr = 0) \to 1, && \text{if } |D^0| = 0,\\
&(ii) && 2Lr \xrightarrow{d} \chi^2_{|D^0|}, && \text{if } |D^0| > 0 \text{ is fixed},\\
&(iii) && (2|D^0|)^{-1/2}(2Lr - |D^0|) \xrightarrow{d} N(0,1), && \text{if } |D^0| \to \infty.
\end{aligned}$$

The degrees of freedom $|D^0|$, as specified in Assumption 4, is usually unknown, and hence that we propose an estimate $\hat{D}^0$ of D^0, defined as the largest subset of F such that $F \cup \hat{E}^0$ forms a DAG. Here $\hat{D}^0$ is estimated based on an estimate $\hat{\boldsymbol{U}}^0_{H_0}$ under H_0 from (7).

The next corollary says that $|\hat{D}^0|$ consistently estimates $|D^0|$.

Corollary 1 (Substitution of D^0 by $\hat{D}^0$). Under the assumptions of Theorem 1, $P(|\hat{D}^0| = |D^0|) \to 1$, as $n \to \infty$. Then the result of Theorem 1 continues to hold with D^0 replaced by $\hat{D}^0$.

On the ground of Theorem 1, we proceed our constrained likelihood ratio test with following empirical rule: (1) fail to reject H_0 if $|\hat{D}^0| = 0$, (2) use $\chi^2_{|\hat{D}^0|}$ for Lr if $1 \le |\hat{D}^0| < 25$, (3) use normal distribution for Lr if $|\hat{D}^0| \ge 25$, as a chi-squared distribution can be well approximated by a normal distribution when its degrees of freedom is sufficiently large.

3.2. Test of a Directed Pathway

A direct pathway is specified by an index set F in a consecutive manner, where a common segment is shared by two consecutive indices of $F = \{(i_1,i_2),(i_2,i_3),\ldots,(i_{|F|},i_{|F|+1})\}$, for instance, (i_1,i_2) and (i_2,i_3) are shared by i_2. Now consider

$$\begin{aligned}
H_0 &: U_{ij} = 0;\ \text{for some } (i,j) \in F \ \text{ versus}\\
H_a &: U_{ij} \ne 0;\ \text{for all } (i,j) \in F
\end{aligned}$$

with unspecified nuisance parameters U_{ij}; $(i,j) \in F^c$ and $\sigma^2 > 0$. Rejection of H_0 suggests the presence of a directed pathway specified by F.

In this situation, H_0 in (8) dramatically differs from that in (7). Now the constrained likelihood ratio statistic for H_0 versus H_a is modified to account for a directed pathway at some indices: $Lr = l(\hat{\boldsymbol{U}}^{H_a}, \hat{\sigma}^2_{H_a}) - \max_{k=1}^{|F|} l(\hat{\boldsymbol{U}}^{H_0}(k), \hat{\sigma}_{H_0}(k))$, where

$$\begin{aligned}
(\hat{\boldsymbol{U}}^{H_0}(k), \hat{\sigma}^2_{H_0}(k)) &= \underset{(\boldsymbol{U},\sigma^2,\boldsymbol{\Lambda})}{\arg\max}\; l(\boldsymbol{U},\sigma^2),\\
&\quad \text{subj to } (6),\; U_{i_k,i_{k+1}} = 0; (i_k, i_{k+1}) \in F,\\
&\sum_{(i,j)\in F^c} \mathrm{J}_\tau(|U_{ij}|) \le \kappa,\\
(\hat{\boldsymbol{U}}^{H_a}, \hat{\sigma}^2_{H_a}) &= \underset{(\boldsymbol{U},\sigma^2,\boldsymbol{\Lambda})}{\arg\max}\; l(\boldsymbol{U},\sigma^2), \text{ subj to } (6),\\
&\sum_{(i,j)\in F^c} \mathrm{J}_\tau(|U_{ij}|) \le \kappa. \qquad (8)
\end{aligned}$$

Next, Assumption 5 is a modified version of Assumption 4.

Assumption 5 (Restriction of the index set E^0).

$$\frac{|E^0| \log p}{n} \to 0, \quad \text{as } n \to \infty.$$

Theorem 2 (Sampling distribution of Lr for testing a directed pathway under H_0). Under Assumptions 1, 2, 3, and 5, there exist tuning parameters $\kappa = |E^0 \setminus F|$ and $\tau \le C_{\min} c_1 / 4p$ such that under H_0, as $n \to \infty$,

$$\begin{array}{lll}
(i) & P(Lr = 0) \to 1 & \text{if } E^0 \cup F \text{ does not form a DAG;}\\
(ii) & 2Lr \xrightarrow{d} \min\{X_1, \ldots, X_d\} & \text{if } E^0 \cup F \text{ forms a DAG and}\\
 & & d \text{ is fixed;}\\
(iii) & 2d^2 Lr \xrightarrow{d} \Gamma & \text{if } E^0 \cup F \text{ forms a DAG but}\\
 & & d \to \infty,
\end{array}$$

where $d = |\{(i,j) \in F : U^0_{ij} = 0\}|$ is the number of breakpoints in the hypothesized pathway, $X_1, \ldots, X_d$ are independently identically distributed χ^2_1 variables, and Γ is the generalized Gamma distribution with density $\sqrt{\frac{1}{2\pi x}} \exp(-\sqrt{2x/\pi})$ for $x > 0$.

The degrees of freedom d is estimated by $\hat{d} = \max(|\{(i,j) \in F : \hat{U}_{ij} = 0\}|, 1)$, where $\hat{\boldsymbol{U}}$ is the constrained maximum likelihood estimate under H_a in (8) but with $F = \emptyset$. The next corollary says that $\hat{d}$ consistently estimates d.

Corollary 2 (Substitution of d by $\hat{d}$). Under the assumptions of Theorem 2, $P(\hat{d} = d) \to 1$, as $n \to \infty$. Then the result of Theorem 2 continues to hold with d replaced by $\hat{d}$.

In practice, we proceed with our test as follows: (1) fail to reject H_0 if $\hat{E}^0 \cup D$ does not form a DAG, (2) use the distribution of $\min\{X_1, \ldots, X_d\}$ for Lr if $\hat{E}^0 \cup D$ forms a DAG but $\hat{d} < 25$, (3) use the generalized gamma distribution for Lr if $\hat{d} \ge 25$.

3.3. Power Analysis

This section analyzes the local power of the CLR tests for (7) and (8) separately.

In (7), consider a local alternative H_a: $U_{ij} = U^0_{ij} + \delta^n_{ij}$ for $(i,j) \in F$, with $U^0_{ij} = 0$ for $(i,j) \in F$, where $\boldsymbol{\delta}^n = (\delta^n_{ij})_{p\times p}$ satisfies: $\boldsymbol{\delta}^n_{F^c} = 0$, and $\|\boldsymbol{\delta}^n\|_F = n^{-1/2}h$ if $|D^0|$ is fixed and $\|\boldsymbol{\delta}^n\|_F = |D^0|^{1/4} n^{-1/2} h$ if $|D^0| \to \infty$, where $\boldsymbol{\delta}^n = (\delta^n_{ij})_{p\times p}$ and $\|\cdot\|_F$ is the Frobenius norm of a matrix. Now define the power function $\pi_n(h)$ as 1 with $\boldsymbol{U}^n = \boldsymbol{U}^0$ and $h = 0$ if $|D^0| = 0$, $P_{\boldsymbol{U}^n}(2Lr \ge \chi^2_{|D_0|,1-\alpha})$ if $|D^0| > 0$ is fixed, and $P_{\boldsymbol{U}^n}\Big((2|D^0|)^{-1/2}(2Lr - |D^0|) \ge z_{1-\alpha}\Big)$ if $|D^0| \to \infty$, where $\boldsymbol{U}^n = \boldsymbol{U}^0 + \boldsymbol{\delta}^n$, with $\boldsymbol{U}^0$ being an adjacency matrix such that $U^0_{ij} = 0$ for any $(i,j) \in D$, and $\chi^2_{|D^0|,1-\alpha}$ and $z_{1-\alpha}$ are the $(1-\alpha)$th quantiles of the distributions of $\chi^2_{|D^0|}$ and $N(0,1)$, respectively.

The next theorem confirms that the CLR test for (7) has the power function tending to 1 when the sample size tends to infinity.

Theorem 3 (Local limiting power for graph linkages). For any $\boldsymbol{U}^0$ satisfying Assumptions 1–4 such that $\boldsymbol{U}^n$ induces a DAG, we have that $\lim_{h\to\infty} \liminf_{n\to\infty} \pi_n(h) = 1$.

In (8), consider a local alternative H_a: $U_{ij} = U^0_{ij} + \delta^n_{ij}$ for $(i,j) \in A$, where $|\delta^n_{ij}| = \frac{h}{\sqrt{n}}$ for $(i,j) \in A$ and $A = \{(i,j) \in F : U^0_{ij} = 0\}$. Let the power function $\tilde{\pi}_n(h)$ be 1 with $\boldsymbol{U}^n = \boldsymbol{U}^0$ and $h = 0$ if $E^0 \cup F$ is cyclic, $P_{\boldsymbol{U}^n}(2Lr \ge \Gamma_{d,1-\alpha}) = 1$ if $E^0 \cup F$ is acyclic and d is fixed, and $P_{\boldsymbol{U}^n}(2d^2 Lr \ge \Gamma_{1-\alpha}) = 1$ if $E^0 \cup F$ is acyclic and $d \to \infty$, where $\Gamma_{d,1-\alpha}$ and $\Gamma_{1-\alpha}$ are the $(1-\alpha)$th quantiles of the minimal of the chi-squares and Γ distributions in Theorem 2.

Theorem 4 gives a parallel result for (8).

Theorem 4 (Local limiting power for a directed pathway). For any $\boldsymbol{U}^0$ satisfying Assumptions 1, 2, 3, 5, such that $\boldsymbol{U}^n$ induces a DAG, we have that $\lim_{h\to\infty} \liminf_{n\to\infty} \tilde{\pi}_n(h) = 1$.

4. Optimization and Computation

This section develops a strategy to compute the CLR tests for (7) and (8). To treat the nonconvex minimization in (7) or (8), we develop a difference convex programming approach to iteratively relax the nonconvex constraints through a sequence of convex set approximations. Then each convex subproblem is solved by an alternating direction method of multipliers (Boyd et al. 2011). The process iterates until a termination criterion is met.

Specifically, we decompose J_τ into a difference of two convex functions $\mathrm{J}_\tau(z) = |z|/\tau - \max(|z|/\tau - 1, 0) \equiv f_1(z) - f_2(z)$. On this ground, a convex approximation at $(m+1)$th iteration is constructed by replacing $f_2(z)$ with its affine majorization $f_2(z^{[m]}) + \nabla f_2(z^{[m]})^T (z - z^{[m]})$ at the solution $z^{[m]}$ at mth iteration, where $\nabla f_2(z^{[m]}) = \frac{\mathrm{Sign}(z^{[m]})}{\tau} \mathbb{I}(|z^{[m]}| > \tau)$ is a subgradient of f_2 at $z^{[m]}$. This leads to a convex subproblem at the $(m+1)$th iteration,

$$
\begin{aligned}
&\max_{(\boldsymbol{U},\sigma^2,\boldsymbol{\Lambda})} \quad l(\boldsymbol{U},\sigma^2)\\
&= -\textstyle\sum_{j=1}^{p}\left(\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_{ij}-\sum_{k\neq j}U_{jk}y_{ik}\right)^2+\frac{n}{2}\log\sigma^2\right)\\
&\text{subj to} \quad \textstyle\sum_{(i,j)\notin E_1}|U_{ij}|\mathbb{I}(|\hat{U}_{ij}^{[m]}|\le\tau)\\
&\le \tau(\kappa-\textstyle\sum_{(i,j)\notin E_1}\mathbb{I}(|\hat{U}_{ij}^{[m]}|>\tau)), \quad \boldsymbol{U}_{E_2}=0,\\
&\tau\lambda_{jk}+\tau\mathbb{I}(i\neq k)\\
&-\tau\lambda_{ik}\ge|U_{ij}|\mathbb{I}(|\hat{U}_{ij}^{[m]}|\le\tau)+\tau\mathbb{I}(|\hat{U}_{ij}^{[m]}|>\tau),\\
&i,j,k=1,\dots,p, i\neq j, \qquad (9)
\end{aligned}
$$

where $E_1 = E_2 = F$ for H_0 in (7), $E_1 = F$ and $E_2 = \emptyset$ for H_a in (7), $E_1 = F$ and $E_2 = \{(i_k, i_{k+1})\}$ for H_0 in (8), and $E_1 = F$, $E_2 = \emptyset$ for H_a in (8), and $\hat{\boldsymbol{U}}^{[m]}$ is the solution of (9) at its previous mth iteration.

To solve (9), we employ an alternating direction method of multipliers (Boyd et al. 2011). Note that (9) can proceed through profiling by maximizing with respect to σ^2 and $(\boldsymbol{U},\boldsymbol{\Lambda})$ separately. Plugging an expression $\hat{\sigma}^2 = (np)^{-1}\sum_{j=1}^{p}\sum_{i=1}^{n}(y_{ij} - \sum_{k\neq j} y_{ik}U_{jk})^2$ into (9), it reduces to minimization of $\mathrm{RSS}(\boldsymbol{U}) \equiv \frac{1}{2}\sum_{j=1}^{p}\sum_{i=1}^{n}(y_{ij} - \sum_{k\neq j} y_{ik}U_{jk})^2$ in $(\boldsymbol{U},\boldsymbol{\Lambda})$ subject to the constraints in (9). Now consider an equivalent form of (9) by introducing a nonnegative multiplier μ

$$
\begin{aligned}
&\min_{(\boldsymbol{U},\boldsymbol{\Lambda})} \quad \mathrm{RSS}(\boldsymbol{U})+\mu\sum_{(i,j)\notin E_1}|U_{ij}|\mathbb{I}(|\hat{U}_{ij}^{[m]}|\le\tau),\\
&\quad \text{subj to} \quad \boldsymbol{U}_{E_2}=0, \qquad (10)\\
&U_{ij}\mathbb{I}(|\hat{U}_{ij}^{[m]}|\le\tau)\le\tau\lambda_{jk}+\tau\mathbb{I}(i\neq k)-\tau\lambda_{ik}-\tau\mathbb{I}(|\hat{U}_{ij}^{[m]}|>\tau), \qquad (11)\\
&U_{ij}\mathbb{I}(|\hat{U}_{ij}^{[m]}|\le\tau)\ge-\tau\lambda_{jk}-\tau\mathbb{I}(i\neq k)+\tau\lambda_{ik}\\
&\quad +\tau\mathbb{I}(|\hat{U}_{ij}^{[m]}|>\tau), \quad i,j,k=1,\dots,p, i\neq j. \qquad (12)
\end{aligned}
$$

To solve (10), we separate its differentiable from non-differentiable parts by introducing a decoupling matrix $\boldsymbol{V}$ for $\boldsymbol{U}$, in addition to slack variables $\boldsymbol{\xi} = \{\xi_{ijk}^1, \xi_{ijk}^2\}_{p\times p\times p}$ to convert inequality to equality constraints. This yields

$$
\begin{aligned}
&\min_{(\boldsymbol{U},\boldsymbol{V},\boldsymbol{\Lambda},\boldsymbol{\xi})} \quad \mathrm{RSS}(\boldsymbol{U})+\mu\sum_{(i,j)\notin E_1}|V_{ij}|\mathbb{I}(|\hat{U}_{ij}^{[m]}|\le\tau),\\
&\text{subj to} \quad \boldsymbol{U}_{E_2}=0, \quad \boldsymbol{U}-\boldsymbol{V}=\boldsymbol{0},\\
&V_{ij}\mathbb{I}(|\hat{U}_{ij}^{[m]}|\le\tau)+\tau\mathbb{I}(|\hat{U}_{ij}^{[m]}|>\tau)\\
&+\xi_{ijk}^1-\tau\lambda_{jk}-\tau\mathbb{I}(i\neq k)+\tau\lambda_{ik}=0,\\
&V_{ij}\mathbb{I}(|\hat{U}_{ij}^{[m]}|\le\tau)-\tau\mathbb{I}(|\hat{U}_{ij}^{[m]}|>\tau)\\
&-\xi_{ijk}^2+\tau\lambda_{jk}+\tau\mathbb{I}(i\neq k)-\tau\lambda_{ik}=0,\\
&\xi_{ijk}^1,\xi_{ijk}^2\ge0, \quad i,j,k=1,\dots,p, i\neq j.
\end{aligned}
$$

Following Boyd et al. (2011), an introduction of scaled dual variables $\boldsymbol{\alpha} = \{\alpha_{ijk}^1, \alpha_{ijk}^2\}_{p\times p\times p}$ and $\boldsymbol{W} = \{w_{ij}\}_{p\times p}$ leads to an augmented Lagrangian,

$$
\begin{aligned}
&L_\rho(\boldsymbol{V},\boldsymbol{U},\boldsymbol{\Lambda},\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{W})\\
&=\mathrm{RSS}(\boldsymbol{U})+\mu\sum_{(i,j)\notin E_1}|V_{ij}|\mathbb{I}(|\hat{U}_{ij}^{[m]}|\le\tau)+\frac{\rho}{2}\|\boldsymbol{U}-\boldsymbol{V}+\boldsymbol{W}\|_F^2\\
&+\frac{\rho}{2}\sum_k\sum_{i\neq j}\Big(V_{ij}\mathbb{I}(|\hat{U}_{ij}^{[m]}|\le\tau)\\
&+\tau\mathbb{I}(|\hat{U}_{ij}^{[m]}|>\tau)+\xi_{ijk}^1-\tau\lambda_{jk}-\tau\mathbb{I}(i\neq k)+\tau\lambda_{ik}+\alpha_{ijk}^1\Big)^2\\
&+\frac{\rho}{2}\sum_k\sum_{i\neq j}\Big(V_{ij}\mathbb{I}(|\hat{U}_{ij}^{[m]}|\le\tau)\\
&-\tau\mathbb{I}(|\hat{U}_{ij}^{[m]}|>\tau)-\xi_{ijk}^2+\tau\lambda_{jk}+\tau\mathbb{I}(i\neq k)-\tau\lambda_{ik}+\alpha_{ijk}^2\Big)^2,
\end{aligned}
$$

where the minimization is solved iteratively. At $(s+1)$th iteration of ADMM, we update following steps:

$$
\begin{aligned}
&\boldsymbol{V}^{[s+1]}=\arg\min_{\boldsymbol{V}} L_\rho(\boldsymbol{V},\boldsymbol{U}^{[s]},\boldsymbol{\Lambda}^{[s]},\boldsymbol{\xi}^{[s]},\boldsymbol{\alpha}^{[s]},\boldsymbol{W}^{[s]}),\\
&\boldsymbol{U}^{[s+1]}=\arg\min_{\boldsymbol{U}} L_\rho(\boldsymbol{V}^{[s+1]},\boldsymbol{U},\boldsymbol{\Lambda}^{[s]},\boldsymbol{\xi}^{[s]},\boldsymbol{\alpha}^{[s]},\boldsymbol{W}^{[s]}),\\
&(\boldsymbol{\Lambda}^{[s+1]},\boldsymbol{\xi}^{[s+1]})=\arg\min_{\boldsymbol{\Lambda}} L_\rho(\boldsymbol{V}^{[s+1]},\boldsymbol{U}^{[s+1]},\boldsymbol{\Lambda},\boldsymbol{\xi},\boldsymbol{\alpha}^{[s]},\boldsymbol{W}^{[s]}),\\
&\boldsymbol{W}^{[s+1]}=\boldsymbol{W}^{[s]}+\boldsymbol{U}^{[s]}-\boldsymbol{V}^{[s]}, \qquad (13)
\end{aligned}
$$

where $\alpha_{ijk}^{1[s+1]} = V_{ij}^{[s]}\mathbb{I}(|U_{ij}^{[m]}| \le \tau) + \tau\mathbb{I}(|U_{ij}^{[m]}| > \tau) + \xi_{ijk}^{1[s]} - \tau\lambda_{jk}^{[s]} - \tau\mathbb{I}(i \neq k) + \tau\lambda_{ik}^{[s]} + \alpha_{ijk}^{1[s]}$, $\alpha_{ijk}^{2[s+1]} = V_{ij}^{[s]}\mathbb{I}(|U_{ij}^{[m]}| \le \tau) - \tau\mathbb{I}(|U_{ij}^{[m]}| > \tau) - \xi_{ijk}^{2[s]} + \tau\lambda_{jk}^{[s]} + \tau\mathbb{I}(i \neq k) - \tau\lambda_{ik}^{[s]} + \alpha_{ijk}^{2[s]}$, and an updating formula is given in the Appendix to facilitate computation.

The strategy for computing $\hat{\boldsymbol{U}}^{H_0}$ and $\hat{\boldsymbol{U}}^{H_a}$ is summarized in Algorithm 1.

Algorithm 1. **Step 1: (Initialization)** Fix E_1 and E_2 in (9). Initiate an estimate $(\boldsymbol{U}^{[0]}, \boldsymbol{\Lambda}^{[0]})$ satisfying (6). Set $E^{[0]} = \{(i,j) : |U_{ij}^{[0]}| > \tau\}$.

Step 2: (ADMM) At mth iteration, compute $(\boldsymbol{U}^{[m]}, \boldsymbol{\Lambda}^{[m]})$ by ADMM through (13).

Step 3: (Check for acyclicity) Let $E^{[m]} = \{(i,j) : |U_{ij}^{[m]}| > \tau\}$. If $E^{[m]}$ constitutes a cycle in the graph, sort $|U_{ij}^{[m]}|$ decreasingly; $(i,j) \in E^{[m]}$. For each $(i,j) \in E^{[m]}$ in order, if $E^{[m]} \setminus \{(i,j)\}$ induces a DAG, update $E^{[m]}$ by $E^{[m]} \setminus \{(i,j)\}$. Otherwise, keep $E^{[m]}$ intact. Here a cycle detection algorithm such as depth-first search is applied (Cormen et al. 2001).

Step 4: (Termination) Repeat Steps 2 and 3 until a termination criterion is met, that is, $\mathrm{RSS}(\boldsymbol{U}^{[m-1]}) - \mathrm{RSS}(\boldsymbol{U}^{[m]}) \le \varepsilon$. The final solution $\boldsymbol{U}^{[m^*]}$ is the corresponding estimate under H_0 or H_a, where m^* is the smallest index at termination.

Importantly, Step 3 in Algorithm 1 ensures that $\boldsymbol{U}^{[m^*]}$ satisfies the acyclicity condition by removing the weakest link in an existing cycle, hence that it yields a DAG.

Concerning Algorithm 1, we note that its complexity over five blocks in one iteration is roughly of order $p^3 + np^2$. In terms of convergence speed, based on our limited numerical experience, the ADMM component converges with a modest accuracy within a few thousand iterations, while the difference convex programming component usually terminates within ten steps.

Theorem 5 (Convergence of Algorithm 1). For $\rho > 0$ and sufficiently small $\tau > 0$, Algorithm 1 yields a local minimizer

$\hat{\boldsymbol{U}}$, which satisfies the optimality condition for some multipliers $\mu \geq 0$ and $\{\nu_{ijk} \geq 0\}_{i,j,k=1,\ldots,p;i\neq j}$,

$$\partial_{ij} l(\boldsymbol{U}) + \frac{\mu}{\tau}\partial_{ij} J_\tau(|U_{ij}|) + \frac{\sum_{1\leq k\leq p} \nu_{ijk}}{\tau}\partial_{ij} J_\tau(|U_{ij}|) = 0;$$
$$i,j = 1,\ldots,p, i \neq j,$$

where ∂_{ij} denotes the subgradient (Rockafellar and Wets 2011).

As indicated by Theorem 5, Algorithm 1 yields a local minimizer. However, as showed in Table 3, it can yield a global minimizer or a good local minimizer. In fact, the probability that the solution of Algorithm 1 agrees with the oracle estimator is high, which is a global minimizer asymptotically, see Lemmas 1 and 2 in the Appendix. This aspect has been recognized (Tao 2005) for a difference convex algorithm. Note, however, that a global minimizer can be attained if Breiman and Culter's outer approximation method (a version of difference convex algorithm) can attain a global minimizer at the expense of slow convergence (Breiman and Cutler 1993).

5. Numerical Examples

This section examines operating characteristics of the proposed tests and compares against oracle tests in regard to the size and power in simulated and real data examples. Here the oracle tests for (7) and (8) are $Lr_{\mathrm{OR}} = l(\hat{\boldsymbol{U}}_{E^0\cup F}, \hat{\sigma}_{E^0\cup F}) - l(\hat{\boldsymbol{U}}_{E^0\setminus F}, \hat{\sigma}_{E^0\setminus F})$ and $Lr_{\mathrm{OR}} = l(\hat{\boldsymbol{U}}_{E^0\cup F}, \hat{\sigma}_{E^0\cup F}) - \max_{k=1}^{|F|} l(\hat{\boldsymbol{U}}_{E^0\cup F\setminus\{(i_k,i_pathwayk+1)\}}, \hat{\sigma}_{E^0\cup F\setminus\{(i_k,i_{k+1})\}}$, based on the constrained maximum likelihood estimates $(\hat{\boldsymbol{U}}_{E^0\setminus F}, \hat{\sigma}_{E^0\setminus F})$ and $(\hat{\boldsymbol{U}}_{E^0\cup F\setminus\{(i_k,i_{k+1})\}}, \hat{\sigma}_{E^0\cup F\setminus\{(i_k,i_{k+1})\}})$ assuming that the true graph structure would be known in advance.

In simulations, we consider three different types of DAGs, known as random, hub, and chain graphs, as displayed in Figure 1, and report the empirical size and power of a test. For the size of a test, we compute the percentage of times rejecting H_0 out of 1000 simulations when H_0 is true. For the power of a test, we examine four alternatives H_a hypotheses. The empirical power of a test is the percentage of times rejecting H_0 out of 100 simulations when H_a is true.

For the proposed method, tuning parameters (μ, τ) are estimated by maximizing the 5-fold predictive likelihood $\hat{l}(\mu, \tau)$, defined as

$$\frac{1}{5}\sum_{l=1}^{5}\sum_{j=1}^{p}\Big(\frac{1}{2\hat{\sigma}^2_{-l}(\mu,\tau)}$$
$$\sum_{y_i\in\mathcal{D}_l}\big(y_{ij} - \sum_{k\neq j}\hat{U}^{-l}_{jk}(\mu,\tau)y_{ik}\big)^2 + \frac{|\mathcal{D}_l|}{2}\log\hat{\sigma}^2_{-l}(\mu,\tau)\Big),$$

where $\hat{\boldsymbol{U}}^{-l}(\mu, \tau)$ and $\hat{\sigma}^2_{-l}(\mu, \tau)$ are the estimates under H_a based on a random partition of the original data with roughly equal parts, $\mathcal{D}_l$; $l = 1,\ldots,5$. Then the optimal tuning parameters $(\hat{\mu}, \hat{\tau}) = \arg\min \hat{l}(\mu, \tau)$ are used to compute the final estimator based on original data $\mathcal{D}$. In our simulations, μ is selected from $10^{0.5l}$ and τ is selected from $0.05 + 0.1l$; $l = 0,\ldots,4$.

Numerical results are produced using an R package "clrdag", implementing the proposed constrained likelihood ratio tests based on Algorithm 1.

5.1. Simulated Examples

Example 1 (Test for linkage). In (1), consider a random DAG of p nodes in the absence of no graph structures, defined by adjacency matrix $\boldsymbol{U}$, with $\sigma = 1$. The graph is displayed in Figure 1. For the value of $\boldsymbol{U}$, upper off-diagonals are set as a random sample from $\{0, 1\}$ using the Bernoulli distribution with success probability $2/p$, and the rest of entries of $\boldsymbol{U}$ are set to be zero. For the linkage test in (7), with $(i_0, j_0) = (2, p)$, we examine two different H_0 hypotheses: (i) $H_0 : U_{i_0j_0} = 0$; (ii) $H_0 : U_{ij} = 0$; $i = 2,\ldots,16, j = p-1, p$. Whereas the first concerns one single entry of $\boldsymbol{U}$, the second focuses on the last two columns. Here $|D^0| = |F|$. Moreover, for the power analysis, we look at four alternatives $H_a : U_{i_0j_0} = 0.1l$; $l = 1, 2, 3$; $U_{ij} = 0$ otherwise.

Example 2 (Test for linkage). This example considers consider a hub graph DAG of p nodes in a set-up as in Example 1, see Figure 1 for a display of a network. To define adjacency matrix $\boldsymbol{U}$, we set $U_{1j} = 1$; $j = 2,\ldots,p$, and set remaining entries to be zero. The other set-up remains the same as in Example 1.

As a result, the neighborhood of the first node is dense, while the overall graph remains sparse.

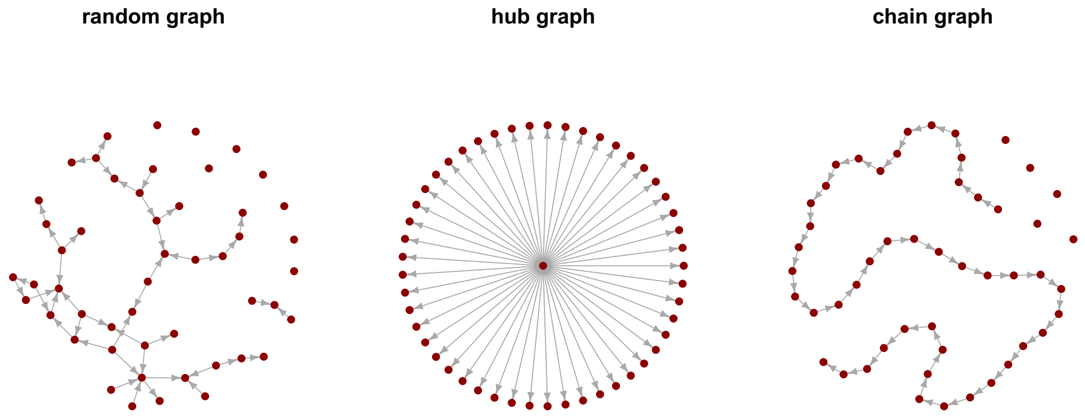

Figure 1. Three types of graphs used in Examples 1–3.

Table 1. Empirical size and power of the proposed CLR test for testing linkages (7) in Examples 1 and 2, where the chi-squared or normal tests with $\alpha = 0.05$ are used based on $|\hat{D}^0|$.

			CLR		Oracle LR	
Graph	$\|D^0\|$	(p, n)	Size	Power	Size	Power
Random	1	$(50, 500)$	0.053	$(0.56, 0.99, 1.00)$	0.051	$(0.57, 0.99, 1.00)$
		$(100, 500)$	0.054	$(0.57, 1.00, 0.99)$	0.056	$(0.59, 1.00, 1.00)$
		$(200, 500)$	0.039	$(0.54, 0.98, 1.00)$	0.050	$(0.62, 1.00, 1.00)$
		$(250, 200)$	0.060	$(0.46, 0.98, 1.00)$	0.054	$(0.55, 0.98, 1.00)$
	30	$(50, 500)$	0.050	$(0.19, 0.62, 1.00)$	0.052	$(0.22, 0.67, 1.00)$
		$(100, 500)$	0.062	$(0.17, 0.74, 0.99)$	0.066	$(0.20, 0.76, 1.00)$
		$(200, 500)$	0.047	$(0.18, 0.70, 0.96)$	0.063	$(0.19, 0.81, 0.99)$
		$(250, 200)$	0.055	$(0.09, 0.66, 0.96)$	0.051	$(0.11, 0.75, 0.99)$
Hub	1	$(50, 500)$	0.049	$(0.65, 0.95, 1.00)$	0.048	$(0.65, 1.00, 1.00)$
		$(100, 500)$	0.043	$(0.65, 0.93, 1.00)$	0.053	$(0.64, 0.99, 1.00)$
		$(200, 500)$	0.053	$(0.63, 0.96, 0.99)$	0.047	$(0.65, 1.00, 1.00)$
		$(250, 200)$	0.043	$(0.54, 1.00, 0.99)$	0.038	$(0.58, 1.00, 0.99)$
	30	$(50, 500)$	0.048	$(0.16, 0.78, 1.00)$	0.054	$(0.15, 0.83, 1.00)$
		$(100, 500)$	0.059	$(0.23, 0.75, 0.99)$	0.057	$(0.19, 0.82, 0.99)$
		$(200, 500)$	0.057	$(0.20, 0.58, 0.91)$	0.055	$(0.20, 0.62, 1.00)$
		$(250, 200)$	0.041	$(0.14, 0.53, 0.90)$	0.040	$(0.15, 0.60, 1.00)$

Table 2. Empirical size and power of the proposed likelihood ratio test for testing a directed pathway (8) in a chain graph in Example 3, where the minimum $\hat{d}$ chi-square test with $\alpha = 0.05$ is used.

	CLR		Oracle LR	
(p, n)	Size	Power	Size	Power
$(50, 500)$	0.045	$(0.84, 1.00, 1.00)$	0.050	$(0.91, 1.00, 1.00)$
$(100, 500)$	0.057	$(0.78, 0.94, 0.99)$	0.051	$(0.90, 1.00, 1.00)$
$(200, 500)$	0.052	$(0.73, 0.90, 0.94)$	0.046	$(0.91, 1.00, 1.00)$
$(250, 200)$	0.060	$(0.59, 0.80, 0.92)$	0.053	$(0.80, 0.97, 1.00)$

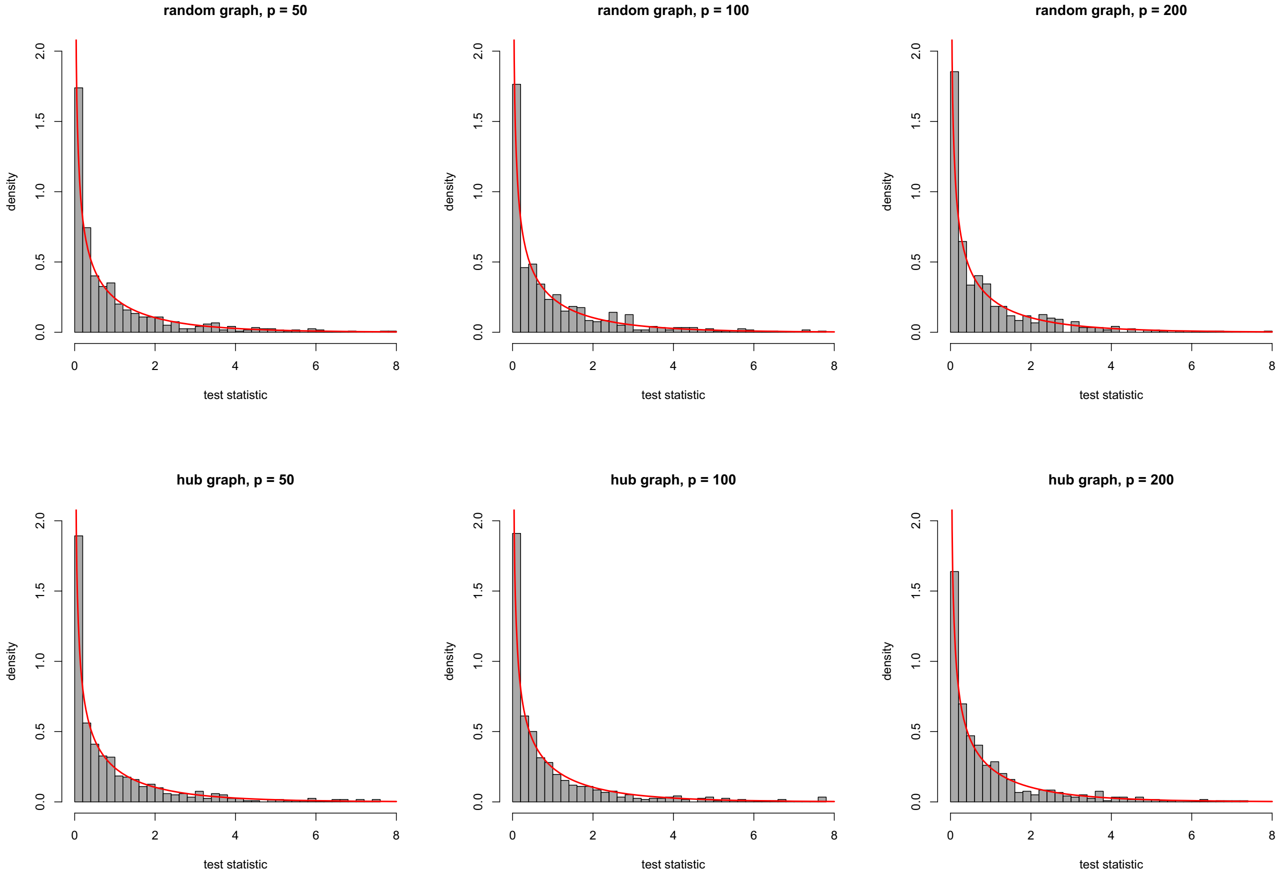

Figure 2. Empirical null distribution of the proposed CLR linkage test based on the chi-squared approximation with $n = 500$ and $|D^0| = 1$.

Example 3 (Test for a pathway). This example focuses on testing a pathway in a chain DAG of p nodes in (1) with $\sigma = 1$, see Figure 1 for a display of a graph. First, we construct a directed pathway from nodes 1 to 50 with other node variables being independent. Toward this end, let $U_{i(i+1)} = 1; i = 1, \dots, 49$, and $U_{ij} = 0$ otherwise. Second, randomly sample $\nu_0 = 5$ edges from $\{(i, i+1) : i = 1, \dots, 49\}$ without replacement. Now $\boldsymbol{U}_A = 0$, where A is the set of sampled edges.

For the pathway test in (8), let $F = \{(i, i+1) : i = 1, \dots, 49\}$ and consider: $H_0 : U_{i(i+1)} = 0$ for some $(i, i+1) \in F$ versus $H_a : U_{i(i+1)} \neq 0$ for all $(i, i+1) \in F$. For the power calculations, we examine four alternatives $H_a : U_{i(i+1)} = 1$ for $(i, i+1) \in F \setminus S$ and $U_{i(i+1)} = 0.1l$; $(i, i+1) \in S$; $l = 1, 2, 3$, and $U_{ij} = 0$ otherwise.

Example 4 (Oracle rate of CLR). This example illustrates that Algorithm 1 can yield a global or a good local optimum for (7) with a reasonably good chance. In particular, we compare the constrained likelihood ratio Lr in (7) against the corresponding oracle constrained likelihood ratio Lr_{OR} to see if our test can reconstruct the oracle test. Note that $Lr_{\mathrm{OR}} = l(\hat{\boldsymbol{U}}_{E^0 \cup F}, \hat{\sigma}_{E^0 \cup F}) - l(\hat{\boldsymbol{U}}_{E^0 \setminus F}, \hat{\sigma}_{E^0 \setminus F})$, where $(\hat{\boldsymbol{U}}_{E^0 \cup F}, \hat{\sigma}_{E^0 \cup F})$ and $(\hat{\boldsymbol{U}}_{E^0 \setminus F}, \hat{\sigma}_{E^0 \setminus F})$ are maximizers of the constrained likelihood in (7) asymptotically. In the setting of Examples 1 and 2, we generate random and hub graphs with p nodes and consider the chi-squared linkage test over the hypothesized index set $F = \{(2, p)\}$, where $H_0 : U_{2p} = 0$. Now we define the oracle rate of CLR as the chance that Lr and Lr_{OR} are sufficiently close (tol $\leq 10^{-4}$).

5.1.1. Power and Empirical Size of the CLR Tests

As indicated in Tables 1 and 2, the tests perform well for testing linkages of a DAG in Examples 1 and 2 as well as a directed pathway of a DAG in Example 3. Specifically, the empirical sizes are close to the nominal level 0.05 in all scenarios. Moreover, the power functions of the tests exhibit desirable properties in that it increases as the sample size n increases when p is held fixed, and/or increases to 1 as the level of difficulty of a test decreases, determined by the alternative hypothesis H_a. This is consistent with the result of Theorems 1–4. Overall, the proposed tests perform well when comparing against the corresponding oracle tests.

5.1.2. Asymptotic Approximations for Testing Linkages (7) and a Directed Pathway (8)

For (7), as suggested by Figures 2 and 3, the chi-squared and normal approximations appear adequate for $|D^0| = 1$ and $|D^0| = 30$ in Examples 1 and 2, respectively. For (8), the approximation by minimum of d chi-squared variables seems adequate for $d = 5$ in Example 3, as suggested by Figure 4.

5.1.3. Globality of Lr in (7)

As suggested by Table 3 in Example 4, the proposed CLR test has good agreement rates with the oracle test, exceeding 95%. Roughly, the constrained likelihood estimator based on Algorithm 1 are likely to attain an approximate global optimum of nonconvex minimization in (7). This aspect of a difference

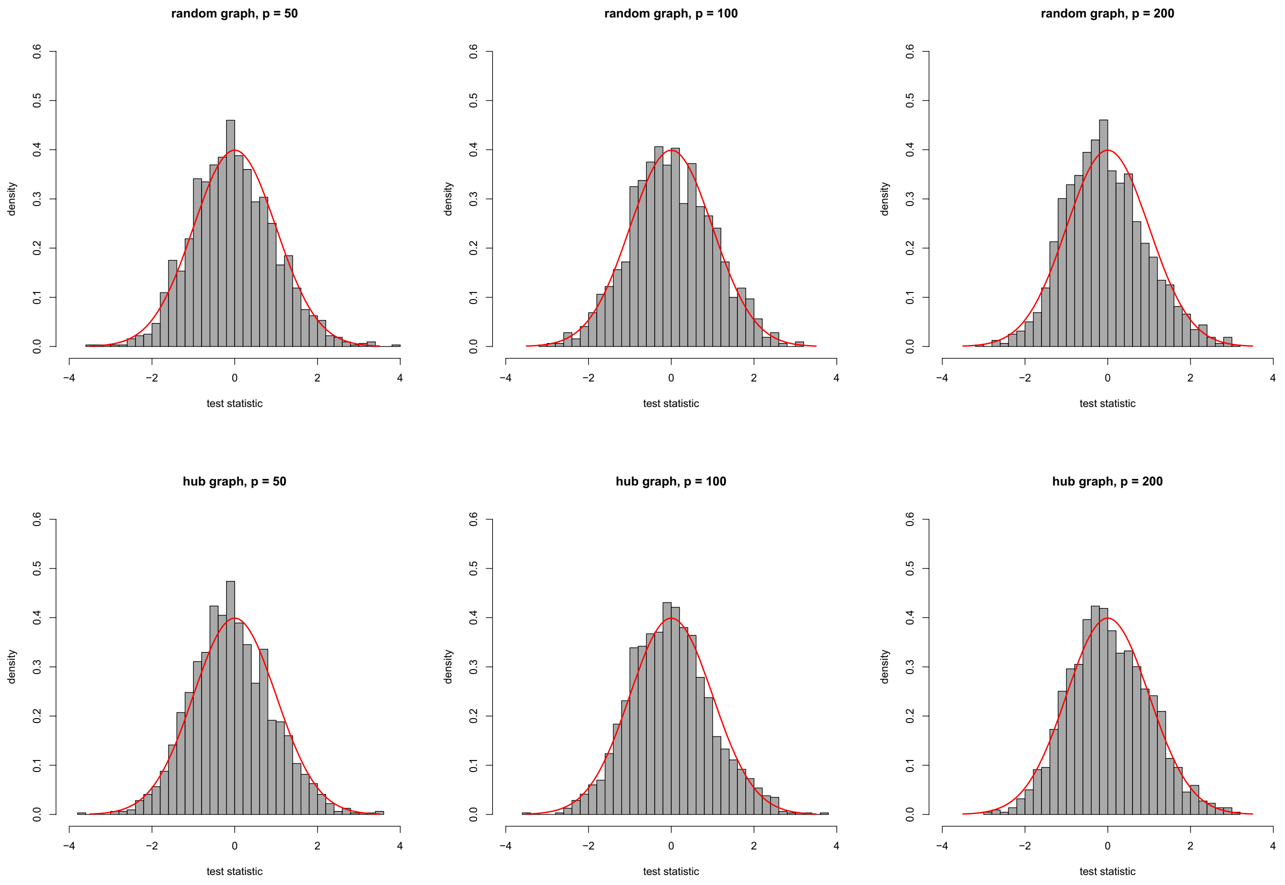

Figure 3. Empirical null distribution of the proposed CLR linkage test based on the normal approximation with $n = 500$ and $|D^0| = 30$.

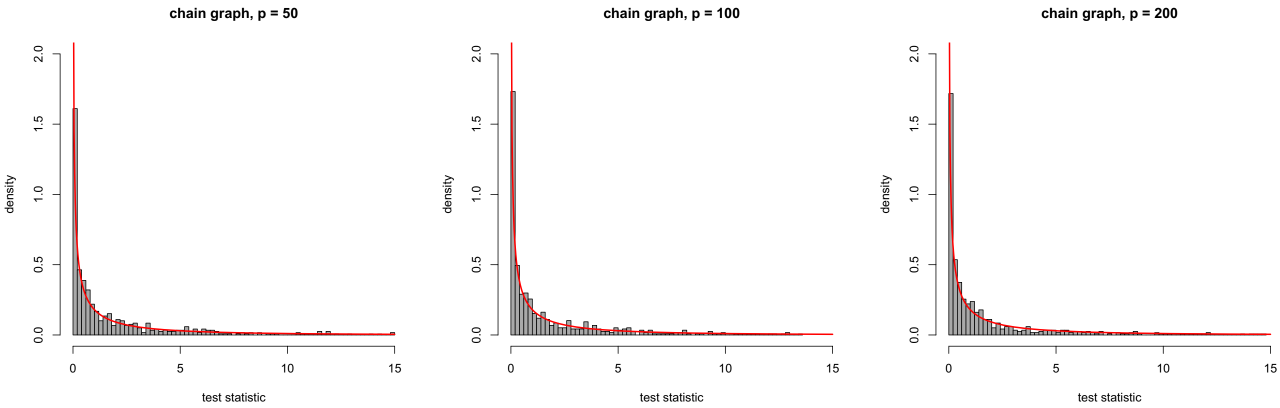

Figure 4. Empirical null distribution of the proposed CLR pathway test based on the minimum $\hat{d}$ chi-squared approximation with $n = 500$ and $d = 5$.

Table 3. Percentage of agreement between the proposed CLR test and the oracle test in (7) based on 100 simulation replications.

(p, n)	Random	Hub
(10, 200)	0.99	0.99
(30, 600)	0.95	1.00
(50, 1000)	0.97	0.99

convex algorithm agrees with findings of Tao (2005) for different nonconvex problems.

5.2. Gene Networks for Alzheimer's Disease

This section applies the proposed tests to analyze the late-onset Alzheimer's disease (AD) dataset (Webster et al. 2009), containing expression levels of 8560 genes in the brain tissue for 176 AD patients and 187 healthy controls (HCs). Our primary goal is to infer gene pathways related to AD, then contrast the corresponding pathways between the AD and HC groups to highlight some gene-gene interactions possibly differentiating these two groups.

From the KEGG database (Kanehisa and Goto 2000), we extract the AD reference pathway (map05010), including 99 genes in the AD data. Then we conduct a differential gene expression analysis on these genes with a two-sample t-test, detecting $p = 21$ genes differentially expressed (at the significance level 0.01) between the AD and HC groups. After some data preprocessing, the gene expression levels of the 21 genes for 154 AD and 182 HC individuals are approximately normally distributed.

The existing literature suggests that the Bcl-2-associated death promoter (BAD) gene is involved in neurodegeneration (Foster et al. 2001). Thus, for this dataset we focus attention on a regulatory subnetwork associated with gene BAD in the AD reference pathway map, particularly three genetic pathways/subnetworks, (P_1) CALM3 → PPP3CA → BAD, (P_2) CALM3 → PPP3CB → BAD, and (P_3) CALM3 → PPP3R1 → BAD. We consider testing of three hypotheses for simultaneous presence of the three pathways specified by P_l; $l = 1, 2, 3$: $H_0 : U_{ij} = 0$ for some $(i, j) \in \mathrm{P}_l$ versus $H_a : U_{ij} \neq 0$ for all $(i, j) \in \mathrm{P}_l$.

The p-values and significant pathways under the significance level $\alpha = 0.01$ after the Holm–Bonferroni adjustment are displayed in Figure 5. As a comparison, the reconstructed networks, using constrained maximum likelihood in (7) with $F = \emptyset$, are displayed in Figure 6. Interestingly, both pathways CALM3 → PPP3CA → BAD and CALM3 → PPP3CB → BAD are present in the reconstructed network for the AD group, but the proposed test suggests that only one pathway CALM3 → PPP3CA → BAD is significant (p-value = 0.001735). Besides

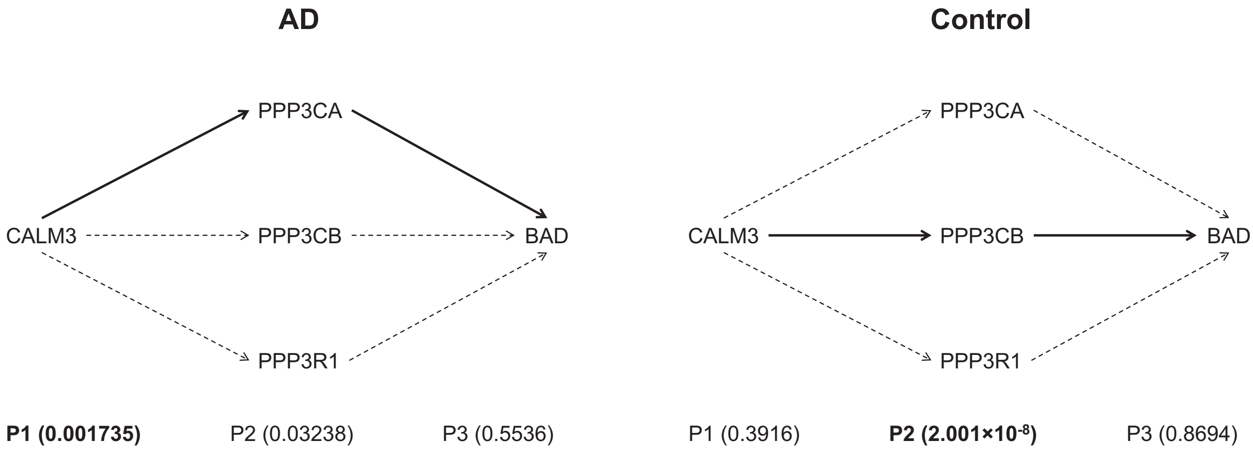

Figure 5. A subnetwork associated with BAD, consisting of dashed pathways denoted by P_1–P_3, for multiple hypothesis testing, where solid pathways are significant under level $\alpha = 0.01$ after the Holm–Bonferroni correction and adjusted p-values of these pathways are given in parentheses for multiplicity.

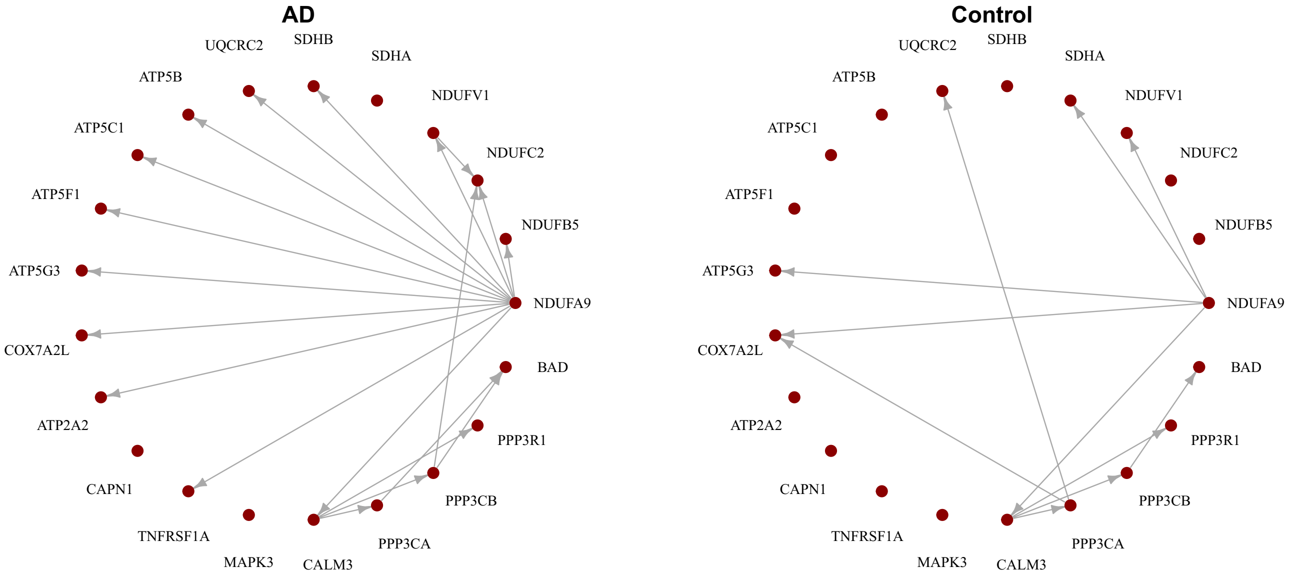

Figure 6. Reconstructed gene networks for the Alzheimer disease and healthy groups.

this difference, the results of the testing and reconstruction mostly agree: the pathway CALM3 → PPP3CB → BAD is identified for the HC group (p-value = 2.001×10^{-8}), whereas the last pathway CALM3 → PPP3R1 → BAD is not present for both groups. Our result suggests that the regulatory subnetwork associated with BAD for AD patients differs from that for healthy individuals. In summary, our analysis suggest that the two pathways/subnetworks CALM3 → PPP3CA → BAD and CALM3 → PPP3CB → BAD may differentiate these two groups, though more follow-up studies are needed.

For model checking, the residual plots and Q–Q plots in Figure 7 suggest no strong evidence against the normality assumption and the equal-variance assumption for the error terms in the structure equations model (1).

Appendix A

A.1. Analytical Updating Formulas for (13)

***V* direction:** Elementwise minimization yields that

$$V_{ij}^{[s+1]} = \begin{cases} \text{Soft}_{\mu/\rho(2p+1)}\left(\gamma_{ij}^{[s]}\right) & \text{if } (i,j) \notin E_1, |U_{ij}^{[m]}| \le \tau, \\ \gamma_{ij}^{[s]} & \text{if } (i,j) \in E_1, |U_{ij}^{[m]}| \le \tau, \\ U_{ij}^{[s]} + W_{ij}^{[s]} & \text{if } |U_{ij}^{[m]}| > \tau, \end{cases}$$

where $\text{Soft}_r(\cdot)$ denotes r-soft-thresholding and $\gamma_{ij}^{[s]} = (2p+1)^{-1}\Big[U_{ij}^{[s]} + W_{ij}^{[s]} - \sum_k \left(\xi_{ijk}^{1[s]} - \xi_{ijk}^{2[s]} + \alpha_{ijk}^{1[s]} + \alpha_{ijk}^{2[s]}\right)\Big]$.

***U* direction:** The minimizer satisfies $\boldsymbol{U}_{\cdot j}^{[s+1]} = (\boldsymbol{U}_{S_j}, \boldsymbol{0})^T; j = 1, \dots, p$ with

$$(\boldsymbol{Y}_{S_j}^T \boldsymbol{Y}_{S_j} + \rho \boldsymbol{I}_{|S_j| \times |S_j|}) \boldsymbol{U}_{S_j} = \boldsymbol{Y}_{S_j}^T \boldsymbol{Y}_j + \rho(\boldsymbol{V}_{S_j}^{[s+1]} - \boldsymbol{W}_{S_j}^{[s]}),$$

where $S_j = \{(i,j) : i \ne j \text{ and } (i,j) \notin E_2\}$.

$(\boldsymbol{\Lambda}, \boldsymbol{\xi})$ direction: The updating formula of $\boldsymbol{\Lambda}$ is

$$\boldsymbol{\Lambda}^{[s+1]} = \boldsymbol{M}_{p \times p} \boldsymbol{L}_{p \times p}^{[s+1]},$$

where

$$\boldsymbol{M}_{p \times p} = \frac{1}{\tau} \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & \frac{2}{p} & \frac{1}{p} & \dots & \frac{1}{p} \\ 1 & \frac{1}{p} & \frac{2}{p} & \dots & \frac{1}{p} \\ \vdots & \vdots & & \ddots & \vdots \\ 1 & \frac{1}{p} & \dots & \frac{1}{p} & \frac{2}{p} \end{pmatrix},$$

$$\begin{aligned} L_{1j}^{[s+1]} &= 1, \\ L_{ik}^{[s+1]} &= \frac{1}{2}\Big[\tau + \frac{1}{2}\sum_j \Big(2\tau \mathbb{I}(|\hat{U}_{ij}^{[m]}| > \tau) \\ &\quad + \xi_{ijk}^{1[s+1]} + \xi_{ijk}^{2[s+1]} + \alpha_{ijk}^{1[s]} - \alpha_{ijk}^{2[s]}\Big) \\ &\quad - \frac{1}{2}\sum_j \Big(2\tau \mathbb{I}(|\hat{U}_{ji}^{[m]}| > \tau) + \xi_{jik}^{1[s+1]} + \xi_{jik}^{2[s+1]} \\ &\quad + \alpha_{jik}^{1[s]} - \alpha_{jik}^{2[s]}\Big)\Big]; \quad i \ne k, \\ L_{kk}^{[s+1]} &= \frac{1}{2}\Big[- (p-1)\tau + \frac{1}{2}\sum_j \Big(2\tau \mathbb{I}(|\hat{U}_{kj}^{[m]}| > \tau) \\ &\quad + \xi_{kjk}^{1[s+1]} + \xi_{ijk}^{2[s+1]} + \alpha_{kjk}^{1[s]} - \alpha_{kjk}^{2[s]}\Big) \\ &\quad - \frac{1}{2}\sum_j \Big(2\tau \mathbb{I}(|\hat{U}_{jk}^{[m]}| > \tau) + \xi_{jkk}^{1[s+1]} \\ &\quad + \xi_{jkk}^{2[s+1]} + \alpha_{jkk}^{1[s]} - \alpha_{jkk}^{2[s]}\Big)\Big]. \end{aligned}$$

The updating formula of $\boldsymbol{\xi}$ is

$$\begin{aligned} \xi_{ijk}^{1[s+1]} &= \max(0, \tau\lambda_{jk}^{[s]} + \tau\mathbb{I}(i \ne k) \\ &\quad - \tau\lambda_{ik}^{[s]} - V_{ij}^{[s+1]}\mathbb{I}(|\hat{U}_{ij}^{[m]}| \le \tau) - \tau\mathbb{I}(|\hat{U}_{ij}^{[m]}| > \tau) - \alpha_{ijk}^{1[s]}), \\ \xi_{ijk}^{2[s+1]} &= \max(0, \tau\lambda_{jk}^{[s]} + \tau\mathbb{I}(i \ne k) - \tau\lambda_{ik}^{[s]} \\ &\quad + V_{ij}^{[s+1]}\mathbb{I}(|\hat{U}_{ij}^{[m]}| \le \tau) - \tau\mathbb{I}(|\hat{U}_{ij}^{[m]}| > \tau) + \alpha_{ijk}^{2[s]}); \end{aligned}$$

$i, j, k = 1, \dots, p$.

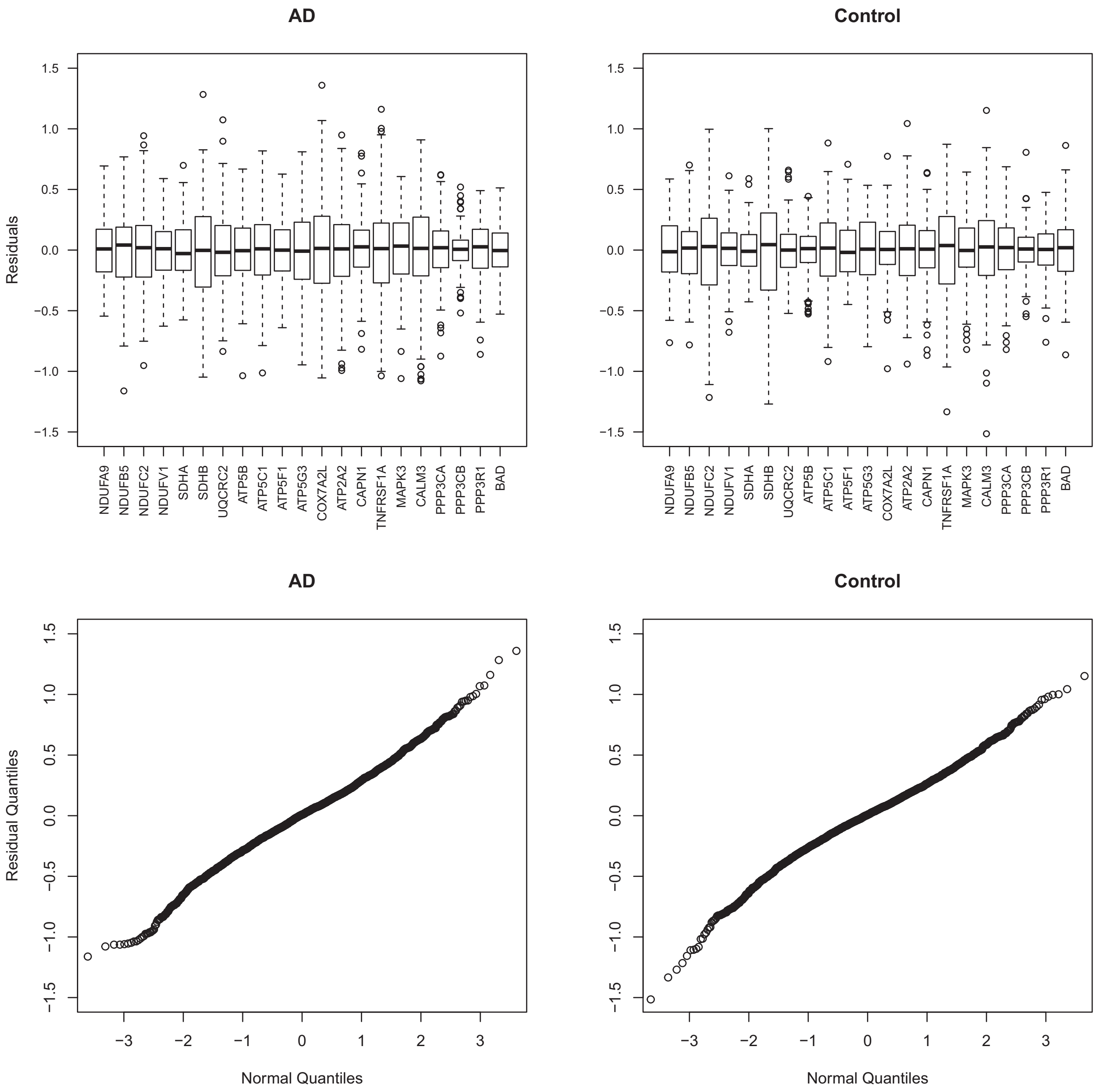

Figure 7. Diagnostic plots of the reconstructed networks. First row: Side-to-side boxplots of residuals of each gene for the AD and control groups. Second row: Normal quantile-quantile plots of residuals of the AD and control groups.

A.2. Technical Proofs

For simplicity, we write $l(\boldsymbol{\Omega})$ as $l(\boldsymbol{U}, \sigma)$ in what follows.

Lemma 1 (Test for linkages). Assume Assumptions 1–3 are met. If $\kappa = |E^0 \setminus D|$ and $\tau \leq C_{\min} c_1/4p$, then under H_0, as $n \to \infty$,

$$\max\left(P(\hat{\boldsymbol{\Omega}}_{H_0} \neq \hat{\boldsymbol{\Omega}}_{E^0}), P(\hat{\boldsymbol{\Omega}}_{H_a} \neq \hat{\boldsymbol{\Omega}}_{E^0 \cup D^0})\right)$$
$$\leq \exp(-c_2 c_4 n C_{\min}/2 + 2\log((p^2 - p + 1) + |D| \log 2 + 3) \to 0.$$

Proof of Lemma 1. Without of loss of generality, we only prove the bound for $P(\hat{\boldsymbol{\Omega}}_{H_a} \neq \hat{\boldsymbol{\Omega}}_{E^0 \cup D^0})$ as the proof for other cases is similar.

Let $S_1^\tau = \{(i,j) \in F^c : |U_{ij}| \geq \tau\}$. When $\kappa = |E^0 \setminus F| = |E^0|$, $\sum_{(i,j)\in F^c} \mathrm{J}_\tau(|U_{ij}|) \leq |E^0|$, so $|F_1^\tau| \leq |E^0|$. If $\hat{E}_{H_a}^\tau \setminus F = E^0$, then $\sum_{(i,j)\in F^c} |\hat{U}_{ij}^{H_a}| \mathbb{I}(|\hat{U}_{ij}^{H_a}| < \tau) = 0$, which implies $\hat{\boldsymbol{U}}^{H_a} = \hat{\boldsymbol{U}}_{E^0 \cup D^0}$. It therefore suffices to prove the case when $\hat{E}_{H_a}^\tau \setminus F \neq E^0$.

Recall that $\boldsymbol{\Omega} = \boldsymbol{\Omega}_S = (\boldsymbol{I} - \boldsymbol{U}_S)^T (\boldsymbol{I} - \boldsymbol{U}_S)/\sigma^2$ for any $S = S_1 \cup S_2$, where $S_1 \subset F^c$ and $S_2 \subset F$. Partition S_1 into two parts, $S_1 = (S_1 \cap E^0) \cup (S_1 \setminus E^0)$. Let $\boldsymbol{k} = (k_1, k_2, k_3)$ and $B_{\boldsymbol{k}} = \{\boldsymbol{\Omega}_\tau : S_1 \neq E^0, |S_1 \cap E^0| = k_1, |S_1 \setminus E^0| = k_2, S_2 \subset D, |F_2| = k_3, c_4(|E^0| - k_1)C_{\min} - c_6 p \tau^{c_5} \leq h^2(\boldsymbol{\Omega}_\tau, \boldsymbol{\Omega}^0)\}$; $k_1 = 0, \ldots, |E^0| - 1$, $k_2 = 1, \ldots, |E^0| - k_1$, $k_3 = 0, \ldots, |F|$. Thus, $B_{\boldsymbol{k}}$ consists of elements with $\binom{|E^0|}{k_1}\binom{p(p-1)-|E^0|}{k_2}\binom{|F|}{k_3}$ different supports. Note that $\{\boldsymbol{\Omega}_\tau : S_1 \neq E^0, |S_1| \leq |E^0|, c_4(|E^0| - k_1)C_{\min} - c_6 p \tau^{c_5} \leq h^2(\boldsymbol{\Omega}_\tau, \boldsymbol{\Omega}^0)\} \subset \bigcup_{k_1=0}^{|E^0|} \bigcup_{k_2=1}^{|E^0|-k_1} \bigcup_{k_3=0}^{|F|} B_{\boldsymbol{k}}$. Then

$$P(\hat{\boldsymbol{\Omega}}_{H_a} \neq \hat{\boldsymbol{\Omega}}_{E^0}) \leq P^*\Big(\sup_{\boldsymbol{\Omega}_{S_1 \cup S_2} : S_1 \neq E^0, |S_1| \leq |E^0|} l(\boldsymbol{\Omega}_{S_1 \cup S_2}) - l(\hat{\boldsymbol{\Omega}}_{E^0}) \geq 0)$$

$$
\begin{aligned}
&\leq P^*\Big(\sup_{\mathbf{\Omega}_{S_1\cup S_2}: S_1\neq E^0, |S_1|\leq |E^0|} l(\mathbf{\Omega}_{S_1\cup S_2}) - l(\mathbf{\Omega}^0) \geq 0\Big)\\
&\leq \sum_{k_1=0}^{|E^0|-1}\sum_{k_2=1}^{|E^0|-k_1}\sum_{k_3=0}^{|F|}\\
&\quad P^*\Big(\sup_{\mathbf{\Omega}_{S_1\cup S_2}\in B_{\boldsymbol{k}}} l(\mathbf{\Omega}_{S_1\cup S_2}) - l(\mathbf{\Omega}^0) \geq 0\Big) \equiv I,
\end{aligned}
$$

where P^* denotes the outer probability.

We apply Theorem 1 of Wong and Shen (1995) to bound I. Toward this end, we verify Condition (3.1) there for the bracketing entropy over $B_{\boldsymbol{k}}$. Let $\mathcal{F}_{\boldsymbol{k}} = \{p^{1/2}(\mathbf{\Omega},\cdot) : \mathbf{\Omega} \in B_{\boldsymbol{k}}\}$, where $p(\mathbf{\Omega},\cdot)$ is the normal density function with mean zero and precision matrix $\mathbf{\Omega}$. Note that

$$
\begin{aligned}
&\int \sup_{\overline{\mathbf{\Omega}}:\|\overline{\mathbf{\Omega}}-\mathbf{\Omega}\|<\delta} |p^{1/2}(\overline{\mathbf{\Omega}},\boldsymbol{x}) - p^{1/2}(\mathbf{\Omega},\boldsymbol{x})|^2 d\boldsymbol{x}\\
&\leq \int \sup_{\overline{\mathbf{\Omega}}:\|\overline{\mathbf{\Omega}}-\mathbf{\Omega}\|<\delta} \left(1-\exp\left(-\frac{1}{2}\boldsymbol{x}^T(\overline{\mathbf{\Omega}}-\mathbf{\Omega})\boldsymbol{x}\right)\right)^2 p(\mathbf{\Omega},\boldsymbol{x})d\boldsymbol{x}\\
&\leq \int \sup_{\overline{\mathbf{\Omega}}:\|\overline{\mathbf{\Omega}}-\mathbf{\Omega}\|<\delta} c''\|\overline{\mathbf{\Omega}}-\mathbf{\Omega}\|^2 p(\mathbf{\Omega},\boldsymbol{x})d\boldsymbol{x} \leq c''\delta^2.
\end{aligned}
$$

By Lemma 2.1 of Ossiander (1987), the bracketing L_2-entropy is bounded by the L_2-metric entropy $H(u,\mathcal{F}_{\boldsymbol{k}})$ of $\mathcal{F}_{\boldsymbol{k}}$. By Kolmogorov and Tikhomirov (1959), for $u \geq \varepsilon^2$,

$$
\begin{aligned}
H(u,\mathcal{F}_{\boldsymbol{k}}) &\leq c_0|S|\log\left(e\frac{p(p-1)}{|S|}\right)\\
&\quad + c_0|S|\log\frac{\min(c_2^{1/2},1)}{u} \leq c_0|S|\log p\log(1/u),
\end{aligned}
$$

where $\binom{a}{b} \leq a^b$ has been applied. Let $\varepsilon = \min(1, \sqrt{2c_0}c_4^{-1}\log(\sqrt{2}/c_3)\sqrt{(|E^0|+|F|)/n})$. Then $\sup_{\boldsymbol{k}}\int_{\varepsilon^2/2^8}^{\sqrt{2}\varepsilon} H^{1/2}(s/c_3,\mathcal{F}_{\boldsymbol{k}})ds \leq \varepsilon\sqrt{2(|E^0|+|F|)}\log(\sqrt{2}/c_3) \leq c_4 n^{1/2}\varepsilon^2$. By Assumption 2, $C_{\min} \geq \varepsilon^2$.

An application of Theorem 1 of Wong and Shen (1995) yields that for some constant $c_2 > 0$,

$$
\begin{aligned}
I &\leq 4\sum_{k_1=0}^{|E^0|-1}\sum_{k_2=1}^{|E^0|-k_1}\sum_{k_3=0}^{|F|}\binom{|E^0|}{k_1}\binom{p(p-1)-|E^0|}{k_2}\binom{|F|}{k_3}\\
&\quad \times \exp\Big(-c_2 n(c_4(|E^0|-k)C_{\min} - c_6 p\tau^{c_5})\Big)\\
&\leq \sum_{k=1}^{|E^0|} 4\exp\Big(-c_2 n(kc_4 C_{\min} - c_6 p\tau^{c_5})\\
&\quad + k(\log|E^0| + \log(p(p-1)-|E^0|+1)) + |F|\log 2\Big)\\
&\leq \sum_{k=1}^{|E^0|} 4\exp\Big(-kc_2c_4 nC_{\min}/2\\
&\quad + k(\log|E^0| + \log(p(p-1)-|E^0|+1)) + |F|\log 2\Big)\\
&\leq \exp(-c_2c_4nC_{\min}/2 + 2\log((p^2-p+1) + |F|\log 2 + 3).
\end{aligned}
$$

Hence, $P(\hat{\mathbf{\Omega}}_{H_a} \neq \hat{\mathbf{\Omega}}_{E^0\cup D^0}) \leq \exp(-c_2c_4nC_{\min}/2 + 2\log((p^2-p+1) + |F|\log 2 + 3)$.

Similarly, we bound $P(\hat{\mathbf{\Omega}}_{H_0} \neq \hat{\mathbf{\Omega}}_{E^0})$. This completes the proof. □

Proof of Theorem 1. By Lemma 1, under H_0 and Assumptions 1–3, $P(\hat{\boldsymbol{U}}^{H_0} = \hat{\boldsymbol{U}}_{E^0}, \hat{\boldsymbol{U}}^{H_a} = \hat{\boldsymbol{U}}_{D^0\cup E^0}) \to 1$, implying that $P(\hat{D}^0 = D^0) \to 1$. When $D^0 = \emptyset$, then $P(Lr = 0) \to 1$, establishing (i).

Now consider the case of $D^0 \neq \emptyset$. For (ii) and (iii), it suffices to focus on event $\{\hat{\boldsymbol{U}}^{H_0} = \hat{\boldsymbol{U}}_{E^0}, \hat{\boldsymbol{U}}^{H_a} = \hat{\boldsymbol{U}}_{D^0\cup E^0}\}$, where $\hat{\boldsymbol{U}}_S$ denotes the maximizer of the likelihood $l(\boldsymbol{U},\sigma)$ with the support index S. Let $\boldsymbol{U}_{\cdot j}$ denote the jth column of matrix $\boldsymbol{U}$. Then $\hat{\boldsymbol{U}}_{\cdot j}^{H_0} = (\boldsymbol{Y}^T_{\mathrm{pa}_j^0}\boldsymbol{Y}_{\mathrm{pa}_j^0})^{-1}\boldsymbol{Y}^T_{\mathrm{pa}_j^0}\boldsymbol{Y}_j$ and $\hat{\boldsymbol{U}}_{\cdot j}^{H_a} = (\boldsymbol{Y}^T_{\mathrm{pa}_j^1}\boldsymbol{Y}_{\mathrm{pa}_j^1})^{-1}\boldsymbol{Y}^T_{\mathrm{pa}_j^1}\boldsymbol{Y}_j$, where pa_j^0 and pa_j^1 denote the parent variables of Y_j in E^0 and $E^0 \cup D^0$, respectively. Note that

$$
\begin{aligned}
&\log\frac{\sum_{j=1}^p\|\boldsymbol{Y}_j - \boldsymbol{Y}\boldsymbol{U}_{\cdot j}\|^2}{\sum_{j=1}^p\|\boldsymbol{Y}_j - \boldsymbol{Y}\boldsymbol{U}^0_{\cdot j}\|^2}\\
&\quad = \log\left(1 - \frac{\sum_{j=1}^p\{2\boldsymbol{e}_j^T\boldsymbol{Y}(\boldsymbol{U}_{\cdot j}-\boldsymbol{U}^0_{\cdot j}) - \|\boldsymbol{Y}(\boldsymbol{U}_{\cdot j}-\boldsymbol{U}^0_{\cdot j})\|^2\}}{\sum_{j=1}^p\|\boldsymbol{e}_j\|^2}\right)\\
&\quad = -\sum_{k=1}^{\infty}\frac{1}{k}\left(\frac{\sum_{j=1}^p\{2\boldsymbol{e}_j^T\boldsymbol{Y}(\boldsymbol{U}_{\cdot j}-\boldsymbol{U}^0_{\cdot j}) - \|\boldsymbol{Y}(\boldsymbol{U}_{\cdot j}-\boldsymbol{U}^0_{\cdot j})\|^2\}}{\sum_{j=1}^p\|\boldsymbol{e}_j\|^2}\right)^k,
\end{aligned}
$$

provided that $\{\sum_{j=1}^p(2\boldsymbol{e}_j^T\boldsymbol{Y}(\boldsymbol{U}_{\cdot j}-\boldsymbol{U}^0_{\cdot j}) - \|\boldsymbol{Y}(\boldsymbol{U}_{\cdot j}-\boldsymbol{U}^0_{\cdot j})\|^2) < \sum_{j=1}^p\|\boldsymbol{e}_j\|^2\}$. On the event $\bigcap_{j=1}^p\{\boldsymbol{e}_j^T\boldsymbol{P}_{\mathrm{pa}_j^1}\boldsymbol{e}_j < \|\boldsymbol{e}_j\|^2\}$, we have that

$$
\begin{aligned}
2Lr &= np\log\frac{\sum_{j=1}^p\|\boldsymbol{Y}_j - \boldsymbol{Y}\hat{\boldsymbol{U}}^{H_0}_{\cdot j}\|^2}{\sum_{j=1}^p\|\boldsymbol{Y}_j - \boldsymbol{Y}\boldsymbol{U}^0_{\cdot j}\|^2} - np\log\frac{\sum_{j=1}^p\|\boldsymbol{Y}_j - \boldsymbol{Y}\hat{\boldsymbol{U}}^{H_a}_{\cdot j}\|^2}{\sum_{j=1}^p\|\boldsymbol{Y}_j - \boldsymbol{Y}\boldsymbol{U}^0_{\cdot j}\|^2}\\
&= \frac{1}{\sigma^2}\sum_{j=1}^p\boldsymbol{e}_j^T\{\boldsymbol{P}_{\mathrm{pa}_j^1} - \boldsymbol{P}_{\mathrm{pa}_j^0}\}\boldsymbol{e}_j + \Big(\frac{np}{\sum_{j=1}^p\|\boldsymbol{e}_j\|^2} - \frac{1}{\sigma^2}\Big)\\
&\quad \sum_{j=1}^p\boldsymbol{e}_j^T\{\boldsymbol{P}_{\mathrm{pa}_j^1} - \boldsymbol{P}_{\mathrm{pa}_j^0}\}\boldsymbol{e}_j\\
&\quad - np\sum_{k=2}^{\infty}\frac{1}{k}\left[\left(\frac{\sum_{j=1}^p\boldsymbol{e}_j^T\boldsymbol{P}_{\mathrm{pa}_j^1}\boldsymbol{e}_j}{\sum_{j=1}^p\|\boldsymbol{e}_j\|^2}\right)^k - \left(\frac{\sum_{j=1}^p\boldsymbol{e}_j^T\boldsymbol{P}_{\mathrm{pa}_j^0}\boldsymbol{e}_j}{\sum_{j=1}^p\|\boldsymbol{e}_j\|^2}\right)^k\right]\\
&\equiv T_1 + T_2 + T_3,
\end{aligned}
$$

where $\boldsymbol{P}_S = \boldsymbol{Y}_S(\boldsymbol{Y}_S^T\boldsymbol{Y}_S)^{-1}\boldsymbol{Y}_S^T$ is the projection over subset S of predictors.

Let $Y_{j_1} \preceq \cdots \preceq Y_{j_p}$ be a partial order of DAG G^0. Taking the iterated expectation, we simplify the characteristic function $\phi(t)$ of T_1 for any real t:

$$
\begin{aligned}
\phi(t) &= E\left\{E\left[\exp\left(it\sigma^{-2}\sum_{j=1}^p\boldsymbol{e}_j^T\{\boldsymbol{P}_{\mathrm{pa}_j^1} - \boldsymbol{P}_{\mathrm{pa}_j^0}\}\boldsymbol{e}_j\right)\middle|\boldsymbol{Y}_{\{1,\ldots,p\}\backslash\{j_p\}}\right]\right\}\\
&= (1-2it)^{-\frac{|\mathrm{pa}^1_{j_p}|-|\mathrm{pa}^0_{j_p}|}{2}}\\
&\quad E\left(\exp\left(it\sigma^{-2}\sum_{j=1}^{p-1}\boldsymbol{e}_j^T\{\boldsymbol{P}_{\mathrm{pa}_j^1} - \boldsymbol{P}_{\mathrm{pa}_j^0}\}\boldsymbol{e}_j\right)\right) = (1-2it)^{-\frac{|D^0|}{2}}.
\end{aligned}
$$

Thus, T_1 follows $\chi^2_{|D^0|}$. Moreover, it follows immediately that $\boldsymbol{e}_j^T\boldsymbol{P}_{\mathrm{pa}_j^1}\boldsymbol{e}_j/\sigma^2$ are $\chi^2_{|\mathrm{pa}_j^1|}$ distributed.

For T_2, we have $\sqrt{np}\Big(\frac{np}{\sum_{j=1}^p\|\boldsymbol{e}_j\|^2} - \frac{1}{\sigma^2}\Big) \xrightarrow{d} N(0, 2\sigma^{-4})$. Then it follows from Assumption 4 that $T_2/\sqrt{|D^0|} \xrightarrow{p} 0$.

To bound T_3, note that, there exists some constant $C' > 0$ such that as $n \to \infty$,

$$|T_3| \le C'T_1 \sum_{k=1}^{\infty} \left(\frac{\sum_{j=1}^p \boldsymbol{e}_j^T \boldsymbol{P}_{\mathrm{pa}_j^1} \boldsymbol{e}_j}{\sum_{j=1}^p \|\boldsymbol{e}_j\|^2} \right)^k \le C'T_1 \sum_{k=1}^{\infty} \left(\max_{1 \le j \le p} \frac{\boldsymbol{e}_j^T \boldsymbol{P}_{\mathrm{pa}_j^1} \boldsymbol{e}_j}{\|\boldsymbol{e}_j\|^2} \right)^k$$
$$= C'T_1 \left(\max_{1 \le j \le p} \frac{\boldsymbol{e}_j^T \boldsymbol{P}_{\mathrm{pa}_j^1} \boldsymbol{e}_j}{\|\boldsymbol{e}_j\|^2} \right) \left(1 - \max_{1 \le j \le p} \frac{\boldsymbol{e}_j^T \boldsymbol{P}_{\mathrm{pa}_j^1} \boldsymbol{e}_j}{\|\boldsymbol{e}_j\|^2} \right)^{-1},$$

on an event $\bigcap_{j=1}^p \{\boldsymbol{e}_j^T \boldsymbol{P}_{\mathrm{pa}_j^1} \boldsymbol{e}_j < \|\boldsymbol{e}_j\|^2\}$, where the inequality $\frac{a_1+\cdots+a_p}{b_1+\cdots+b_p} \le \max_{1 \le j \le p} \frac{a_j}{b_j}$ is used; $0 \le a_j \le b_j; j = 1, \ldots, p$. By Lemma 1 of Laurent and Massart (2000), for any $s > 0$, $P(\chi_r^2 > r + 2\sqrt{rs} + 2s) \le e^{-s}$ and $P(\chi_r^2 < r - 2\sqrt{rs}) \le e^{-s}$. Set $s = nt^2/16\sigma^4 < n$ for any $0 < t < 4\sigma^2$. By Assumption 4,

$$P\Big(\max_{1 \le j \le p} \left| \frac{\|\boldsymbol{e}_j\|^2}{n} - \sigma^2 \right| > t \Big)$$
$$\le P\Big(\exists j : |\|\boldsymbol{e}_j\|^2 - n\sigma^2| > 4\sigma^2 \sqrt{ns} \Big) \le \exp(2(\log p - s)) \to 0,$$

as $n \to \infty$. Hence, $\|\boldsymbol{e}_j\|^2/n \xrightarrow{p} \sigma^2$. An application of Lemma 1 of Laurent and Massart (2000) with $s = (|E^0| + |D^0|) \log p$ leads to

$$P\Big(\max_{1 \le j \le p} \boldsymbol{e}_j^T \boldsymbol{P}_{\mathrm{pa}_j^1} \boldsymbol{e}_j > 5\sigma^2 s \Big)$$
$$\le P\Big(\exists j : \boldsymbol{e}_j^T \boldsymbol{P}_{\mathrm{pa}_j^1} \boldsymbol{e}_j > \sigma^2 (|\mathrm{pa}_j^1| + 2\sqrt{|\mathrm{pa}_j^1| s} + 2s) \Big)$$
$$\le \exp(\log p - s).$$

By Assumption 4, $\max_{1 \le j \le p} \boldsymbol{e}_j^T \boldsymbol{P}_{\mathrm{pa}_j^1} \boldsymbol{e}_j = O_p((|E^0| + |F|) \log p)$. It follows that

$$|T_3| \le C'T_1 \max_{1 \le j \le p} \left(\frac{\boldsymbol{e}_j^T \boldsymbol{P}_{\mathrm{pa}_j^1} \boldsymbol{e}_j}{n} \right) \left(\frac{n}{\|\boldsymbol{e}_j\|^2} \right)$$
$$= O_p \left(|D^0|(|D^0| + |E^0|) \frac{\log p}{n} \right).$$

Hence, $T_3/\sqrt{|D^0|} \xrightarrow{p} 0$.

Finally, note that $P(\exists j : \boldsymbol{e}_j^T \boldsymbol{P}_{\mathrm{pa}_j^1} \boldsymbol{e}_j \ge \|\boldsymbol{e}_j\|^2) \to 0$. Therefore, when $|D^0|$ is fixed, $Lr \xrightarrow{d} \chi^2_{|D^0|}$; when $|D^0| \to \infty$, $(2|D^0|)^{-1/2}(Lr - |D^0|) \xrightarrow{d} N(0, 1)$. This completes the proof. □

Lemma 2 (Test of pathway). Assume Assumptions 1–3 are met. Let $A = \{(i_k, i_{k+1}) \in F : U^0_{i_k, i_{k+1}} = 0\}$ and $F_{-k} = F \backslash \{(i_k, i_{k+1})\}$. If $\kappa = |E^0 \backslash F|$ and $\tau = C_{\min} c_1 / 4p$, then under H_0,

$$P\Big(\max_{k:(i_k, i_{k+1}) \notin A} l(\hat{\boldsymbol{\Omega}}(k)) - l(\boldsymbol{\Omega}^0) \ge 0 \Big) \to 0,$$
$$\max \Big(P(\exists k : \hat{E}_{H_0}(k) \ne E^0 \cup F_{-k}; (i_k, i_{k+1}) \in A),$$
$$P(\hat{E}_{H_a} \ne E^0 \cup F) \Big) \to 0.$$

Proof of Lemma 2. The basic idea of the proof is similar to that of Lemma 1.

Recall that $\boldsymbol{\Omega}_S = (\boldsymbol{I} - \boldsymbol{U}_S)^T (\boldsymbol{I} - \boldsymbol{U}_S)/\sigma^2$ for any S that forms a DAG. Now, partition S into four parts, $S = E_1 \cup E_2 \cup E_3 \cup E_4$, where $E_1 = S_1 \cap E^0$, $E_2 = S_1 \backslash F$, $E_3 = S_2 \cap E^0$, and $E_4 = S_2 \backslash F$, where S_1 and S_2 are defined the same as in Assumption 2. Let $\kappa_1 = |E^0 \backslash F|$ and $\kappa_2 = |E^0 \cap F|$. Let $\boldsymbol{k} = (k_1, k_2, k_3, k_4)$ and $B_{\boldsymbol{k}} = \{\boldsymbol{\Omega}_\tau : S_1 \cup S_2 \not\supset E_0, |S_1| \le |E^0 \setminus F|, |E_i| = k_i; i = 1, 2, 3, 4, c_4(|E^0| - k_1 - k_3)C_{\min} - c_6 p \tau^{c_5} \le h^2(\boldsymbol{\Omega}_\tau, \boldsymbol{\Omega}^0)\}$, where $0 \le k_1 \le \kappa_1$, $0 \le k_2 \le \kappa_1 - k_1$, $0 \le k_3 \le \kappa_2$, $0 \le k_4 \le |F| - \kappa_2$, and $k_1 + k_3 < \kappa_1 + \kappa_2 = |E^0|$. Thus, $B_{\boldsymbol{k}}$ contains elements with $\binom{\kappa_1}{k_1}\binom{p(p-1)-|F|-\kappa_1}{k_2}\binom{\kappa_2}{k_3}\binom{|F|-\kappa_2}{k_4}$ different supports. Note that $\{\boldsymbol{\Omega}_\tau : S_1 \cup S_2 \not\supset E_0, |S_1| \le |E^0 \setminus F|, c_4(|E^0| - k_1 - k_3)C_{\min} - c_6 p \tau^{c_5} \le h^2(\boldsymbol{\Omega}_\tau, \boldsymbol{\Omega}^0)\} \subset \bigcup_{\boldsymbol{k}} B_{\boldsymbol{k}}$. Then

$$P\Big(\max_{k:(i_k, i_{k+1}) \notin A} l(\hat{\boldsymbol{\Omega}}(k)) - l(\boldsymbol{\Omega}^0) \ge 0 \Big)$$
$$\le \sum_{\boldsymbol{k}} P^*(\sup_{\boldsymbol{\Omega}_{S_1 \cup S_2} \in B_{\boldsymbol{k}}} l(\boldsymbol{\Omega}_{S_1 \cup S_2}) - l(\boldsymbol{\Omega}^0) \ge 0) \equiv I',$$
$$P(\hat{E}_{H_a} \ne E^0 \cup F) \le I',$$
$$P(\exists k : \hat{E}_{H_0}(k) \ne E^0 \cup F_{-k}; (i_k, i_{k+1}) \in A) \le I'.$$

To apply Theorem 1 of Wong and Shen (1995) to bound I', we verify Condition (3.1) of Wong and Shen (1995) using the similar argument of Lemma 1. Let $\varepsilon = \min(1, \sqrt{2c_0} c_4^{-1} \log(\sqrt{2}/c_3) \sqrt{(|E^0| + |F|)/n})$. Then

$$\sup_{\boldsymbol{k}} \int_{\varepsilon^2/2^8}^{\sqrt{2}\varepsilon} H^{1/2}(s/c_3, \mathcal{F}_{\boldsymbol{k}}) ds$$
$$\le \varepsilon \sqrt{2(|E^0| + |D|) \log(\sqrt{2}/c_3)} \le c_4 n^{1/2} \varepsilon^2.$$

By Assumption 5, $C_{\min} \ge \varepsilon^2$.

By Theorem 1 of Wong and Shen (1995), we have, for some constant $c_2 > 0$,

$$I' \le \sum_{\boldsymbol{k}} \binom{\kappa_1}{k_1} \binom{p(p-1) - |F| - \kappa_1}{k_2} \binom{\kappa_2}{k_3} \binom{|F| - \kappa_2}{k_4}$$
$$\times 4 \exp(-c_2 n (c_4(|E^0| - k_1 - k_3)C_{\min} - c_6 p \tau^{c_5}))$$
$$\le \sum_{k=1}^{|E^0|} 2^{|F|+2} \exp(-k(c_2 c_4 n C_{\min}/2 - \log|E^0|$$
$$- 2\log(p(p-1) - |F|)))$$
$$\le 5 \exp(-c_2 c_4 n C_{\min}/2 + \log|E^0| + 2\log(p(p-1) - |F|)$$
$$+ |F| \log 2) \to 0.$$

Then desired result follows from the bound of I'. This completes the proof. □

Proof of Theorem 2. It follows from Lemma 2 that $P(\hat{E}_{H_a} \setminus F = E^0 \setminus F) \to 1$ and $P(\hat{E}_{H_0}(k) \setminus F = E^0 \setminus F) \to 1$ for k such that $U^0_{i_k, i_{k+1}} = 0$. Consider the event $\{\hat{E}_{H_a} \setminus F = E^0 \setminus F\}$.

For (i), suppose $E^0 \cup F$ does not form a DAG. Then there exists $1 \le k \le |F|$ such that $(\hat{\boldsymbol{U}}^{H_a})_{i_k, i_{k+1}} = 0$. Thus, $l(\hat{\boldsymbol{\Omega}}_{H_a}) = l(\hat{\boldsymbol{\Omega}}_{H_0}(k))$, establishing (i).

For (ii) and (iii), suppose $E^0 \cup F$ forms a DAG. Then for $(i_k, i_{k+1}) \in E^0 \cap F$ and $(i_{k'}, i_{k'+1}) \in F \backslash E^0$, we have $P\Big(l(\hat{\boldsymbol{\Omega}}_{H_0}(k)) < l(\hat{\boldsymbol{\Omega}}_{H_0}(k')) \Big) \to 1$. It suffices to consider edges $A = \{(i_k, i_{k+1}) \in F : U^0_{i_k, i_{k+1}} = 0\}$. Recall that pa_j^1 denotes the parent variables of Y_j in $\hat{E}_{H_a}$. Similarly as in the proof of Theorem 1, for any $(i_k, i_{k+1}) \in A$,

$$2\Big(l(\hat{\boldsymbol{\Omega}}_{H_a}) - l(\hat{\boldsymbol{\Omega}}_{H_0}(k)) \Big) = \frac{1}{\sigma^2} \boldsymbol{e}_{i_{k+1}}^T \big(\boldsymbol{P}_{\mathrm{pa}^1_{i_{k+1}}} - \boldsymbol{P}_{\mathrm{pa}^1_{i_{k+1}} \backslash \{i_k\}} \big) \boldsymbol{e}_{i_{k+1}}$$
$$+ \left(\frac{np}{\sum_{j=1}^p \|\boldsymbol{e}_j\|^2} - \frac{1}{\sigma^2} \right) \boldsymbol{e}_{i_{k+1}}^T \big(\boldsymbol{P}_{\mathrm{pa}^1_{i_{k+1}}} - \boldsymbol{P}_{\mathrm{pa}^1_{i_{k+1}} \backslash \{i_k\}} \big) \boldsymbol{e}_{i_{k+1}}$$
$$+ np \sum_{k=2}^{\infty} \frac{1}{k} \left[\left(\frac{\boldsymbol{e}_{i_{k+1}}^T \boldsymbol{P}_{\mathrm{pa}^1_{i_{k+1}}} \boldsymbol{e}_{i_{k+1}}}{\sum_{j=1}^p \|\boldsymbol{e}_j\|^2} \right)^k - \left(\frac{\boldsymbol{e}_{i_{k+1}}^T \boldsymbol{P}_{\mathrm{pa}^1_{i_{k+1}} \backslash \{i_k\}} \boldsymbol{e}_{i_{k+1}}}{\sum_{j=1}^p \|\boldsymbol{e}_j\|^2} \right)^k \right]$$

$$\equiv T_1(k) + T_2(k) + T_3(k).$$

Let ψ be the characteristic function of the joint distribution of $T_1(k)$; $(i_k, i_{k+1}) \in A$. Taking the iterated expectation, as in the proof of Theorem 1, we factorize ψ as follows: $\psi(\boldsymbol{t}) = \prod_{(i_k,i_{k+1})\in A}(1 - 2it_k)^{-1/2}$. It follows immediately that $T_1(k)$; $(i_k, i_{k+1}) \in A$ are asymptotically independently χ_1^2 distributed.

To bound $\max_k |T_2(k)|$, note that $\sqrt{np}\Big(\frac{np}{\sum_{j=1}^p \|\boldsymbol{e}_j\|^2} - \frac{1}{\sigma^2}\Big) \xrightarrow{d} N(0, 2\sigma^{-4})$. Similarly as in the proof of Theorem 1, we have $\max_k T_1(k) = O_p(\log p)$. Hence, $\max_k d^2|T_2(k)| \xrightarrow{p} 0$.

It remains to bound $|T_3(k)|$. By Lemma 1 of Laurent and Massart (2000), $\|\boldsymbol{e}_j\|^2/n \xrightarrow{p} \sigma^2$ provided that $\frac{\log p}{n} \to 0$. Consequently, $|T_3(k)| \leq C'|T_1(k)|\sum_{l=1}^{\infty}\left(\max_k \frac{\boldsymbol{e}_{i_{k+1}}^T \boldsymbol{P}_{\mathrm{pa}^1_{i_{k+1}}} \boldsymbol{e}_{i_{k+1}}}{\|\boldsymbol{e}_{i_{k+1}}\|^2}\right)^l \equiv |T_1(k)|\Delta$ for some constant $C' > 0$, where $\Delta = O_p\Big(|E^0|\frac{\log p}{n}\Big)$ when $\frac{\log p}{n} \to 0$.

When $d = |A|$ is fixed, $2Lr \xrightarrow{d} \Gamma_d$ and Γ_d has the same distribution as $\min\{X_1, \ldots, X_d\}$.

If $d \to \infty$, then

$$\begin{aligned}(1-\Delta)d^2 \min_k T_1(k) - d^2 \max_k |T_2(k)| &\leq 2d^2 Lr\\ &\leq (1+\Delta)d^2 \min_k T_1(k) + d^2 \max_k |T_2(k)|.\end{aligned}$$

By Assumption 6, $\Delta \xrightarrow{p} 0$ and $\max_k d^2|T_2(k)| \xrightarrow{p} 0$. Thus, it suffices to show that $\min_k d^2 T_1(k) \xrightarrow{d} \Gamma$. Denote F_{χ^2} as the distribution function of χ_1^2 random variable. Then for every $x > 0$, $\lim_{d\to\infty} d\log\Big(1 - F_{\chi^2}\left(x/d^2\right)\Big) = -\lim_{d\to\infty} d\sum_{k=1}^{\infty}\frac{F^k_{\chi^2}(x/d^2)}{k} = \sqrt{\frac{2x}{\pi}}$. Hence, the limiting distribution function of $2d^2Lr$ is $(1 - \exp(-\sqrt{2x/\pi}))\mathbb{I}(x > 0)$. This completes the proof. □

Lemma 3. Assume that $q < n$ is allowed to depend on n and $\boldsymbol{Y}_{i\cdot} \sim N(\boldsymbol{0}, \boldsymbol{\Omega}^{-1})$ be independently and identically distributed vectors in $\mathbb{R}^q$; $i = 1, \ldots, n$. If Assumption 1 is satisfied, then, for any $0 < \delta < 1$, there exists a constant $c > 0$ such that

$$\begin{aligned}P\Big(\exists S : \|n^{-1}\boldsymbol{Y}_S^T(\boldsymbol{I} - \boldsymbol{P}_{S^c})\boldsymbol{Y}_S - (\boldsymbol{\Omega}_{S,S})^{-1}\|_2\\ > c\|\boldsymbol{\Omega}^{-1}\|_2 \log\left(\frac{2}{\delta}\right)\sqrt{\frac{q}{n}}\Big) \leq \delta,\end{aligned}$$

where $S \subset \{1, \ldots, q\}$ and $\|\cdot\|_2$ denotes the matrix 2-norm.

Proof of Lemma 3. First, it follows from Proposition 2.1 of Vershynin (2012) that $P(\|n^{-1}\boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{\Omega}^{-1}\|_2 > c\log\left(\frac{2}{\delta}\right)\sqrt{\frac{q}{n}}) \leq \delta$. Let $S \subset \{1, \ldots, q\}$. Note that

$$\begin{aligned}&n^{-1}\boldsymbol{Y}_S^T(\boldsymbol{I} - \boldsymbol{P}_{S^c})\boldsymbol{Y}_S = n^{-1}(\boldsymbol{Y}_S^T\boldsymbol{Y}_S - \boldsymbol{Y}_S^T\boldsymbol{Y}_{S^c}(\boldsymbol{Y}_{S^c}^T\boldsymbol{Y}_{S^c})^{-1}\boldsymbol{Y}_{S^c}^T\boldsymbol{Y}_S)\\ &= ((\boldsymbol{\Omega}^{-1})_{SS} + \boldsymbol{\Delta}_{SS}) - ((\boldsymbol{\Omega}^{-1})_{SS^c} + \boldsymbol{\Delta}_{SS^c})\\ &((\boldsymbol{\Omega}^{-1})_{S^cS^c} + \boldsymbol{\Delta}_{S^cS^c})^{-1}((\boldsymbol{\Omega}^{-1})_{S^cS} + \boldsymbol{\Delta}_{S^cS})\\ &= (\boldsymbol{\Omega}^{-1})_{SS} - (\boldsymbol{\Omega}^{-1})_{SS^c}(\boldsymbol{\Omega}^{-1})_{S^cS^c}(\boldsymbol{\Omega}^{-1})_{S^cS}\\ &+ O(\|\boldsymbol{\Omega}^{-1}\|_2\boldsymbol{\Delta}) = (\boldsymbol{\Omega}_{SS})^{-1} + O(\|\boldsymbol{\Omega}^{-1}\|_2\boldsymbol{\Delta}),\end{aligned}$$

where $\boldsymbol{\Delta}_{SS} = n^{-1}\boldsymbol{Y}^T\boldsymbol{Y} - (\boldsymbol{\Omega}^{-1})_{SS}$, $\boldsymbol{\Delta}_{SS^c} = n^{-1}\boldsymbol{Y}_S^T\boldsymbol{Y}_{S^c} - (\boldsymbol{\Omega}^{-1})_{SS^c} = \boldsymbol{\Delta}_{S^cS}^T$ and $\boldsymbol{\Delta} = n^{-1}\boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{\Omega}^{-1}$. The fact that $\max(\|\boldsymbol{M}_{SS}\|_2, \|\boldsymbol{M}_{SS^c}\|_2) \leq \|\boldsymbol{M}\|_2$ for any $\boldsymbol{M} \in \mathbb{R}^{q\times q}$ has been applied. The desired result follows immediately. □

Proof of Theorem 3. Without loss of generality, assume $\sigma = 1$ and $Y_1 \preceq \cdots \preceq Y_p$ is the partial order of DAG G^0.

For (i), if $|D^0| = 0$, then for any $E \subset F$, $E^0 \cup E$ is cyclic. However, since $\boldsymbol{U}^n$ is acyclic, then $\boldsymbol{U}^n = \boldsymbol{U}^0$ and $h = 0$, establishing (i).

For (ii), suppose $|D^0| > 0$ is fixed. Define $D_j^0 = \{(i,j) \in D^0\}$ and note that

$$\begin{aligned}l(\boldsymbol{U}^n, \sigma^2) - l(\boldsymbol{U}^0, \sigma^2) &= \sum_{j=1}^p \boldsymbol{\eta}_{D_j^0}^T\sqrt{n}\mathrm{vec}_{D_j^0}(\boldsymbol{\delta})\\ &- \frac{1}{2}\sum_{j=1}^p \sqrt{n}\mathrm{vec}_{D_j^0}(\boldsymbol{\delta})^T\left(\frac{1}{n}\boldsymbol{Y}_{\mathrm{pa}_j^1}^T\boldsymbol{Y}_{\mathrm{pa}_j^1}\right)_{D_j^0,D_j^0}\\ &\sqrt{n}\mathrm{vec}_{D_j^0}(\boldsymbol{\delta}),\end{aligned}$$

where $\boldsymbol{\eta}_{D_j^0} = n^{-1/2}\frac{\partial l(\boldsymbol{U},\sigma)}{\partial \boldsymbol{U}_{D_j^0}}\ |_{\boldsymbol{U}=\boldsymbol{U}^0} = n^{-1/2}\boldsymbol{Y}^T(\boldsymbol{Y}_j - \boldsymbol{Y}\boldsymbol{U}^0_{\cdot j}) = n^{-1/2}\boldsymbol{Y}_{D_j^0}^T\boldsymbol{e}_j$. The local power of the proposed test is

$$\begin{aligned}&P_{\boldsymbol{U}^n}(2Lr > \chi^2_{|D^0|,1-\alpha})\\ &= E_{\boldsymbol{U}^0}\Big(\mathbb{I}(2Lr > \chi^2_{|D^0|,1-\alpha})\exp(l(\boldsymbol{U}^n, \sigma^2) - l(\boldsymbol{U}^0, \sigma^2))\Big)\\ &= E_{\boldsymbol{U}^0}\Bigg(\mathbb{I}\left(2Lr > \chi^2_{|D^0|,1-\alpha}\right)\exp\Bigg(\sum_{j=1}^p \boldsymbol{\eta}_{D_j^0}^T\sqrt{n}\mathrm{vec}_{D_j^0}(\boldsymbol{\delta}) -\\ &\frac{1}{2}\sum_{j=1}^p \sqrt{n}\mathrm{vec}_{D_j^0}(\boldsymbol{\delta})^T\left(\frac{1}{n}\boldsymbol{Y}_{\mathrm{pa}_j^1}^T\boldsymbol{Y}_{\mathrm{pa}_j^1}\right)_{D_j^0,D_j^0}\sqrt{n}\mathrm{vec}_{D_j^0}(\boldsymbol{\delta})\Bigg)\Bigg).\end{aligned}$$

Under the assumptions of Theorem 1, with probability tending to 1 under $P_{\boldsymbol{U}^0}$, we have $2Lr = \sum_{j=1}^p \boldsymbol{e}_j^T(\boldsymbol{P}_{\mathrm{pa}_j^1} - \boldsymbol{P}_{\mathrm{pa}_j^0})\boldsymbol{e}_j + o_p(1) = \sum_{j=1}^p\sum_{r=1}^{|D_j^0|}(\boldsymbol{a}_{jr}^T\boldsymbol{e}_j)^2 + o_p(1)$, where $\boldsymbol{P}_{\mathrm{pa}_j^1} - \boldsymbol{P}_{\mathrm{pa}_j^0} = \sum_{r=1}^{|D_j^0|}\boldsymbol{a}_{jr}\boldsymbol{a}_{jr}^T$. Let $\boldsymbol{A}(j) = (\boldsymbol{a}_{j1}, \ldots, \boldsymbol{a}_{j,|D_j^0|})^T \in \mathbb{R}^{|D_j^0|\times n}$; $j = 1, \ldots, p$. Then

$$\begin{aligned}&\begin{pmatrix}\boldsymbol{A}(j)\boldsymbol{e}_j\\ \boldsymbol{\eta}_{D_j^0}\end{pmatrix}\Bigg|\boldsymbol{Y}_{\{1,\ldots,j-1\}}\\ &\equiv \begin{pmatrix}\boldsymbol{Z}_1(j)\\ \boldsymbol{Z}_2(j)\end{pmatrix} \sim N\left(\boldsymbol{0}, \begin{pmatrix}\boldsymbol{I}_{|D^0|\times|D^0|} & n^{-1/2}\boldsymbol{A}(j)\boldsymbol{Y}_{D_j^0}\\ n^{-1/2}\boldsymbol{Y}_{D_j^0}^T\boldsymbol{A}(j)^T & n^{-1}\boldsymbol{Y}_{D_j^0}\boldsymbol{Y}_{D_j^0}\end{pmatrix}\right).\end{aligned}$$

By simple matrix manipulations, $\boldsymbol{Z}_2 \mid \boldsymbol{Z}_1 = \boldsymbol{z}_1 \sim N\Big(n^{-1/2}\boldsymbol{Y}_{D_j^0}^T\boldsymbol{A}(j)^T\boldsymbol{z}_1, n^{-1}\boldsymbol{Y}_{D_j^0}^T(\boldsymbol{I}_{n\times n} - \boldsymbol{A}(j)^T\boldsymbol{A}(j))\boldsymbol{Y}_{D_j^0}\Big)$. After changing the measure, we have that $\liminf_{n\to\infty} P_{\boldsymbol{U}^n}\left(2Lr > \chi^2_{|D^0|,1-\alpha}\right)$ is lower-bounded by

$$\begin{aligned}&\liminf_{n\to\infty} E_{\boldsymbol{U}^0}\Bigg(\mathbb{I}\left(\sum_{j=1}^p\|\boldsymbol{A}(j)\boldsymbol{e}_j\|^2 > \chi^2_{|D^0|,1-\alpha}\right)\\ &\times\exp\Bigg(\sum_{j=1}^p \boldsymbol{\eta}_{D_j^0}^T\sqrt{n}\mathrm{vec}_{D_j^0}(\boldsymbol{\delta}) - \frac{1}{2}\\ &\sum_{j=1}^p \sqrt{n}\mathrm{vec}_{D_j^0}(\boldsymbol{\delta})^T\left(\frac{1}{n}\boldsymbol{Y}_{\mathrm{pa}_j^1}^T\boldsymbol{Y}_{\mathrm{pa}_j^1}\right)_{D_j^0,D_j^0}\sqrt{n}\mathrm{vec}_{D_j^0}(\boldsymbol{\delta})\Bigg)\Bigg)\\ &= \liminf_{n\to\infty}\\ &E\Bigg[\exp\Bigg(-\frac{1}{2}\sum_{j=1}^p \sqrt{n}\mathrm{vec}_{D_j^0}(\boldsymbol{\delta})^T\left(\frac{1}{n}\boldsymbol{Y}_{\mathrm{pa}_j^1}^T\boldsymbol{Y}_{\mathrm{pa}_j^1}\right)_{D_j^0,D_j^0}\sqrt{n}\mathrm{vec}_{D_j^0}(\boldsymbol{\delta})\Bigg)\end{aligned}$$

$$\times E\Big(\mathbb{I}\Big(\|\boldsymbol{Z}_1\|^2 + \sum_1^{p-1} \|\boldsymbol{A}(j)\boldsymbol{e}_j\|^2 > \chi^2_{|D^0|,1-\alpha}\Big)$$

$$E\Big(\exp(\sqrt{n}\text{vec}_{D^0}(\boldsymbol{\delta})^T\boldsymbol{Z}_2) \mid \boldsymbol{Z}_1\Big)\Big|\boldsymbol{Y}_{1,\ldots,p-1}\Big)\Big]$$

$$= \liminf_{n\to\infty}$$

$$E\Big[\exp\Big(-\frac{1}{2}\sum_{j=1}^{p} \sqrt{n}\text{vec}_{D_j^0}(\boldsymbol{\delta})^T(\frac{1}{n}\boldsymbol{Y}^T_{\text{pa}_j^1}\boldsymbol{Y}_{\text{pa}_j^1})_{D_j^0,D_j^0}\sqrt{n}\text{vec}_{D_j^0}(\boldsymbol{\delta})\Big)$$

$$\times E\Big(\mathbb{I}\Big(\|\boldsymbol{Z}_1\|^2 + \sum_1^{p-1} \|\boldsymbol{A}(j)\boldsymbol{e}_j\|^2 > \chi^2_{|D^0|,1-\alpha}\Big)$$

$$\exp(\text{vec}_{D^0}(\boldsymbol{\delta})^T\boldsymbol{Y}^T_{\text{pa}_j^1}\boldsymbol{A}(p)^T\boldsymbol{Z}_1)\Big|\boldsymbol{Y}_{1,\ldots,p-1}\Big)\Big]$$

$$= \liminf_{n\to\infty} E\Big[P\Big(\|\boldsymbol{Z}_1 + \boldsymbol{A}(p)\boldsymbol{Y}_{\text{pa}_j^1}\text{vec}_{D^0}(\boldsymbol{\delta})\|^2$$

$$+ \sum_1^{p-1} \|\boldsymbol{A}(j)\boldsymbol{e}_j\|^2 > \chi^2_{|D^0|,1-\alpha}|\boldsymbol{Y}_{1,\ldots,p-1}\Big)\Big],$$

where $\|\boldsymbol{Z}_1 + \boldsymbol{A}(p)\boldsymbol{Y}_{D_j^0}\text{vec}_{D^0}(\boldsymbol{\delta})\|^2|\boldsymbol{Y}_{1,\ldots,p-1} \sim \chi^2_{|D^0|}(\|\boldsymbol{A}(p)\boldsymbol{Y}_{D_j^0}$ $\text{vec}_{D^0}(\boldsymbol{\delta})\|^2)$, and $\chi^2_r(\mu^2)$ denotes the non-central chi-squared distribution with the degrees of freedom r and non-centrality parameter μ^2. By Lemma 3, for any $0 < t < 1$,

$$P\Big(\forall j :\text{vec}_{D^0}(\boldsymbol{\delta})^T\boldsymbol{Y}^T_{D_j^0}\boldsymbol{A}(p)^T\boldsymbol{A}(p)\boldsymbol{Y}_{D_j^0}\text{vec}_{D^0}(\boldsymbol{\delta})$$

$$\geq (1-t)\sqrt{n}\text{vec}_{D_j^0}(\boldsymbol{\delta})^T(\boldsymbol{\Omega}_{D_j^0,D_j^0})^{-1}\sqrt{n}\text{vec}_{D_j^0}(\boldsymbol{\delta})\Big) \to 1,$$

under Assumption 4. Thus, proceeding above calculation for $j = p-1,\ldots,1$ yields that

$$\liminf_{n\to\infty} P_{\boldsymbol{U}^n}\left(2Lr > \chi^2_{|D^0|,1-\alpha}\right) \geq P\Big(\chi^2_{|D^0|}\Big((1-t)$$

$$\sum_1^{p}\boldsymbol{h}_j^T(\boldsymbol{\Omega}_{D_j^0,D_j^0})^{-1}\boldsymbol{h}_j\Big) \geq \chi^2_{|D^0|,1-\alpha}\Big).$$

This leads to (ii).

Finally, when $|D^0| \to \infty$, a similar argument yields that for any $0 < t < 1$,

$$\liminf_{n\to\infty} P_{\boldsymbol{U}^n}\left((2|D^0|)^{-1/2}(2Lr - |D^0|) > z_{1-\alpha}\right)$$

$$\geq \liminf_{n\to\infty} P_{\boldsymbol{U}^0}\left(\chi^2_{|D^0|}\Big((1-t)\sum_1^{p}\boldsymbol{h}_j^T(\boldsymbol{\Omega}_{D_j^0,D_j^0})^{-1}\boldsymbol{h}_j\Big)\right.$$

$$\left. > |D^0| + |D^0|^{-1/2}z_{1-\alpha}\right)$$

$$= P\left(N(0,1) > \frac{z_{1-\alpha} - (1-t)\lambda^{-1}_{\max}(\boldsymbol{\Omega})h^2}{\sqrt{2}}\right),$$

where the fact that $\frac{\chi^2_r(\mu^2)-r-\mu^2}{\sqrt{2(r+2\mu^2)}} \xrightarrow{d} N(0,1)$ is used. This completes the proof. □

Proof of Theorem 4. Suppose that $E \cup F$ is cyclic. Then the condition that $\boldsymbol{U}^n$ is acyclic implies that $\boldsymbol{U}^n = \boldsymbol{U}^0$ and $h = 0$, establishing (i).

Suppose $d > 0$ is fixed. Let $Lr(k) = l(\hat{\boldsymbol{\Omega}}_{H_a}) - l(\hat{\boldsymbol{\Omega}}_{H_0}(k))$; $(i_k, i_{k+1}) \in A$. For notational simplicity, let $B = \{k : (i_k, i_{k+1}) \in A\}$ and assume $B = \{1,\ldots,|A|\}$ without loss of generality. Using the argument in Theorem 3, for $k \in B$,

$$\begin{aligned}2Lr(k) &= \sigma^{-2}\boldsymbol{e}^T_{i_{k+1}}\left\{\boldsymbol{P}_{\text{pa}^1_{i_{k+1}}} - \boldsymbol{P}_{\text{pa}^1_{i_{k+1}}\setminus\{i_k\}}\right\}\boldsymbol{e}_{i_{k+1}} + o_p(1)\\ &= (\boldsymbol{a}_k^T\boldsymbol{e}_{i_{k+1}})^2 + o_p(1),\end{aligned}$$

and

$$\begin{aligned}&\begin{pmatrix}\boldsymbol{a}_k^T\boldsymbol{e}_{i_{k+1}}\\ \boldsymbol{\eta}_{i_{k+1}}\end{pmatrix}\Bigg|\boldsymbol{Y}_{i_1,\ldots,i_k}\\ &\equiv \begin{pmatrix}\boldsymbol{Z}_1(k)\\ \boldsymbol{Z}_2(k)\end{pmatrix} \sim N\left(\boldsymbol{0}, \begin{pmatrix}1 & n^{-1/2}\boldsymbol{a}_k^T\boldsymbol{Y}_{i_k}\\ n^{-1/2}\boldsymbol{a}_k^T\boldsymbol{Y}_{i_k} & n^{-1}\boldsymbol{Y}^T_{i_k}\boldsymbol{Y}_{i_k}\end{pmatrix}\right).\end{aligned}$$

Hence, taking iterated expectation as in the proof of Theorem 3 yields that

$$\begin{aligned}&\liminf_{n\to\infty} P_{\boldsymbol{U}^n}(2Lr > \Gamma_{d,1-\alpha})\\ &\quad\geq \liminf_{n\to\infty} P_{\boldsymbol{U}^0}(f(\chi^2_1(h^2/\Omega_{i_1i_1}),\ldots,\chi^2_1(h^2/\Omega_{i_di_d})) > \Gamma_{d,1-\alpha}),\end{aligned}$$

where $f(\boldsymbol{x}) = \min_{1\leq k\leq d} x_k$; $\boldsymbol{x} = (x_1,\ldots,x_d) \in \mathbb{R}^d$. This establishes (ii).

Similarly, when $d \to \infty$,

$$\begin{aligned}&\liminf_{n\to\infty} P_{\boldsymbol{U}^n}(2Lr \geq d^{-2}\Gamma_{1-\alpha})\\ &\quad= \liminf_{n\to\infty} P_{\boldsymbol{U}^0}(d^2f(\chi^2_1(h^2/\Omega_{i_1i_1}),\ldots,\chi^2_1(h^2/\Omega_{i_di_d})) \geq \Gamma_{1-\alpha}).\end{aligned}$$

This completes the proof. □

Proof of Theorem 5. The proof involves two parts: convergence of ADMM and convergence of DC programming.

For the ADMM part, we show that (13) reduces to a two-block ADMM. Then the desired result follows from Section 3.2.1 of Boyd et al. (2011). Recall that the convex subproblem considered at $(m+1)$th DC iteration is

$$\begin{aligned}\min_{(\boldsymbol{U},\boldsymbol{V},\boldsymbol{\Lambda},\boldsymbol{\xi})}\quad &\text{RSS}(\boldsymbol{U}) + \mu\sum_{(i,j)\notin E_1}|V_{ij}|\mathbb{I}(|\hat{U}^{[m]}_{ij}| \leq \tau),\\ \text{subj to}\quad &\boldsymbol{U}_{E_2} = 0, \quad \boldsymbol{U} - \boldsymbol{V} = \boldsymbol{0},\\ &V_{ij}\mathbb{I}(|\hat{U}^{[m]}_{ij}| \leq \tau) + \tau\mathbb{I}(|\hat{U}^{[m]}_{ij}| > \tau)\\ &+ \xi^1_{ijk} - \tau\lambda_{jk} - \tau\mathbb{I}(i \neq k) + \tau\lambda_{ik} = 0,\\ &V_{ij}\mathbb{I}(|\hat{U}^{[m]}_{ij}| \leq \tau) - \tau\mathbb{I}(|\hat{U}^{[m]}_{ij}| > \tau)\\ &- \xi^2_{ijk} + \tau\lambda_{jk} + \tau\mathbb{I}(i \neq k) - \tau\lambda_{ik} = 0,\\ &\xi^1_{ijk}, \xi^2_{ijk} \geq 0, \quad i,j,k = 1,\ldots,p, i \neq j.\end{aligned}$$

Let $g(\boldsymbol{V}) = \mu\sum_{(i,j)\notin E_1}|V_{ij}|\mathbb{I}(|\hat{U}^{[m]}_{ij}| \leq \tau)$ and let $h(\boldsymbol{U},\boldsymbol{\Lambda},\boldsymbol{\xi}) = \text{RSS}(\boldsymbol{U})$. Treating $(\boldsymbol{U},\boldsymbol{\Lambda},\boldsymbol{\xi})$ as a variable, (13) exactly minimizes the objective function in $(\boldsymbol{U},\boldsymbol{\Lambda},\boldsymbol{\xi})$-direction at each ADMM iteration. Note that g and h are convex, proper, and closed. Moreover, the linear constraints implies the existence of Lagrangian multipliers for the unaugmented Lagrangian L_0 defined in Section 3.2 of Boyd et al. (2011), which further implies the existence of a saddle point of L_0. Moreover, (13) reduces to a standard two-block ADMM that satisfies Assumptions 1 and 2 of Boyd et al. (2011). Then convergence is established.

For the DC programming part, note that the Karush-Kuhn-Tucker conditions imply that there exists Lagrangian multipliers $\mu \geq 0$ and

$\boldsymbol{\nu} = \{\nu_{ijk} \geq 0\}_{i,j,k=1,\ldots,p;i\neq j}$ such that $(\boldsymbol{U}^{[m^*]}, \boldsymbol{\Lambda}^{[m^*]})$ minimizes the Lagrange function, where m^* is the iteration index at termination,

$$f(\boldsymbol{U}, \boldsymbol{\Lambda}) = \mathrm{RSS}(\boldsymbol{U}) + \mu \left(\sum_{(i,j)\notin E_1} \mathrm{J}_\tau(U_{ij}) - \kappa \right) + \sum_{i,j,k=1,\ldots,p;i\leq j} \nu_{ijk} \left(\mathrm{J}_\tau(U_{ij}) - \lambda_{ik} - \mathbb{I}(j \neq k) + \lambda_{jk} \right),$$

with respect to $\boldsymbol{U}$. For the constrained MLEs defined in (7) and (8), $0 \leq f(\boldsymbol{U}^{[m]}, \boldsymbol{\Lambda}^{[m]}) = f^{[m+1]}(\boldsymbol{U}^{[m]}, \boldsymbol{\Lambda}^{[m]}) \leq f^{[m]}(\boldsymbol{U}^{[m]}, \boldsymbol{\Lambda}^{[m]}) \leq f^{[m]}(\boldsymbol{U}^{[m-1]}, \boldsymbol{\Lambda}^{[m-1]}) = f(\boldsymbol{U}^{[m-1]}, \boldsymbol{\Lambda}^{[m-1]})$, where m is the DC iteration index and $f^{[m]}$ is the difference convex objective function at iteration m. By monotonicity, $\lim_{m\to\infty} f(\boldsymbol{U}^{[m]}, \boldsymbol{\Lambda}^{[m]}) = f(\boldsymbol{U}^{[m^*]}, \boldsymbol{\Lambda}^{[m^*]})$. The finite termination property follows from strict decreasingness of $f^{[m]}(\boldsymbol{U}^{[m]}, \boldsymbol{\Lambda}^{[m]})$ in m and finite possible values of the subgradient of the trailing convex function. At termination $f(\boldsymbol{U}^{[m^*]}, \boldsymbol{\Lambda}^{[m^*]}) = f(\boldsymbol{U}^{[m^*-1]}, \boldsymbol{\Lambda}^{[m^*-1]})$; otherwise the iteration continues. It can be verified that $(\boldsymbol{U}^{[m^*]}, \boldsymbol{\Lambda}^{[m^*]})$ satisfies the desired local optimality condition. This completes the proof. □

Acknowledgments

The authors would like to thank the editor, the associate editor, and two anonymous referees for helpful comments and suggestions.

Funding

Research supported in part by NSF grants DMS-1712564, DMS-1721216, NIH grants 1R01GM081535-01, 1R01GM126002, HL65462, and R01HL105397.

References

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011), "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *Foundations and Trends in Machine Learning*, 3, 1–122. [4,5,15]

Breiman, L., and Cutler, A. (1993), "A Deterministic Algorithm for Global Optimization," *Mathematical Programming*, 58, 179–199. [6]

Bühlmann, P., Peters, J., and Ernest, J. (2014), "CAM: Causal Additive Models, High-Dimensional Order Search and Penalized Regression," *The Annals of Statistics*, 42, 2526–2556. [1]

Cheng, J., Greiner, R., Kelly, J., Bell, D., and Liu, W. (2002), "Learning Bayesian Networks From Data: An Information-Theory Based Approach," *Artificial intelligence*, 137, 43–90. [1]

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2001), *Introduction to Algorithms*, Cambridge, MA: The MIT Press. [5]

Edwards, D. (2012), *Introduction to Graphical Modelling*, New York: Springer Science & Business Media. [1,2]

Foster, T. C., Sharrow, K. M., Masse, J. R., Norris, C. M., and Kumar, A. (2001), "Calcineurin Links Ca^{2+} Dysregulation With Brain Aging," *Journal of Neuroscience*, 21, 4066–4073. [9]

Fu, F., and Zhou, Q. (2013), "Learning Sparse Causal Gaussian Networks With Experimental Intervention: Regularization and Coordinate Descent," *Journal of the American Statistical Association*, 108, 288–300. [1]

Gabay, D., and Mercier, B. (1975), *A Dual Algorithm for the Solution of Nonlinear Variational Problems via Finite Element Approximation*, Rocquencourt: Institut de recherche d'informatique et d'automatique. [2]

Geyer, C. J. (1994), "On the Asymptotics of Constrained M-Estimation," *The Annals of Statistics*, 22, 1993–2010. [1]

Gu, J., Fu, F., and Zhou, Q. (2017), "Penalized Estimation of Directed Acyclic Graphs From Discrete Data," *Statistics and Computing*, 29, 161–176. [1]

Horst, R., and Tuy, H. (2013), *Global Optimization: Deterministic Approaches*, New York: Springer Science & Business Media. [2]

Kanehisa, M., and Goto, S. (2000), "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, 28, 27–30. [9]

Kolmogorov, A., and Tikhomirov, V. (1959), "ε-Entropy and ε-Capacity of Sets in Function Spaces," *Uspekhi Matematicheskikh Nauk*, 14, 3–86. [12]

Laurent, B., and Massart, P. (2000), "Adaptive Estimation of a Quadratic Functional by Model Selection," *Annals of Statistics*, 28, 1302–1338. [13,14]

Liu, J. (2001), *Monte Carlo Strategies in Statistical Computing*, New York: Springer. [1]

Liu, Z., Zhang, M., Xu, G., Huo, C., Tan, Q., Li, Z., and Yuan, Q. (2017), "Effective Connectivity Analysis of the Brain Network in Drivers During Actual Driving Using Near-Infrared Spectroscopy," *Frontiers in Behavioral Neuroscience*, 11, 211. [1]

Luo, R., and Zhao, H. (2011), "Bayesian Hierarchical Modeling for Signaling Pathway Inference From Single Cell Interventional Data," *The Annals of Applied Statistics*, 5, 725. [1]

Núñez-Antón, V. A., and Zimmerman, D. L. (2009), *Antedependence Models for Longitudinal Data*, Boca Raton, FL: Chapman and Hall/CRC. [2]

Ossiander, M. (1987), "A Central Limit Theorem Under Metric Entropy With L_2 Bracketing," *The Annals of Probability*, 15, 897–919. [12]

Peters, J., and Bühlmann, P. (2014), "Identifiability of Gaussian Structural Equation Models With Equal Error Variances," *Biometrika*, 101, 219–228. [2]

Peters, J., Bühlmann, P., and Meinshausen, N. (2016), "Causal Inference by Using Invariant Prediction: Identification and Confidence Intervals," *Journal of the Royal Statistical Society*, Series B, 78, 947–1012. [1]

Rockafellar, R., and Wets, R. (2011), *Variational Analysis* (Vol. 317), Berlin, Heidelberg: Springer. [6]

Rothenhäusler, D., Bühlmann, P., and Meinshausen, N. (2019), "Causal Dantzig: Fast Inference in Linear Structural Equation Models With Hidden Variables Under Additive Interventions," *The Annals of Statistics*, 47, 1688–1722. [1]

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005), "Causal Protein-Signaling Networks Derived From Multiparameter Single-Cell Data," *Science*, 308, 523–529. [1]

Shen, X., Pan, W., and Zhu, Y. (2012), "Likelihood-Based Selection and Sharp Parameter Estimation," *Journal of American Statistical Association*, 107, 223–232. [2]

Tao, P. D. (2005), "The DC (Difference of Convex Functions) Programming and DCA Revisited With DC Models of Real World Nonconvex Optimization Problems," *Annals of Operations Research*, 133, 23–46. [6,9]

Vershynin, R. (2012), "How Close Is the Sample Covariance Matrix to the Actual Covariance Matrix?," *Journal of Theoretical Probability*, 25, 655–686. [14]

Webster, J. A., Gibbs, J. R., Clarke, J., Ray, M., Zhang, W., Holmans, P., Rohrer, K., Zhao, A., Marlowe, L., Kaleem, M., and McCorquodale III, D. S. (2009), "Genetic Control of Human Brain Transcript Expression in Alzheimer Disease," *The American Journal of Human Genetics*, 84, 445–458. [2,9]

Wong, W. H., and Shen, X. (1995), "Probability Inequalities for Likelihood Ratios and Convergence Rates of Sieve MLEs," *The Annals of Statistics*, 23, 339–362. [12,13]

Yuan, Y., Shen, X., Pan, W., and Wang, Z. (2019), "Constrained Likelihood for Reconstructing a Directed Acyclic Gaussian Graph," *Biometrika*, 106, 109–125. [1,2]