HHS Public Access

Author manuscript

J Am Stat Assoc. Author manuscript; available in PMC 2020 August 11.

Published in final edited form as:

J Am Stat Assoc. 2020 ; 115(529): 217–230. doi:10.1080/01621459.2018.1540986.

On High-Dimensional Constrained Maximum Likelihood Inference

Yunzhang Zhu[a], **Xiaotong Shen**[b], **Wei Pan**[c]

[a]Department of Statistics, Ohio State University, Columbus, OH

[b]School of Statistics, University of Minnesota, Minneapolis, MN

[c]Division of Biostatistics, University of Minnesota, Minneapolis, MN

Abstract

Inference in a high-dimensional situation may involve regularization of a certain form to treat overparameterization, imposing challenges to inference. The common practice of inference uses either a regularized model, as in inference after model selection, or bias-reduction known as "debias." While the first ignores statistical uncertainty inherent in regularization, the second reduces the bias inbred in regularization at the expense of increased variance. In this article, we propose a constrained maximum likelihood method for hypothesis testing involving unspecific nuisance parameters, with a focus of alleviating the impact of regularization on inference. Particularly, for general composite hypotheses, we unregularize hypothesized parameters whereas regularizing nuisance parameters through a L_0-constraint controlling the degree of sparseness. This approach is analogous to semiparametric likelihood inference in a high-dimensional situation. On this ground, for the Gaussian graphical model and linear regression, we derive conditions under which the asymptotic distribution of the constrained likelihood ratio is established, permitting parameter dimension increasing with the sample size. Interestingly, the corresponding limiting distribution is the chi-square or normal, depending on if the co-dimension of a test is finite or increases with the sample size, leading to asymptotic similar tests. This goes beyond the classical Wilks phenomenon. Numerically, we demonstrate that the proposed method performs well against it competitors in various scenarios. Finally, we apply the proposed method to infer linkages in brain network analysis based on MRI data, to contrast Alzheimer's disease patients against healthy subjects. Supplementary materials for this article are available online.

Keywords

Brain networks; Generalized Wilks phenomenon; High-dimensionality; L_0-regularization; (p, n)-asymptotics; Similar tests

CONTACT Yunzhang Zhu zhu.219@osu.edu Department of Statistics, Ohio State University, Columbus, OH 43210.

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

Supplementary Materials

The technical details of the counter example in Section 2.2 and the proofs of Lemma 2–9 are provided.

1. Introduction

High-dimensional analysis has become increasingly important in modern statistics, where a model's size may greatly exceed the sample size. For instance, in studying the brain activity, a brain network is often examined, which consists of structurally and functionally interconnected regions at many scales. At the macroscopic level, networks can be studied noninvasively in healthy and disease subjects with functional MRI (fMRI) and other modalities such as MEG and EEG. In such a situation, inferring the structure of a network becomes critically important, which is one kind of high-dimensional inference. Yet, high-dimensional inference remains largely under-studied. In this article, we develop a full likelihood inferential method, particularly for a Gaussian graphical model and high-dimensional linear regression.

In the literature, a great deal of effort has been devoted to estimation. For the linear model, many methods focus on estimation with sparsity-inducing convex and nonconvex regularization such as Lasso, SCAD, MCP, and TLP (Tibshirani 1996; Fan and Li 2001; Zhang 2010; Shen, Pan, and Zhu 2012), among others. For the Gaussian graphical model, methods include the regularized likelihood approach (Rothman et al. 2008; Friedman, Hastie, and Tibshirani 2008; Yuan and Lin 2007; Fan, Feng, and Wu 2009; Shen, Pan, and Zhu 2012) and the nodewise regression approach (Meinshausen and Bühlmann 2006), and their extensions, such as conditional Gaussian graphical (Li, Chun, and Zhao 2012; Yin and Li 2013) and multiple Gaussian graphical models (Zhu, Shen, and Pan 2014; Lin et al. 2017). Despite progress, there is a paucity of inferential methods for high-dimensional models, although some have been recently proposed in Zhang and Zhang (2014), Van de Geer et al. (2014), Javanmard and Montanari (2014), and Janková and Van de Geer (2017), where CI are constructed based on a bias-reduction method called "debias" (Zhang and Zhang 2014). One potential issue of this kind of approach is not asymptotically similar with its null distribution depending on unknown nuisance parameters to be estimated, and most critically the variance is likely to increase after debias, resulting in an increased length of a CI.

In this article, we propose a maximum likelihood method subject to certain constraints for hypothesis testing involving unspecific nuisance parameters, referred to as the constrained maximum likelihood ratio (CMLR) test, which regularizes the degree of sparsity of un-hypothesized parameters in a high-dimensional model, whereas hypothesized parameters are not regularized. This is an analogy of semiparametric inference with respect to the parametric component, which enables to alleviate the inherited bias problem due to regularization. For computation, we employ a surrogate of the L_0-function, a truncated L_1-function, for the constraints. On this ground, we develop the CMLR test, which is asymptotically similar with its null distribution independent of unspecific nuisance parameters. Moreover, we derive the asymptotic distributions of the test in the presence of growing parameter dimensions for the Gaussian graphical model and linear model. Most importantly, the corresponding distribution for the CMLR test statistic converges to the chi-square distribution when the co-dimension, or the difference in dimensionality between the full and null spaces, is finite, and converges to normal (after proper centering and scaling) when the co-dimension tends to infinity. This occurs in a situation roughly when

$\frac{(|A^0| + |B|)\log p}{n^{1/2}} \to 0$ and $\frac{\sqrt{|B|}(|A^0| + |B|)}{n} \to 0$, respectively, in the Gaussian graphical model and linear regression, where $|B|$ and $|A^0|$ are the numbers of the hypothesized parameters and the nonzero unhypothesized parameters. Such a critical assumption is in contrast to a requirement of $\frac{\log p}{n} \to 0$ for sparse feature selection Shen et al. (2013), which has been used in Portnoy (1988) for the maximum likelihood estimation in a different context. Empirically, the asymptotic approximation becomes inadequate when departure from this assumption occurs in a less sparse situation. To our knowledge, our result is the first of this kind, providing a multivariate likelihood test in the presence of high-dimensional nuisance parameters. This is in contrast to a univariate debias test Zhang and Zhang (2014), Van de Geer et al. (2014), Javanmard and Montanari (2014), and Janková and Van de Geer (2017). When specializing the CMLR test to a single parameter in the Gaussian graphical model and linear regression, we show that it has asymptotic power, that is, no less than that of the debias test; see, Theorem 3. This is anticipated since the debias test does not capture all the information contained in the likelihood, whereas the full likelihood takes into account component to component dependencies. This aspect is illustrated by our second numerical example in which a null hypothesis involves a row (column) of offdiagonals of the precision matrix. Of course, a multivariate likelihood test as ours may require stronger conditions than a univariate non-likelihood test, which is analogous to the classical situation of the maximum likelihood versus the method of moments in inference. Throughout this article, we shall focus our attention to the CMLR test as opposed to the corresponding Wald test based on the constrained maximum likelihood, which not asymptotically similar, given that it is rather challenging to invert a high-dimensional Fisher information matrix.

Computationally, we relax the nonconvex minimization using an L_0-surrogate function by solving a sequence of convex relaxations as in Shen, Pan, and Zhu (2012). For each convex relaxation, we employ the alternating direction method of multipliers algorithm Boyd et al. (2011), permitting a treatment of problems of medium to large size. Moreover, we study the operating characteristics of the proposed inference method and compare against the debias methods through numerical examples. In simulations, we demonstrate that the proposed method performs well under various scenarios, and compares favorably against its competitors. Finally, we apply the proposed method to confirm that a reduced level of connectivity is observed in certain brain regions in the default mode network (DMN) but an increased level in others for Alzheimer's disease (AD) patients as compared to healthy subjects.

The rest of the article is organized as follows. Section 2 proposes a constrained likelihood ratio test, and gives specific conditions under which the asymptotic approximation of the sampling distribution of the test is valid for the Gaussian graphical model and linear regression. Section 3 performs the power analysis for the CMLR test. Section 4 discusses computational strategies for the proposed test. Section 5 performs numerical studies, followed by an application of the tests to detect the structural changes in brain network analysis for AD subjects versus healthy subjects in Section 6. Section 7 is devoted to technical proofs.

2. Constrained Likelihood Ratios

Given an iid sample $X_1,...,X_n$ from a probability distribution with density $p_{\boldsymbol{\theta}}$, consider a testing problem $H_0 : \theta_i = 0;\ i \in B$ versus $H_a : \theta_i \neq 0$ for some $i \in B$, with unspecific nuisance parameters θ_j for $j \in B^c$, possibly high-dimensional, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d$, and $B \subseteq \{1,...,d\}$. Here, we allow the dimension of $\boldsymbol{\theta}$ and size of $|B|$ to grow as a function of the sample size n. For a problem of this type, we construct a constrained likelihood ratio with a sparsity constraint on nuisance parameters θ_{B^c}. Specifically, define

$$\widehat{\boldsymbol{\theta}}^{(0)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} L_n(\boldsymbol{\theta}) \quad \text{subj to: } \sum_{i \notin B} p_\tau(|\theta_i|) \leq K \text{ and } \boldsymbol{\theta}_B = 0 \tag{1}$$

$$\widehat{\boldsymbol{\theta}}^{(1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} L_n(\boldsymbol{\theta}) \quad \text{subj to: } \sum_{i \notin B} p_\tau(|\theta_i|) \leq K, \tag{2}$$

where $L_n(\boldsymbol{\theta}) = \sum_{i=1}^n \log p_{\boldsymbol{\theta}}(X_i)$ is the log-likelihood, $p_{\boldsymbol{\tau}}(x) = \min(x/\boldsymbol{\tau}, 1)$ is the truncated L_1-function Shen, Pan, and Zhu (2012) as a surrogate of the L_0-function, and $(K, \boldsymbol{\tau})$ are nonnegative tuning parameters. In this situation, without the sparsity constraint, $\widehat{\boldsymbol{\theta}}^{(0)}$ and $\widehat{\boldsymbol{\theta}}^{(1)}$ in (1) and (2) are exactly the maximum likelihood estimates under H_0 and H_a, respectively. Now, we define the constrained likelihood ratio as: $\Lambda_n(B) = 2\left(L_n\left(\widehat{\boldsymbol{\theta}}^{(1)}\right) - L_n\left(\widehat{\boldsymbol{\theta}}^{(0)}\right)\right)$. In what is to follow, we derive the asymptotic distribution of $\Lambda_n(B)$ in a high-dimensional situation for the Gaussian graphical model and linear regression. On this ground, an asymptotically similar test is derived, whose null distribution is independent of nuisance parameters.

Tuning parameters K and $\boldsymbol{\tau}$ in (1) and (2) are estimated using a cross-validation (CV) criterion based on the full model (1). Choosing the same values of $(K, \boldsymbol{\tau})$ in (1) and (2) ensures the nestedness property of $\Lambda_n(B) \geq 0$ because the constrained set in (1) is a subset of that in (2). With $K = \infty$, the test statistic $\Lambda_n(B)$ reduces to the classical likelihood ratio test statistic.

2.1. *Asymptotic Distribution of $\Lambda_n(B)$ in Graphical Models*

This subsection is devoted to a Gaussian graphical model, where $\boldsymbol{X}_1,...,\boldsymbol{X}_n$ follow from a p-dimensional normal distribution $N(\mathbf{0}, \boldsymbol{\Omega}^{-1})$, with $\boldsymbol{\Omega}$ a precision matrix, or the inverse of the covariance matrix $\boldsymbol{\Sigma}$. In this case, $\boldsymbol{\theta} = \boldsymbol{\Omega}$. The log-likelihood is $L_n(\boldsymbol{\theta}) = L_n(\boldsymbol{\Omega}) = \frac{n}{2}\log\det(\boldsymbol{\Omega}) - \frac{n}{2}\operatorname{tr}(\boldsymbol{\Omega}\boldsymbol{S})$, where $\boldsymbol{S} = n^{-1}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^\top$ is the sample covariance matrix, and $\operatorname{tr}(\cdot)$ denotes the trace of a matrix.

In the foregoing testing framework, the null and alternative hypotheses can be written as: $H_0 : \boldsymbol{\Omega}_B = \mathbf{0}$ versus $H_a : \boldsymbol{\Omega}_B \neq \mathbf{0}$ for some prespecified index set B. Then the constrained log-likelihood ratio becomes $\Lambda_n(B) = 2(L_n(\widehat{\boldsymbol{\Omega}}^{(1)}) - L_n(\widehat{\boldsymbol{\Omega}}^{(0)}))$, where $\widehat{\boldsymbol{\Omega}}^{(0)}$ and $\widehat{\boldsymbol{\Omega}}^{(1)}$ are the constrained maximum likelihood estimates (CMLE)s based on the null and full spaces of the test.

To establish the asymptotic distribution of $\Lambda_n(B)$, we first introduce some notations to be used. For any symmetric matrix $\boldsymbol{M}$, let $\lambda_{\max}(\boldsymbol{M})$ and $\lambda_{\min}(\boldsymbol{M})$ be the maximum and minimum eigenvalues of $\boldsymbol{M}$, and $\|\boldsymbol{M}\|_F$ be the Frobenius norm of $\boldsymbol{M}$. Let $\setminus$ and $|\cdot|$ denote the set difference and the size of a set. For any vector $\boldsymbol{a} \in \mathbb{R}^m$, let $\left\|\boldsymbol{a}\right\|_2 = \sqrt{a_1^2 + \cdots + a_m^2}$. Denote by $\overline{\boldsymbol{\Omega}}^0_{A \cup B} = \operatorname{argmin}_{\boldsymbol{\Omega} \succ 0: \boldsymbol{\Omega}_{(A \cup B)^c} = 0} K\left(\boldsymbol{\Omega}^0, \boldsymbol{\Omega}\right)$ an approximating point in a space $\{\boldsymbol{\Omega}: \boldsymbol{\Omega} \succ 0, \boldsymbol{\Omega}_{(A \cup B)^c} = 0\}$ to the true $\boldsymbol{\Omega}^0$, where $K(\boldsymbol{\Omega}^0, \boldsymbol{\Omega}) = \frac{1}{2}(\operatorname{tr}(\boldsymbol{\Omega}\boldsymbol{\Sigma}^0) + \log\frac{\det(\boldsymbol{\Omega}^0)}{\det(\boldsymbol{\Omega})} - p)$ is the Kullback–Leibler information. Let $\|\boldsymbol{\Omega}^0 - \boldsymbol{\Omega}\| = \left\|\sqrt{\boldsymbol{\Sigma}^0}\left(\boldsymbol{\Omega} - \boldsymbol{\Omega}^0\right)\sqrt{\boldsymbol{\Sigma}^0}\right\|_F$ be the Fisher-norm between $\boldsymbol{\Omega}^0$ and $\boldsymbol{\Omega}$ Shen (1997). Moreover, let $A^0 = \{i: \theta_i^0 \neq 0\}$ be the support of true parameter $\boldsymbol{\theta}^0$, $\boldsymbol{\kappa}_0 = \lambda_{\max}(\boldsymbol{\Omega}^0)/\lambda_{\min}(\boldsymbol{\Omega}^0)$ be the condition number of $\boldsymbol{\Omega}^0$, and $\kappa_1 = \frac{\bar{\lambda}^2_{\max}}{\lambda^2_{\min}(\boldsymbol{\Omega}^0)}$, where $\bar{\lambda}_{\max} = \max_{A: |A| \leq \left|A^0\right|, A \cap B = \varnothing} \lambda_{\max}(\overline{\boldsymbol{\Omega}}^0_{A \cup B})$. Let $\bar{\lambda}_{\min} = \min_{A: |A| \leq \left|A^0\right|, A \cap B = \varnothing} \lambda_{\min}(\overline{\boldsymbol{\Omega}}^0_{A \cup B})$. Let $\gamma_{\min} = \min_{(i,j) \in A^0} \left|\omega^0_{ij}\right|$ be the minimum nonzero offdiagonals of $\boldsymbol{\Omega}^0$ representing the signal strength. The following technical conditions are made.

***Assumption 1* (*Degree of separation*).**

$$C_{\min} = \min_{A: A \neq A^0, \left|A\right| = \left|A^0\right|, A \cap B = \varnothing} \min\left(\frac{\left\|\boldsymbol{\Omega}^0 - \overline{\boldsymbol{\Omega}}^0_{A \cup B}\right\|^2}{\left|A^0 \setminus A\right|}, 1\right) \geq C_1 \kappa_1 \frac{\left(\left|A^0\right| + \left|B\right|\right)\log p}{n}, \tag{3}$$

where $C_1 > 0$ is a constant.

Assumption 1 requires that the degree of separation $C_{\min}$ exceeds a certain threshold level, roughly $\frac{(|A^0| + |B|)\log p}{n}$, which measures the level of difficulty of the task of removing zero components of the nuisance (un-hypothesized) parameters of $\boldsymbol{\Omega}$ by the constrained likelihood with the L_0-constraint. To better understand (3) of Assumption 1, we consider a sufficient condition of (3) as follows:

Note that $\left\|\boldsymbol{\Omega}^0 - \overline{\boldsymbol{\Omega}}^0_{A \cup B}\right\| \geq \lambda_{\min}\left(\boldsymbol{\Sigma}^0\right)\left\|\boldsymbol{\Omega}^0 - \overline{\boldsymbol{\Omega}}^0_{A \cup B}\right\|_F \geq \lambda^{-1}_{\max}\left(\boldsymbol{\Omega}^0\right)\gamma_{\min}\sqrt{\left|A^0 \setminus A\right|}$. Consequently, a simpler but stronger condition of (3) in terms of $\gamma_{\min}$ is

$$\min(\gamma_{\min}, \lambda_{\max}(\boldsymbol{\Omega}^0)) \geq C_2 \kappa_0 \bar{\lambda}_{\max}\sqrt{\frac{\left(\left|A^0\right| + \left|B\right|\right)\log p}{n}} \tag{4}$$

for some constant $C_2 > 0$.

***Assumption 2* (*Dimension restriction for* $\Lambda_n(B)$).**

Assume that

$$\frac{\kappa_0\left(\left|B\right| + \left|A^0\right|\right)\log p}{\sqrt{n}} \to 0, \text{ as } n \to \infty.$$

Assumption 2 restricts the size p for an asymptotic approximation of the sampling distribution of the likelihood ratio tests, which is closely related to that in Portnoy (1988) for a different problem. Note that if $|A^0| = O(p)$ and $|B| = O(p)$ then Assumption 2 roughly requires that $p\log p/\sqrt{n} \to 0$.

Theorem 1 gives the asymptotic distribution of $\Lambda_n(B)$ when $|B|$ is either fixed or grows with n, referred to as Wilks phenomenon and generalized Wilks phenomenon, respectively.

***Theorem 1* (*Asymptotic sampling distribution of* $\Lambda_n(B)$).**

Under Assumptions 1–2, there exists optimal tuning parameters (K, τ) with $K|A^0|$ and $\tau \le \frac{\bar{\lambda}_{\min}\min(\sqrt{c_{\min}}, c_{\min}^2)}{12\left|A^0\right|}$ such that under H_0

(i) *Wilks phenomenon*: If $\omega_{ij}^0 = 0$ for $(i, j) \in B$ with $|B|$ fixed, then

$$\Lambda_n(B) \xrightarrow{d} \chi^2_{|B|} \text{ as } n \to \infty.$$

(ii) *Generalized Wilks phenomenon*: If $\omega_{ij}^0 = 0$ for $(i, j) \in B$ with $|B| \to \infty$, then

$$(2|B|)^{-1/2}(\Lambda_n(B) - |B|) \xrightarrow{d} N(0, 1) \text{ as } n \to \infty.$$

Concerning Assumptions 1 and 2, we remark that the degree of separation assumption (3) or (4) is necessary for the result of Theorem 1. Without Assumption 1, the result may break down, as suggested by a counter example in Lemma 1 for a parallel condition—Assumption 3 in linear regression in Section 2.2. This is expected because when the constrained likelihood cannot be over-selection consistency when Assumption 1 breaks down in view of the result of Shen, Pan, and Zhu (2012). That means that any under-selected component yields a bias of order $\sqrt{\frac{\log p}{n}}$. As a result, the foregoing results are not generally expected to hold. Moreover, Assumption 2 is intended for joint inference of multiple parameters, for instance, testing zero offdiagonals of one row or column of $\mathbf{\Omega}$ as in the second simulation example of Section 4. These assumptions, as we believe, are needed for multivariate tests based on a full likelihood although we have not proved so, which appear stronger than those required for a univariate debias test based on a pseudo likelihood Janková and Van de Geer (2017). This is primarily due to the full likelihood approach estimating component to component dependencies in lieu of a marginal approach without them, leading to higher efficiency when possible. This is evident from Corollary 1 that the CMLR gives more precise inference than the debias test under these conditions.

The result of Theorem 1 depends on the optimal tuning parameter $K = K^0$ and τ, both of which are unknown in practice. Therefore, K is estimated by cross-validation through tuning, and the exact knowledge of the value K is not necessary, whereas τ is usually set to be a small number, say 10^{-2}, in practice.

2.2. *Asymptotic Distribution of $\Lambda_n(B)$ in Linear Regression*

In linear regression, a random sample $(Y_i, \boldsymbol{x}_i)_{i=1}^n$ follows

$$Y_i = \boldsymbol{\beta}^T \boldsymbol{x}_i + \epsilon_i; \qquad \epsilon_i \sim N(0, \sigma^2); \qquad i = 1, \ldots, n \tag{5}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ and $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$ are p-dimensional vectors of regression coefficients and predictors, and $\boldsymbol{x}_i$ is independent of random error ϵ_i. In (5), it is known *priori* that $\boldsymbol{\beta}$ is sparse in that $\beta_j = 0, j \notin A^0$, and $\beta_j \neq 0, j \in A^0$, where $A^0 \subseteq \{1, 2, \ldots, p\}$.

In this case, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma)$. Our focus is to test $H_0 : \boldsymbol{\beta}_B = \mathbf{0}$ versus $H_a = \boldsymbol{\beta}_B \neq \mathbf{0}$ for some index set B. The log-likelihood is $L_n(\theta) = L_n(\boldsymbol{\beta}, \sigma) = -\frac{1}{2\sigma^2}\left\|Y - X\boldsymbol{\beta}\right\|_2^2 - n\log(\sqrt{2\pi}\sigma)$, and the constrained log-likelihood ratio is accordingly defined as $\Lambda_n(B) = 2(L_n(\hat{\boldsymbol{\beta}}^{(1)}, \hat{\sigma}^{(1)}) - L_n(\hat{\boldsymbol{\beta}}^{(0)}, \hat{\sigma}^{(0)}))$, where $\hat{\boldsymbol{\beta}}^{(0)}$ and $\hat{\boldsymbol{\beta}}^{(1)}$ are the CMLE based on the null and full spaces of the test.

A parallel condition of Assumption 1 is made in Assumption 3.

***Assumption 3* (*Degree of separation condition, Shen et al. 2013*).**

$$\min_{A:\,|A| \le |A^0| \text{ and } A \neq A^0} \inf_{\boldsymbol{\beta}} \frac{\left\|X\boldsymbol{\beta}^0 - X_{A \cup B}\boldsymbol{\beta}_{A \cup B}\right\|_2^2}{n|A^0 \backslash A|} \ge C_0 \sigma^2 \frac{\log p}{n} \tag{6}$$

for some absolute constant C_0 that may depend on the design matrix X.

A parallel result of Theorem 1 is established for linear regression.

***Theorem 2* (*Sampling distribution of $\Lambda_n(B)$*).**

Assume that $\frac{\sqrt{|B|}(|A^0| + |B|)}{n} \to 0$. Under Assumptions 3, there exists optimal tuning parameters (K, τ) with $K = |A^0|$ and $0 < \tau \le \sigma\sqrt{\frac{6}{(n+2)p\lambda_{\max}(X^\top X)}}$ such that under H_0

(i) *Wilks phenomenon*: If $\beta_i = 0$ for $i \in B$ with $|B|$ fixed, then

$$\Lambda_n(B) \xrightarrow{d} \chi^2_{|B|} \text{ as } n \to \infty.$$

(ii) *Generalized Wilks phenomenon*: If $\beta_i = 0$ for $i \in B$ with $|B| \to \infty$, then

$$(2|B|)^{-1/2}(\Lambda_n(B) - |B|) \xrightarrow{d} N(0,1) \text{ as } n \to \infty .$$

Note of worthy is that the requirement $\frac{\sqrt{|B|}\left(\left|A^0\right| + |B|\right)}{n} \to 0$ in linear regression appears weaker than that $\frac{\left(\left|A^0\right| + |B|\right)\log p}{n^{1/2}} \to 0$ in the Gaussian graphical model. This is primarily because the error for the likelihood ratio approximation in the former is smaller in magnitude.

Next we provide a counter example to show that the result in Theorem 2 breaks down when Assumption 3 is violated in the absence of a strong signal strength. In other words, such an assumption is necessary for such a full likelihood approach to gain the test efficiency, which is in contrast to a pseudo-likelihood approach.

***Lemma 1* (*A counter example*).**

In (5), we write $y = \beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}$, where $\boldsymbol{x} = (x_1,...,x_p)$ are independently distributed from $N(\mu_j, 1)$ with $\mu_1 = 0$ and $\mu_j = 1$; $2 \le j \le p$, and ϵ is $N(0, 1 - n^{-1})$, independent of $\boldsymbol{x}$. Assume that $\beta_0 = 0$ and $\boldsymbol{\beta} = (n^{-1/2}, 0,...,0)$, or, $y = n^{-1/2}x_1 + \epsilon$. Then Assumption 3 is violated. Now consider a hypothesis test of $H_0 : \beta_0 = 0$ versus $H_1 : \beta_0 \neq 0$. If $\frac{\log p}{n} \to 0$ as $n, p \to \infty$, then $\Lambda_n(B) \xrightarrow{p} \infty$ as $n, p \to \infty$, with $B = \{0\}$.

3. Power Analysis

This section analyzes the local limiting power function of the CMLR test and compare it with that of the debias test of Janková and Van de Geer (2017) in Gaussian graphical model. To that the null H_0 for fixed index set B for the Gaussian graphical end, we first establish the asymptotic distribution of $\hat{\theta}_B$ under model and linear model. Then, we use those results to carry out a local power analysis for both models.

3.1. Asymptotic Normality

We first introduce some notations before presenting the asymptotic normality results for Gaussian graphical model. Let $\text{vec}_B(\boldsymbol{C}) = (\sqrt{1 + \mathbb{I}(i \neq j)}c_{ij})_{(i,j):(i,j) \text{ or } (j,i) \in B}$ is a sub-vector of vec($\boldsymbol{C}$) excluding components with indices not in B, $\text{vec}(\boldsymbol{C}) = \left(\sqrt{1 + \mathbb{I}(i \neq j)}c_{ij}\right)_{i \le j} \in \mathbb{R}^{\frac{p(p+1)}{2}}$ is a scaled vectorization of a $p \times p$ symmetric matrix $\boldsymbol{C}$ (Alizadeh et al. 1998) and $\mathbb{I}(\cdot)$ is the indicator. For the Fisher information, we need the symmetric Kronecker product Alizadeh et al. (1998) for a $p \times p$ symmetric matrix $\boldsymbol{C}$ to treat derivatives of the log-likelihood with respect to a matrix. Define the symmetric Kronecker product of $\boldsymbol{C}$ $\boldsymbol{C} \otimes_s C \in \mathbb{R}^{\frac{p(p+1)}{2} \times \frac{p(p+1)}{2}}$ as $(\boldsymbol{C} \otimes_s \boldsymbol{C})\text{vec}(\boldsymbol{\Delta}) = \text{vec}(\boldsymbol{C}\boldsymbol{\Delta}\boldsymbol{C})$ for any symmetric matrix $\boldsymbol{\Delta}$, and define the Fisher information matrix for the $\frac{p(p+1)}{2}$-dimensional vector vec($\boldsymbol{\Omega}$) as $\boldsymbol{I} = \nabla^2(-\frac{1}{2}\log\det\boldsymbol{\Omega}^0) = \frac{1}{2}\boldsymbol{\Sigma}^0 \otimes_s \boldsymbol{\Sigma}^0$, c.f., Lemma 2. Given an index set B, we define a $|B| \times |B|$ submatrix $\boldsymbol{I}_{B,B}$ as $\boldsymbol{I}_{B,B} = \left(I_{(i,j),(k,l)}\right)_{(i,j),(k,l) \in B}$, extracting the

corresponding $|B| \times |B|$ submatrix from $\boldsymbol{I}$. Theorem 1 gives the asymptotic distribution of $\mathrm{vec}_B(\widehat{\boldsymbol{\Omega}}^{(1)})$.

***Proposition 1* (*Asymptotic distribution of CMLE* $\widehat{\Omega}^{(1)}$).**

for Gausian graphical model). Under Assumptions 1 and 2, if $|B|$ is fixed, there exists a pair of tuning parameters (K, τ) with $K = |A^0|$ and $\tau \leq \frac{\bar{\lambda}_{\min} \min\left(\sqrt{C_{\min}}, C_{\min}^2\right)}{12|A^0|}$ such that $\widehat{\boldsymbol{\Omega}}^{(1)}$ satisfies

$$\sqrt{n}\ \mathrm{vec}_B\left(\widehat{\boldsymbol{\Omega}}^{(1)} - \boldsymbol{\Omega}^0\right) \xrightarrow{d} N\left(0, \left(\boldsymbol{I}^{-1}_{A^0 \cup B,\, A^0 \cup B}\right)_{B,\,B}\right), \tag{7}$$

where $\left(\boldsymbol{I}^{-1}_{A^0 \cup B,\, A^0 \cup B}\right)_{B,\,B}$ extracts a $|B| \times |B|$ submatrix from $\boldsymbol{I}^{-1}_{A^0 \cup B,\, A^0 \cup B}$.

For linear regression, a similar asymptotic result can be derived.

***Proposition 2* (*Asymptotic distribution of CMLE*).**

Assume that $X^{\top}_{A^0 \cup B} X_{A^0 \cup B}$ is inevitable. Under Assumptions 3, if $|B|$ is fixed, there exists a pair of tuning parameters (K, τ) with $K = |A^0|$ and $\tau \leq \sigma \sqrt{\frac{6}{(n+2)p\lambda_{\max}\left(X^{\top}X\right)}}$ such that $\hat{\boldsymbol{\theta}}^{(1)}_B$ satisfies

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}^{(1)}_B - \boldsymbol{\beta}^0_B\right) \xrightarrow{d} N\left(0, \left(\left(n^{-1} X^{\top}_{A^0 \cup B} X_{A^0 \cup B}\right)^{-1}\right)_{B,\,B}\right), \tag{8}$$

where $\boldsymbol{M}_{B,B}$ extracts a $|B| \times |B|$ submatrix from a matrix $\boldsymbol{M}$.

3.2. Local Power Analysis

Consider a local alternative $H_a\ \theta^n_i = \theta^0_i + (\delta_n)_i;\ i \in B$ with $(\delta_n)_{B^c} = 0$, for any θ_{B^c}, with $\|\delta_n\|_2 = \frac{h}{\sqrt{n}}$ if $|B|$ is fixed, $\|\delta_n\|_2 = \frac{h|B|^{1/4}}{\sqrt{n}}$ if $|B| \to \infty$, for some constant h. Let $\boldsymbol{\theta}^n = \left(\theta^n_1, \ldots, \theta^n_d\right)^T$. Subsequently, we study the behavior of the *local limiting power function* for the proposed CMLR test $\pi_{LR}(h, \theta_{B^c}) = \lim\inf_{n \to \infty} P_{H_a}\left(\Lambda_n(B) \geq \chi^2_{\alpha, |B|}\right)$ if $|B|$ is fixed and $\lim\inf_{n \to \infty} P_{H_a}\left((2|B|)^{-1/2}\Lambda_n(B) - |B| \right) \geq z_\alpha)$ if $|B| \to \infty$. Let the corresponding $\pi_{\mathrm{debias}}(h, \theta_{B^c})$ of the debias test in Janková and Van de Geer (2017) in the Gaussian graphical model as a result for linear regression is similar.

Theorem 3.—If for any $\Theta^n = \Omega^n$ the Assumptions 1 and 2 for the Gaussian graphical model are met and further assume that $|B|^{3/2}/n \to 0$, then for any nuisance parameters Ω_{B^c},

$$\pi_{LR}(h, \Omega_{B^c}) \to \begin{cases} \mathbb{P}\left(\left\| \boldsymbol{Z} + n^{1/2} \boldsymbol{J}_{B,B}^{-1/2} \boldsymbol{\delta}_n \right\|_2^2 \geq \chi^2_{\alpha, |B|}\right) & \text{when } |B| \text{ is fixed,} \\ \mathbb{P}\left(Z + \dfrac{n \boldsymbol{\delta}_n^\top \boldsymbol{J}_{B,B}^{-1} \boldsymbol{\delta}_n}{\sqrt{2|B|}} \geq z_\alpha \right) & \text{when } |B| \to \infty, \end{cases}$$

where $\alpha > 0$ is the level of significance, $\boldsymbol{Z} \sim N(\boldsymbol{0}, \boldsymbol{I}_{|B| \times |B|})$ is a multivariate normal random variable, $Z \sim N(0, 1)$, and $\boldsymbol{J}_{B,B}$ is the asymptotic variance of $\mathrm{vec}_B(\widehat{\boldsymbol{\Omega}}^{(1)})$ in (7). In particular, $\lim_{h \to \infty} \pi_{\mathrm{LR}}(h, \boldsymbol{\Omega}_{B^c}) = 1$. Moreover, in the one-dimensional situation with $|B| = 1$, for any h and Ω_{B^c},

$$\pi_{\mathrm{LR}}(h, \Omega_{B^c}) \geq \pi_{\mathrm{debias}}(h, \Omega_{B^c}) . \tag{9}$$

Theorem 3 suggests that the proposed CMLR test has the desirable power properties, which dominates the corresponding debias tests, which is attributed to optimality of the corresponding CMLE and likelihood ratio, as suggested by Theorem 1. Note that the debias test requires Assumption 2.

Next, we compare the asymptotic variance of our estimator to that of Janková and Van de Geer (2017) for the one-dimensional case with $|B| = 1$. As indicated by Corollary 1, our estimator has asymptotic variance, that is, no larger than that of its debias counterpart.

***Corollary 1* (*Comparison of asymptotic variances*).—**Under the assumption of Theorem 1, the asymptotic covariance matrix of $\left[\sqrt{n}\left(\widehat{\omega}_{ij} - \omega^0_{ij}\right)\right]_{(i,j) \in B}$ is upper bounded by the matrix $\left[\omega^0_{i'j}\omega^0_{ij'} + \omega^0_{jj'}\omega^0_{ii'}\right]_{(i,j) \in B, (i',j') \in B}$, where $\widehat{\omega}_{ij}$ is the ijth element of the CMLE $\widehat{\boldsymbol{\Omega}}$. When specializing the above result to the one-dimensional case, it implies that the asymptotic variance of $\sqrt{n}\left(\widehat{\omega}_{ij} - \omega^0_{ij}\right)$ is no larger than $\left[\omega^0_{ij}\right]^2 + \omega^0_{ii}\omega^0_{jj}$, the asymptotic variance of the regression estimator in Janková and Van de Geer (2017).

A parallel result of Theorem 3 is established for linear regression.

Theorem 4.—If for any $\theta^n = \beta^n$ the Assumptions 1 and 2 for the linear regression model are met. Then

$$\pi_{LR}(h, \beta_{B^c}) \to \begin{cases} \mathbb{P}\left(\left\| \boldsymbol{Z} + n^{1/2} \boldsymbol{A} \boldsymbol{X}_B \boldsymbol{\delta}_n \right\|_2^2 \geq \chi^2_{\alpha, |B|}\right) & \text{if } |B| \text{ is fixed;} \\ \mathbb{P}\left(Z + \dfrac{n\left\| \boldsymbol{A} \boldsymbol{X}_B \delta_n \right\|_2^2}{\sqrt{2 \mid B \mid}} \geq z_\alpha \right) & \text{if } |B| \to \infty . \end{cases} \tag{10}$$

where $\boldsymbol{A} \in \mathbb{R}^{n \times |B|}$ with columns being the eigenvalues of $\boldsymbol{P}_{A^0 \cup B} - \boldsymbol{P}_{A^0}$, $Z \sim N(0, 1)$, and $\boldsymbol{Z}$ is a $|B|$ dimensional normal random vector. Hence, for any nuisance parameters β_{B^c}, $\lim_{h \to \infty} \pi_{\mathrm{LR}}(h, \beta_{B^c}) = 1$.

4. Computation

To compute the CMLEs under the null and full spaces in (1) and (2), we approximately solve constrained nonconvex optimization through difference convex (DC) programming. Particularly, we follow the DC approach of Shen, Pan, and Zhu (2012) to approximate the nonconvex constraint by a sequence of convex constraints based on a difference convex decomposition iteratively. This leads to an iterative method for solving a sequence of relaxed convex problems. The reader may consult Shen, Pan, and Zhu (2012) for convergence of the method.

For (1) and (2), at the mth iteration, we solve

$$\begin{aligned} \max_{\theta} \quad & L_n(\theta) \\ \text{subj to} \quad & \sum_{i \notin A_1} |\omega_{ij}| \mathbb{I}\left(\left|\widehat{\omega}_i^{[m]}\right| \le \tau\right) \\ & \le \tau\left(K - \sum_{i \notin A_1} \mathbb{I}\left(\left|\widehat{\omega}_i^{[m]}\right| > \tau\right)\right), \theta_{A_2} = 0, \end{aligned} \tag{11}$$

to yield $\hat{\theta}^{[m+1]}$, where $A_1 = B$ and $A_2 = \varnothing$ for (1) and $A_1 = A_2 = B$ for (2). Iteration continues until two adjacent iterates are equal. To solve (11), we employ the alternating direction method of multipliers algorithm (Boyd et al. 2011), which amounts to the following iterative updating scheme

$$\begin{aligned} \theta^{[k+1]} = \underset{\theta}{\operatorname{argmin}} (-L_n(\theta) + (\rho/2) \cdot \\ \left\|\theta - \delta^{[k]} + \gamma^{[k]}\right\|_2^2), \end{aligned} \tag{12}$$

$$\begin{aligned} \delta^{(k+1)} &= \mathcal{P}_{\mathcal{F}^{[m]}}\left(\theta^{[k+1]} + \gamma^{[k]}\right), \\ \gamma^{[k+1]} &= \gamma^{[k]} + \theta^{[k+1]} - \delta^{[k+1]}, \end{aligned} \tag{13}$$

where

$$\begin{aligned} \mathcal{F}^{[m]} = \Bigg\{ & \sum_{i \notin A_1} |\theta_i| \mathbb{I}\left(\left|\theta_i^{[m]}\right| \le \tau\right) \\ & \le \tau\left(K - \sum_{(i,j) \notin A_1} \mathbb{I}\left(\left|\theta_i^{[m]}\right| > \tau\right)\right), \theta_{A_2} = 0 \Bigg\}, \end{aligned}$$

$\mathcal{P}_{\mathcal{F}[m]}(\cdot)$ denotes the projection onto the set $\mathcal{F}^{[m]}$ and $\rho > 0$ is fixed or can be adaptively updated using a strategy in Zhu (2017). Note that in both cases, the θ-update (12) can be solved using an analytic formula involving a singular value decomposition for the Gaussian graphical model (see Section 6.5 of Boyd et al. 2011) and solving a linear system for the linear model, while (13) is performed using the L_1-projection algorithm of Liu and Ye (2009) whose complexity is almost linear in a problem's size. Specifically, consider a generic problem of projection onto a weighted L_1-ball subject to equality constraint:

$$\min_{x \in \mathbb{R}^d} \frac{1}{2}\|x-y\|_2^2 \text{ subj to } \sum_{i \notin A} c_i|x_i| \le z \text{ and } x_i, i \in A,$$

where $c_i \ge 0$; $i = 1,..., d$ and A is a subset of $\{1,...,d\}$. The solution of this problem is $x_i^\star = 0$ if $i \in A$; $x_i^\star = y_i$ if $\sum_{i \notin A} c_i|y_i| \le z$; $x_i^\star = \text{sgn}(y_i)\max\left(|y_i| - c_i\lambda^\star, 0\right)$ otherwise, where $\lambda^\star$ is a root of $f(\lambda) = \sum_{i \notin A} c_i \max(|y_i| - c_i\lambda, 0) - z$. This root-finding problem is solved efficiently by bisection.

5. Numerical Examples

This section investigates operating characteristics of the proposed CMLR test with regard to the size and power of a test through simulations and compare with several strong competitors in the literature.

For the Gaussian graphical model, we examine three different types of graphs—a chain graph, a hub graph, and a random graph, as displayed in Figure 1. For a given graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, $\mathbf{\Omega}$ is generated based on connectivity of the graph, that is, $\omega_{ij} \neq 0$ iff there exists a connection between nodes i and j for $i \neq j$. Moreover, we set $\omega_{ij} = 0.3$ if i and j are connected and diagonals equal to $0.3 + c$ with c chosen so that the smallest eigenvalue of the resulting matrix equals to 0.2. Finally, a random sample of size $n = 200$ is drawn from $N(\mathbf{0}, \mathbf{\Omega}^{-1})$.

In what follows, we consider two hypothesis testing problems concerning conditional independence of components of a Gaussian random vector $\boldsymbol{X} = (X_1,...,X_p)$. The first concerns null hypothesis $H_0: \omega_{i_0 j_0} = 0$ versus its alternative $H_a: \omega_{i_0 j_0} \neq 0$; $i_0 \neq j_0$, for testing conditional independence between X_{i_0} and X_{j_0}. The second deals with $H_0: \omega_{i_0 j} = 0$; $1 \le j \neq i_0 \le p$ versus $H_a: \omega_{i_0 j} \neq 0$ for some $j \neq i_0$, for testing conditional independence of component i_0 with the rest. In either case, we apply the proposed CMLR test in Section 2 and compare it with the univariate debias test of Janková and Van de Geer (2017) in terms of the empirical size and power only in the first problem. To our knowledge, no competing methods are available for the second problem in the present situation.

For the size of a test, we calculate its empirical size as the percentage of times rejecting H_0 out of 1000 simulations when H_0 is true. For the power of a test, we consider four different alternatives: $H_a: \omega_{ij} = \omega_{ij}^{(l)}$ for $(i, j) \neq (i_0, j_0)$ and $\omega_{i_0 j_0}^{(l)} = \frac{\omega_{i_0 j_0}{}^l}{4}, l = 1, \ldots, 4$. Under each alternative, we compute the power as the percentage of times rejecting H_0 out of 1000 simulations when H_a is true.

With regard to tuning, we fix $\tau = 0.001$ and propose to use a vanilla cross-validation to choose the optimal tuning parameter K for our test by minimizing a prediction criterion using a 5-fold CV. Specifically, we divide the dataset into five roughly equal parts denoted by $\mathcal{D}_1, \ldots, \mathcal{D}_5$. Define $\widehat{\mathbf{\Sigma}}_l$ and $\widehat{\mathbf{\Sigma}}_{-l}$, respectively, as the sample covariance matrices calculated based on samples in $\mathcal{D}_l$ and $\{\mathcal{D}_1, \ldots, \mathcal{D}_5\} \backslash \mathcal{D}_l$; $l = 1, \ldots, 5$. Similarly, define $\widehat{\mathbf{\Omega}}_{-l}(K)$ to be the

precision matrix calculated based on sample covariance matrix $\widehat{\mathbf{\Sigma}}_{-l}$; $l = 1, \ldots, 5$. The 5-fold CV criterion is $\mathrm{CV}(K) = 5^{-1}\sum_{l=1}^{5}\left(-\log\det\left(\widehat{\mathbf{\Omega}}_{-l}(K)\right) + \mathrm{tr}\left[\widehat{\mathbf{\Sigma}}_{l}\widehat{\mathbf{\Omega}}_{-l}(K)\right] - p\right)$. Then the optimal tuning parameter is obtained by minimizing CV(K) over a set of grids in the domain of K. Finally, $K^{\star} = \arg\min_{K}\mathrm{CV}(K)$ is used to compute the final estimator based on the original data.

For the first testing problem, the nominal size of a test is set to 0.05 for our CMLR test and the univariate debias test of Janková and Van de Geer (2017), denoted as *CMLR-chi-square* and *JG*, where the confidence interval in Janková and Van de Geer (2017) is converted to a two-sided test. For each graph type, three different graph sizes $p = 50, 100, 200$ are examined. As indicated in Table 1, the empirical size of the CMLR test is under or close to the nominal size 0.05. Moreover, as suggested in Table 1, the power of the likelihood ratio test is uniformly higher across all the 12 scenarios with four alternatives and three different dimensions, where the largest improvements are seen for the hub graph, particularly with $p =$ 100, 200 for an amount of improvement of 50% or more. This result is anticipated because the likelihood method is more efficient than a regression approach.

To study operating characteristics of the constrained likelihood test, we focus on the validity of asymptotic approximations based on the chi-square or normal distribution under H_0. For the first problem, Figure 2 indicates that the chi-square approximation on one degree of freedom is adequate for the likelihood ratio test. Similarly, for the second testing problem involving a column/row of Ω, Figure 3 confirms that the normal approximation is again adequate for the CMLR test. Overall, the asymptotic approximations appear adequate.

For the linear model, we perform a parallel simulation study to compare the CMLR test with the debiased lasso test (Zhang and Zhang 2014; Van de Geer et al. 2014) and the method of Zhang and Cheng (2017). In (5), we examine (n, p) = (100, 50), (100, 200), (100, 500), (100, 1000), in which predictors x_{ij} and the error ϵ_i are generated independently from $N(0, 1)$, where $\beta^0 = \left(1, 2, 3, \beta_B^0, \mathbf{0}\right)$ and $\|\beta_B\|_2 = l/10$; $l = 0, 1, \ldots, 4$. Now consider a hypothesis test with null hypothesis $H_0 : \beta_B = \mathbf{0}$ versus its alternative $H_a : \beta_B \neq \mathbf{0}$, where we let $|B| \neq 1, 5, 10$. With regard to size, power, and tuning, we follow the same scheme as in the Gaussian graphical model.

As indicated in Table 2, the empirical size of *CMLR-chi-square* and *CMLR-normal* are close to the target size 0.05, while the former does better than the latter for $|B|$ is small and worse for large $|B|$, which corroborates with the result of Theorem 2. Moreover, the power of *CMLR-chi-square* is uniformly higher across all the three scenarios with four alternatives compared to the other two competing methods. Interestingly, when $|B|$ is large, the method of Zhang and Cheng (2017) seems to control the size closer to the nominal level than the CMLR test, but the situation is just the opposite when $|B|$ is not large. Additional simulations also suggest that similar results are obtained with additional correlation among covariates, which are not displayed in here.

Concerning sensitivity of the choice of tuning parameters (K, τ) for the proposed method, as illustrated in Figure 4, the choice of τ is much less sensitive than that of K. Moreover, when

$K \geq K^0$, both the size and power become less sensitive to a change of K. With regard to the estimated K by cross-validation, the estimator $\widehat{K}$ is close to $K^0 = 3$ in the linear regression example, as suggested by Table 2.

In summary, our simulation results suggest that the proposed method achieves high power compared to its competitors Janková and Van de Geer (2017), Zhang and Zhang (2014), Van de Geer et al. (2014), and Zhang and Cheng (2017). Moreover, the asymptotic approximation seems adequate in all the examples.

6. Brain Network Analysis

Alzheimer's disease is the most common dementia without cure, while the prevalence is projected to continuously increase with an estimated 11% of the US senior population in 2015 to 16% in 2050, costing over 1.1 trillion in 2050 Alzheimer's Association (2016). AD is now widely believed to be a disease with disrupted brain networks, and cortical networks based in structural MRI have been constructed to contrast with that of normal/healthy controls (He, Chen, and Evans 2008). Using the ADNI-1 baseline data (adni.loni.usc.edu), we extracted the cortical thicknesses for $p = 68$ regions of interest (ROIs) based on the Desikan–Killany atlas Desikan et al. (2006). Since previous studies (e.g., Greicius et al. 2004; Montembeault et al. 2015) have identified the DMN to be associated with AD, we will pay particular attention to this subnetwork, which includes 12 ROIs in our dataset. As in He, Chen, and Evans (2008), we first regress the cortical thickness on five covariates (gender, handedness, education, age, and intercranial volume measured at baseline), then use the residuals to estimate precision matrices, for 145 AD patients and 182 normal controls (CNs), respectively. Our approach here differs from previous studies He, Chen, and Evans (2008) and Montembeault et al. (2015) not only in estimating precision matrices, instead of covariance matrices, but also in rigorous inference.

For this data, we consider a hypothesis test of $H_0 : \omega_{ij} = 0$ versus $H_a : \omega_{ij} \neq 0$; $1 \leq i \neq j \leq 12$. For each estimated network for the two groups, significant edges under the overall error rate $\alpha = 0.05$, after Bonferroni correction, are reported for the proposed CMLR test and the debias test of Janková and Van de Geer (2017) or *JG*. As indicated in Figure 5, the CMLR test yields 28 and 33 signif icant edges for the two groups of CN and AD, which is in contrast to 29 and 28 significant edges by the *JG* test. In other words, the CMLR test detects slightly more edges than the *JG* test, which is in agreement of the simulation results in Table 1.

In what follows, we will focus on scientific interpretations of the statistical findings by the CMLR test. As shown in Montembeault et al. (2015), it is confirmed that for the AD patients, as compared to the normal controls, there seems to be reduced connectivity within DMN, but increased connectivity for some other ROIs, that is, the salience network and the executive network reported in Montembeault et al. (2015). Moreover, it seems that connectivity between the left and right brain within DMN somewhat deteriorates for the AD patients. To further explore the latter point, we then separately test the independence between each node in DMN and the other nodes outside DMN using the proposed CMLR test with the standard normal approximation. Specifically, for node i in DMN, we test H_0 :

$\omega_{ij} = 0$ for all $j \notin$ DMN versus $H_a: \omega_{ij} \neq 0$ for some $j \in$ DMN, where DMN denotes the set of 12 nodes in DMN. This amounts to $2 \times 12 = 24$ tests, with 12 tests for each group. Specifically, it is confirmed that for the group AD, only L-parahippocampal (left side) is independent of all the other nodes outside DMN; in contrast, for the CN group, in addition to L-parahippocampal, three other ROIs in DMN, L-medial prefrontal cortex, R-parahippocampal, and R-precuneus are independent of all the other nodes outside DMN.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank the editors, the associate editor, and anonymous referees for helpful comments and suggestions.

Funding

Research supported in part by NSF grants DMS-1415500, DMS-1712564, DMS-1721216, DMS-1712580, DMS-1721445, and DMS-1721445, NIH funding: NIH grants 1R01GM081535-01, 1R01GM126002, HL65462, and R01HL105397.

Appendix

The following lemmas provide some key results to be used subsequently. Detailed proofs of Lemmas 2–8 are provided in a online Supplementary materials due to space limit. Before proceeding, we introduce some notations. Given an index set $A \subseteq \{(i,j): 1 \leq i \leq j \leq p\}$, define CMLE $\widehat{\boldsymbol{\Omega}}_A$ as $\widehat{\boldsymbol{\Omega}}_A = \text{argmax}_{\boldsymbol{\Omega} \succ 0, \boldsymbol{\Omega}_{A^c} = 0} L_n(\boldsymbol{\Omega})$, with $\succ$ indicating positive definiteness of a matrix. Worthy of note is that $\widehat{\boldsymbol{\Omega}}_A$ becomes the oracle estimator when $A = A^0$, where $A^0 = \left\{(i,j): i \leq j, \omega_{ij}^0 \neq 0\right\}$ is the index set including all the indices corresponding to nonzero entries of the true precision matrix $\boldsymbol{\Omega}^0 = (\omega_{ij}^0)_{p \times p}$

Lemma 2.

For any symmetric matrices $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$, $\text{vec}(\boldsymbol{C}_1)^\top \text{vec}(\boldsymbol{C}_2) = \text{tr}(\boldsymbol{C}_1 \boldsymbol{C}_2)$. Moreover, for any positive definite matrix $\boldsymbol{C} \succ 0$,

$$\nabla(\text{logdet}\boldsymbol{C}) = -\text{vec}(\boldsymbol{C}^{-1}),$$

$$\nabla^2(-\text{logdet}\boldsymbol{\Omega}^0) = \boldsymbol{C}^{-1} \otimes_s \boldsymbol{C}^{-1}, \tag{A.1}$$

$$\boldsymbol{I} = \frac{1}{2}\boldsymbol{\Sigma}^0 \otimes_s \boldsymbol{\Sigma}^0, \tag{A.2}$$

$$\text{var}(\text{vec}(\boldsymbol{X}\boldsymbol{X}^\top)) = 4\boldsymbol{I} \text{ with } \boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}^0), \tag{A.3}$$

$$\mathrm{vec}(\boldsymbol{C})^{\top}\boldsymbol{I}\,\mathrm{vec}(\boldsymbol{C}) = \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^0\boldsymbol{C}\boldsymbol{\Sigma}^0\boldsymbol{C})\,. \tag{A.4}$$

Lemma 3.

For any symmetric matrix $\boldsymbol{T}$ and $\nu > 0$

$$\mathbb{P}\left(\left|\mathrm{tr}((\boldsymbol{S}-\boldsymbol{\Sigma}^0)\boldsymbol{T})\right| \geq v\right) \leq 2\exp\left(-n\frac{v^2}{9\|\boldsymbol{T}\|^2 + 8v\|\boldsymbol{T}\|}\right), \tag{A.5}$$

where $\|\boldsymbol{T}\|^2 = \frac{n}{2}\mathrm{var}(\mathrm{tr}((\boldsymbol{S}-\boldsymbol{\Sigma}^0)\boldsymbol{T}))$. Furthermore, for $\boldsymbol{T}_1,\ldots,\boldsymbol{T}_K$ such that $\|\boldsymbol{T}_k\| \leq c_0; k = 1,\ldots,K$ with $c_0 > 0$ and any $\nu > 0$, we have that

$$\begin{aligned}&\mathbb{P}\left(\max_{1\leq k\leq K}\left|\mathrm{tr}\left(\left(\boldsymbol{S}-\boldsymbol{\Sigma}^0\right)\boldsymbol{T}_k\right)\right| \geq v\right)\\ &\quad\leq 2\exp\left(-n\frac{v^2}{9c_0^2 + 8c_0 v} + \log K\right),\end{aligned} \tag{A.6}$$

which implies that $\max_{1\leq k\leq K}\left|\mathrm{tr}\left(\left(\boldsymbol{S}-\boldsymbol{\Sigma}^0\right)\boldsymbol{T}_k\right)\right| = O_p\left(c_0\sqrt{\frac{\log K}{n}}\right)$. Particularly, for any $\nu > 0$ and any index set B,

$$\begin{aligned}&\mathbb{P}\left(\left\|\mathrm{vec}_B\left(\boldsymbol{S}-\boldsymbol{\Sigma}^0\right)\right\|_\infty \geq v\right)\\ &\quad\leq 2\exp\left(-n\frac{v^2\,|}{9\lambda_{\max}^2\left(\boldsymbol{\Sigma}^0\right) + 8v\lambda\max\left(\boldsymbol{\Sigma}^0\right)} + \log\left|B\right|\right),\end{aligned} \tag{A.7}$$

implying that $\left\|\mathrm{vec}_B\left(\boldsymbol{S}-\boldsymbol{\Sigma}^0\right)\right\|_\infty = O_p\left(\lambda_{\max}\left(\boldsymbol{\Sigma}^0\right)\sqrt{\frac{\log|B|}{n}}\right)$.

Lemma 4.

(The Kullback–Leibler divergence and Fisher-norm) For a positive definite matrix $\boldsymbol{\Omega} \in \mathbb{R}^{p\times p}$, a connection between the Kullback–Leibler divergence $K(\boldsymbol{\Omega}^0, \boldsymbol{\Omega})$ and the Fisher-norm $\|\boldsymbol{\Omega}^0 - \boldsymbol{\Omega}\|$ can be established:

$$K\left(\boldsymbol{\Omega}^0, \boldsymbol{\Omega}\right) \geq \min\left(\frac{1}{16\sqrt{2}}, \frac{\sqrt{K\left(\boldsymbol{\Omega}^0, \boldsymbol{\Omega}\right)}}{2\sqrt{6}}\right)\|\boldsymbol{\Omega}^0 - \boldsymbol{\Omega}\|, \tag{A.8}$$

$$K\left(\boldsymbol{\Omega}^0, \boldsymbol{\Omega}\right) \geq \min\left(\frac{1}{16\sqrt{2}}, \frac{\|\boldsymbol{\Omega}^0 - \boldsymbol{\Omega}\|}{24}\right)\|\boldsymbol{\Omega}^0 - \boldsymbol{\Omega}\|\,. \tag{A.9}$$

Lemma 5.

(Rate of convergence of constrained MLE). Let $\tilde{A} \supseteq A^0$ be an index set. For $\widehat{\boldsymbol{\Omega}}_{\tilde{A}}$, we have that

$$\left\|\widehat{\boldsymbol{\Omega}}_{\tilde{A}} - \boldsymbol{\Omega}^0\right\| \leq 12\left\|\boldsymbol{I}_{\tilde{A},\tilde{A}}^{-1/2}\mathrm{vec}\left(\boldsymbol{\Sigma}^0 - \boldsymbol{S}\right)\right\|_2. \tag{A.10}$$

on the event that $\left\{\left\|I_{\tilde{A},\tilde{A}}^{-1/2}\mathrm{vec}_{\tilde{A}}\left(\boldsymbol{\Sigma}^0 - \boldsymbol{S}\right)\right\|_2 < \frac{1}{8\sqrt{2}}\right\}$. Moreover, if $\frac{|\tilde{A}|\log p}{n} \to 0$, then

$$\left\|\widehat{\boldsymbol{\Omega}}_{\tilde{A}} - \boldsymbol{\Omega}^0\right\| = O_p\left(\sqrt{\frac{\left|\tilde{A}\right|\log p}{n}}\right). \tag{A.11}$$

Lemma 6.

(Selection consistency). If

$$\begin{aligned} &K = \left|A^0\right|, \tau \leq \frac{\bar{\lambda}_{\min}\min\left(\sqrt{C_{\min}}, C_{\min}^2\right)}{12\left|A^0\right|}, \text{then} \\ &\max\left(P\left(\widehat{\boldsymbol{\Omega}}^{(0)} \neq \widehat{\boldsymbol{\Omega}}_{A^0}\right), P\left(\widehat{\boldsymbol{\Omega}}^{(1)} \neq \widehat{\boldsymbol{\Omega}}_{A^0 \cup B}\right)\right) \\ &\leq 2\exp\left(\frac{-nC_{\min}}{2560 \times 512} + 2\log p\right) + \exp\left(\frac{-n}{2560} + \left|A^0\right|\log p\right) \\ &+ 2\exp\left(-n\frac{\min\left(\sqrt{\frac{\min(C_{\min}/512, 3/32)}{48\lambda_{\max}^2\left(\left|A^0\right| + \left|B\right|\right)}}, \lambda_{\max}\left(\boldsymbol{\Sigma}^0\right)\right)^2}{18\lambda_{\max}^2\left(\boldsymbol{\Sigma}^0\right)} + 2\log p\right) \\ &\to 0 \end{aligned} \tag{A.12}$$

as $n \to \infty$ under Assumptions 1 and 2, where $\widehat{\boldsymbol{\Omega}}^{(0)}, \widehat{\boldsymbol{\Omega}}^{(1)}$, and $C_{\min}$ are as defined in (1)–(3).

Lemma 7.

Let $\boldsymbol{\gamma}_k = (\gamma_{k1}, ., \cdot \cdot \gamma_{km}) \in \mathbb{R}^m; k = 1, \ldots, n$ be iid random vectors with $\mathrm{var}(\boldsymbol{\gamma}_1) = \boldsymbol{I}_{m \times m}$. If m is fixed, then

$$n^{-1}\left\|\sum_{k=1}^{n}\boldsymbol{\gamma}_k\right\|_2^2 \xrightarrow{d} \chi_m^2, \text{ as } n \to \infty. \tag{A.13}$$

Otherwise, if max (m, $m_2 m/n$, m_3/n, $m_3 m^{3/2}/n^2 \to 0$), where $m_j = \max_{1 \leq i \leq m}\mathbb{E}\gamma_{1i}^{2j}; j = 2, 3$, then

$$\frac{\left\|\sum_{k=1}^{n}\boldsymbol{\gamma}_k\right\|_2^2 - nm}{n\sqrt{2m}} \xrightarrow{d} N(0, 1), \text{ as } n \to \infty. \tag{A.14}$$

Lemma 8.

Let $\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^0)$ and $\gamma = \mathrm{tr}\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{\Sigma}^0\right)\boldsymbol{T}\right)$ with $\boldsymbol{T}$ a symmetric matrix. Then

$$\mathbb{E}\left(\gamma^{2m}\right) \leq (2m-1)!2^{m-1}\left(\mathbb{E}\left(\gamma^2\right)\right)^m \text{ for any integer } m \geq 1. \tag{A.15}$$

Lemma 9.

(Asymptotic distribution for log-likelihood ratios). The log-likelihood ratio statistic $Lr = 2(L_n(\widehat{\boldsymbol{\Omega}}_{\widetilde{A}}) - L_n(\widehat{\boldsymbol{\Omega}}_{A^0}))$, where $\widehat{\boldsymbol{\Omega}}_{\widetilde{A}}$ is the MLE over index set $\widetilde{A}$ with $\widetilde{A} \supseteq A^0$. Denote by κ_0 the condition number of $\boldsymbol{\Sigma}^0$. If $\frac{\kappa_0|\widetilde{A}|\log p}{\sqrt{n}} \to 0$ with $p \geq 2$, then,

$$Lr \xrightarrow{P_0} W_{|B|}, \text{ if } \left|B\right| \text{ is a constant; } \frac{Lr - |B|}{\sqrt{2|B|}} \xrightarrow{P_0} Z, \text{ if } \left|B\right| \to \infty,$$

where $B = \widetilde{A} \backslash A^0$, $W_{|B|}$ follows a chi-square distribution χ^2 on $|B|$ degrees of freedom and $Z \sim \mathcal{N}(0, 1)$, respectively.

Proof of Theorem 1.

By Lemma 6, $\mathbb{P}\left(\widehat{\boldsymbol{\Omega}}^{(0)} = \widehat{\boldsymbol{\Omega}}_{A^0}\right) \to 1$; $\mathbb{P}\left(\widehat{\boldsymbol{\Omega}}^{(1)} = \widehat{\boldsymbol{\Omega}}_{A^0 \cup B}\right) \to 1$, as $n \to \infty$ under Assumptions 1 and 2. Then, the asymptotic distribution of the likelihood ratio follows immediately from Lemma 9. □

Proof of Proposition 1.

Let $\widetilde{A} = A^0 \cup B$. By Lemma 6, $\mathbb{P}\left(\widehat{\boldsymbol{\Omega}}^{(1)} = \widehat{\boldsymbol{\Omega}}_{A^0 \cup B}\right) \to 1$, as $n \to \infty$. Asymptotic normality of $\mathrm{vec}_B\left(\widehat{\boldsymbol{\Omega}}_{A^0 \cup B}\right)$ follows from an expansion of the score equation. Specifically, note that

$$\sqrt{n}\,\mathrm{vec}_B\left(\widehat{\boldsymbol{\Omega}}_{A^0 \cup B} - \boldsymbol{\Omega}^0\right) = \frac{\sqrt{n}}{2}\left[\boldsymbol{I}^{-1}_{\widetilde{A},\widetilde{A}}\right]_{B,\widetilde{A}} \times \left(\mathrm{vec}_{\widetilde{A}}(\boldsymbol{\Lambda}) - \mathrm{vec}_A\left(R\left(\widehat{\boldsymbol{\Delta}}_{\widetilde{A}}\right)\right)\right),$$

where $R\left(\widehat{\boldsymbol{\Delta}}_{\widetilde{A}}\right) = \boldsymbol{\Sigma}^0\sum_{i=2}^{\infty}(-1)^i\left(\widehat{\boldsymbol{\Delta}}_A\boldsymbol{\Sigma}^0\right)^i$. Let $\boldsymbol{J} = \boldsymbol{I}^{-1}_{\widetilde{A},\widetilde{A}}$ be as defined in (B.33) of the online supplementary material. Multiplying $\boldsymbol{J}^{-1/2}_{B,B}$ on both sides of this identity, we obtain

$$\begin{aligned}&\sqrt{n}\boldsymbol{J}^{-1/2}_{B,B}\mathrm{vec}_B\left(\widehat{\boldsymbol{\Omega}}_{A^0 \cup B} - \boldsymbol{\Omega}^0\right)\\ &= \frac{\sqrt{n}}{2}\boldsymbol{J}^{-1/2}_{B,B}\boldsymbol{J}_{B,\widetilde{A}}\left(\mathrm{vec}_{\widetilde{A}}(\boldsymbol{\Lambda}) - \mathrm{vec}_{\widetilde{A}}\left(R\left(\widehat{\boldsymbol{\Delta}}_{\widetilde{A}}\right)\right)\right).\end{aligned} \tag{A.16}$$

Next, we show that the first term tends to $N(\mathbf{0}, \boldsymbol{I}_{|B|\times|B|})$ in distribution and the second term tends to 0 in probability. For the second term, following similar calculations as in (B.34) of the online supplementary material, we have that $\left\|\boldsymbol{J}_{B,B}^{-1/2}\boldsymbol{J}_{B,\widetilde{A}}\boldsymbol{x}\right\|_2^2 = \boldsymbol{x}^\top\boldsymbol{J}\boldsymbol{x} - \boldsymbol{x}^\top\boldsymbol{I}_{A^0,A^0}^{-1}\boldsymbol{x} \le \boldsymbol{x}^\top\boldsymbol{J}\boldsymbol{x} \le \lambda_{\min}^{-2}\left(\boldsymbol{\Sigma}^0\right)\left\|\boldsymbol{x}\right\|_2^2$ for any $\boldsymbol{x}\in\mathbb{R}^{|A|}$. This, together with (B.37) of the online supplementary material, implies that

$$\begin{aligned}&\left\|.5\sqrt{n}\boldsymbol{J}_{B,B}^{-1/2}\boldsymbol{J}_{B,\widetilde{A}}\mathrm{vec}_{\widetilde{A}}\left(R\left(\widehat{\boldsymbol{\Delta}}_A\right)\right)\right\|_2 \le .5\sqrt{n}\left\|\boldsymbol{J}^{1/2}\mathrm{vec}_{\widetilde{A}}\left(R\left(\widehat{\boldsymbol{\Delta}}_{\widetilde{A}}\right)\right)\right\|_2\\ &\le .5\sqrt{n}\lambda_{\min}^{-1}\left(\boldsymbol{\Sigma}^0\right)\left\|R\left(\widehat{\boldsymbol{\Delta}}_{\widetilde{A}}\right)\right\|_2 \le \sqrt{n}\kappa_0\left\|\boldsymbol{\Sigma}^0\widehat{\boldsymbol{\Delta}}_{\widetilde{A}}\right\|_F^2\\ &= O_p\left(\frac{\kappa_0\left|\widetilde{A}\right|\log p}{\sqrt{n}}\right) = o_p(1)\end{aligned} \tag{A.17}$$

under Assumption 2. For the first term, note that

$$\begin{aligned}&\mathrm{cov}\left(\frac{1}{2}\boldsymbol{J}_{B,B}^{-1/2}\boldsymbol{J}_{B,\widetilde{A}}\mathrm{vec}_A\left(\boldsymbol{X}\boldsymbol{X}^\top-\boldsymbol{\Sigma}^0\right),\right.\\ &\left.\frac{1}{2}\boldsymbol{J}_{B,B}^{-1/2}\boldsymbol{J}_{B,\widetilde{A}}\mathrm{vec}_{\widetilde{A}}\left(\boldsymbol{X}\boldsymbol{X}^\top-\boldsymbol{\Sigma}^0\right)\right)\\ &= \boldsymbol{J}_{B,B}^{-1/2}\boldsymbol{J}_{B,\widetilde{A}}\mathrm{cov}\left(\frac{1}{2}\mathrm{vec}_{\widetilde{A}}\left(\boldsymbol{X}\boldsymbol{X}^\top-\boldsymbol{\Sigma}^0\right),\frac{1}{2}\mathrm{vec}_{\widetilde{A}}\left(\boldsymbol{X}\boldsymbol{X}^\top-\boldsymbol{\Sigma}^0\right)\right)\\ &\boldsymbol{J}_{\widetilde{A},B}\boldsymbol{J}_{B,B}^{-1/2}\\ &= \boldsymbol{J}_{B,B}^{-1/2}\boldsymbol{J}_{B,\widetilde{A}}\boldsymbol{I}_{\widetilde{A},\widetilde{A}}\boldsymbol{J}_{\widetilde{A},B}\boldsymbol{J}_{B,B}^{-1/2} = I_{|B|\times|B|}.\end{aligned}$$

where the second last equality uses the property of exponential family Brown (1986). Hence, by the central limit theorem, $\mathrm{vec}_{\widetilde{A}}(\boldsymbol{\Lambda}) \xrightarrow{d} N\left(0, \left[\boldsymbol{I}_{\widetilde{A},\widetilde{A}}^{-1}\right]_{B,B}\right)$. Finally, by Slutsky's Theorem, we obtain that $\sqrt{n}\,\mathrm{vec}_B\left(\widehat{\boldsymbol{\Omega}}_{A^0\cup B} - \boldsymbol{\Omega}^0\right) \xrightarrow{d} N\left(0, \left[\boldsymbol{I}_{\widetilde{A},\widetilde{A}}^{-1}\right]_{B,B}\right)$. This completes the proof. □

Proof of Proposition 2.

By Theorem 3 of Shen et al. (2013), $\mathbb{P}\left(\left\{\hat{\boldsymbol{\beta}}^{(1)} = \hat{\boldsymbol{\beta}}_{A^0\cup B}^{ls}\right\}\right) \to 1$, as $n, p \to \infty$. Hence, with probability tending to 1,

$$\begin{aligned}\hat{\boldsymbol{\beta}}_B^{(1)} &= \mathrm{vec}_B\left(\left(X_{A^0\cup B}^\top X_{A^0\cup B}\right)^{-1}X_{A^0\cup B}^\top Y\right)\\ &= \mathrm{vec}_B\left(\left(X_{A^0\cup B}^\top X_{A^0\cup B}\right)^{-1}X_{A^0\cup B}^\top\left(X_{A^0\cup B}\boldsymbol{\beta}_{A^0\cup B}^0+\epsilon\right)\right)\\ &= \boldsymbol{\beta}_B^0 + \mathrm{vec}_B\left(\left(X_{A^0\cup B}^\top X_{A^0\cup B}\right)^{-1}X_{A^0\cup B}^\top\epsilon\right).\end{aligned}$$

Simple moment generating function calculations show that when $|B|$ is fixed,

$$\mathrm{vec}_B\left(\left(X_{A^0\cup B}^\top X_{A^0\cup B}\right)^{-1}X_{A^0\cup B}^\top\epsilon\right) \sim N\left(\mathbf{0}, \left[\left(X_{A^0\cup B}^\top X_{A^0\cup B}\right)^{-1}\right]_{B,B}\right).$$

Hence, $\sqrt{n}\left(\hat{\boldsymbol{\beta}}_B^{(1)} - \boldsymbol{\beta}_B^0\right) \xrightarrow{d} N\left(\mathbf{0}, \left[\left(n^{-1} X_{A^0 \cup B}^\top X_{A^0 \cup B}\right)^{-1}\right]_{B,B}\right)$.. This completes the proof. □

Proof of Corollary 1.

Let $\widetilde{A} = A^0 \cup B$. The result follows directly from Theorem 1. Specifically, we bound the asymptotic covariance matrix of $\left[\sqrt{n}\left(\hat{\omega}_{ij} - \omega_{ij}^0\right)\right]_{(i,j) \in B}$ for any B of fixed size. Note that the asymptotic covariance matrix of $\sqrt{n}\,\mathrm{vec}_B\left(\hat{\boldsymbol{\Omega}}_{\widetilde{A}} - \boldsymbol{\Omega}^0\right)$ can be bounded: $\left[\boldsymbol{I}_{\widetilde{A},\widetilde{A}}^{-1}\right]_{B,B} \preceq \left[\boldsymbol{I}^{-1}\right]_{B,B} = 2\left[\boldsymbol{\Omega}^0 \otimes_s \boldsymbol{\Omega}^0\right]_{B,B}$. Moreover, for any $(i,j),(i',j') \in B$, $2\left[\boldsymbol{\Omega}^0 \otimes_s \boldsymbol{\Omega}^0\right]_{(i,j),(i',j')}$ can be written as

$$\begin{aligned}&\frac{\sqrt{1+\mathbb{I}(i\neq j)}\sqrt{1+\mathbb{I}(i'\neq j')}}{2}\mathrm{tr}\\&\times\left(\left(\boldsymbol{e}_i\boldsymbol{e}_j^\top+\boldsymbol{e}_j\boldsymbol{e}_i^\top\right)\boldsymbol{\Omega}^0\left(\boldsymbol{e}_{i'}\boldsymbol{e}_{j'}^\top+\boldsymbol{e}_{j'}\boldsymbol{e}_{i'}^\top\right)\boldsymbol{\Omega}^0\right)\\&=\sqrt{1+\mathbb{I}(i\neq j)}\sqrt{1+\mathbb{I}(i'\neq j')}\left(\omega_{i'j}^0\omega_{ij'}^0+\omega_{jj'}^0\omega_{ii'}^0\right).\end{aligned}$$

Using $\mathrm{vec}_B(C) = \left(\sqrt{1+\mathbb{I}(i\neq j)}c_{ij}\right)_{(i,j)\in B}$, the asymptotic variance of $\left[\sqrt{n}\left(\hat{\omega}_{ij} - \omega_{ij}^0\right)\right]_{(i,j)\in B}$ is upper bounded by a $|B| \times |B|$ matrix $\left[\omega_{i'j}^0\omega_{ij'}^0 + \omega_{jj'}^0\omega_{ii'}^0\right]_{(i,j)\in B,(i',j')\in B}$. Particularly, when $B = \{(i,j)\}$, this reduces to an upper bound on the asymptotic variance $\left[\omega_{ij}^0\right]^2 + \omega_{ii}^0\omega_{jj}^0$. This completes the proof. □

Proof of Theorem 2.

By Theorem 3 of Shen et al. (2013), $\mathbb{P}\left(\left\{\hat{\boldsymbol{\beta}}^{(1)} = \hat{\boldsymbol{\beta}}_{A^0\cup B}^{ls}\right\} \cap \hat{\boldsymbol{\beta}}^{(0)} = \hat{\boldsymbol{\beta}}_{A^0}^{ls}\right) \to 1$, as $n, p \to \infty$, by Assumption 1, where $\hat{\boldsymbol{\beta}}_A^{ls}$ is the least square estimate over A. Hence, in what follows, we focus our attention to event $\left\{\hat{\boldsymbol{\beta}}^{(1)} = \hat{\boldsymbol{\beta}}_{A^0\cup B}^{ls}\right\} \cap \left\{\hat{\boldsymbol{\beta}}^{(0)} = \hat{\boldsymbol{\beta}}_{A^0}^{ls}\right\}$.

Easily, after profiling out σ, we have $\Lambda_n(B) = n\left(\log\left(\left\|y - X\hat{\boldsymbol{\beta}}^{(0)}\right\|_2^2\right) - \log\left(\left\|y - X\hat{\boldsymbol{\beta}}^{(1)}\right\|_2^2\right)\right)$. Then an application of Taylor's expansion of $\log(1-x)$ yields that

$$\begin{aligned}&n\left(\log\left(\|y - X\boldsymbol{\beta}\|_2^2\right) - \log\left(\left\|y - X\boldsymbol{\beta}^0\right\|_2^2\right)\right)\\&= -n\sum_{i=1}^{\infty}\frac{\left(2\boldsymbol{\epsilon}^\top X\boldsymbol{\delta} - \|X\boldsymbol{\delta}\|_2^2\right)^i}{i\|\boldsymbol{\epsilon}\|_2^{2i}}\end{aligned} \tag{A.18}$$

where $\boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^0$. Moreover, on the event $\left\{\hat{\boldsymbol{\beta}}^{(1)} = \hat{\boldsymbol{\beta}}_{A^0\cup B}^{ls}\right\} \cap \left\{\hat{\boldsymbol{\beta}}^{(0)} = \hat{\boldsymbol{\beta}}_{A^0}^{ls}\right\}$,

$$\hat{\boldsymbol{\beta}}^{(1)} = \boldsymbol{\beta}^0 + \left(X_{A^0 \cup B}^\top X_{A^0 \cup B}\right)^{-1} X_{A^0 \cup B}^\top \boldsymbol{\epsilon} \text{ and}$$
$$\hat{\boldsymbol{\beta}}^{(0)} = \boldsymbol{\beta}^0 + \left(X_{A^0}^\top X_{A^0}\right)^{-1} X_{A^0}^\top \boldsymbol{\epsilon},$$

implying that $X\left(\hat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^0\right) = \boldsymbol{P}_{A^0 \cup B}\boldsymbol{\epsilon}$ and $X\left(\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^0\right) = \boldsymbol{P}_{A^0}\boldsymbol{\epsilon}$. Consequently, replacing $\boldsymbol{\delta} = \hat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^0$, the right-hand of (A.18) reduces to

$$-n\sum_{i=1}^{\infty} \frac{\left(\boldsymbol{\epsilon}^\top \boldsymbol{P}_{A^0 \cup B}\boldsymbol{\epsilon}\right)^i}{i\|\boldsymbol{\epsilon}\|_2^{2i}} = -\frac{n}{\|\boldsymbol{\epsilon}\|_2^2} \times \left(\boldsymbol{\epsilon}^\top \boldsymbol{P}_{A^0 \cup B}\boldsymbol{\epsilon} + \sum_{i=2}^{\infty} \frac{\left(\boldsymbol{\epsilon}^\top \boldsymbol{P}_{A^0 \cup B}\boldsymbol{\epsilon}\right)^i}{i\|\boldsymbol{\epsilon}\|_2^{2(i-1)}}\right).$$

Similarly, replacing $\boldsymbol{\delta}$ by $\hat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^0$, (A.18) becomes $-\frac{n}{\|\boldsymbol{\epsilon}\|_2^2}\left(\boldsymbol{\epsilon}^\top \boldsymbol{P}_{A^0}\boldsymbol{\epsilon} + \sum_{i=2}^{\infty} \frac{\left(\boldsymbol{\epsilon}^\top \boldsymbol{P}_{A^0}\boldsymbol{\epsilon}\right)^i}{i\|\boldsymbol{\epsilon}\|_2^{2(i-1)}}\right)$.

Taking the difference leads to that $\Lambda_n(B) = \frac{n\boldsymbol{\epsilon}^\top(\boldsymbol{P}_{A^0 \cup B} - \boldsymbol{P}_{A^0})\boldsymbol{\epsilon}}{\|\boldsymbol{\epsilon}\|_2^2} + R(\boldsymbol{\epsilon})$, where $R(\boldsymbol{\epsilon})$ is

$$\sum_{i=2}^{\infty} \frac{\left(\boldsymbol{\epsilon}^\top \boldsymbol{P}_{A^0 \cup B}\boldsymbol{\epsilon}\right)^i - \left(\boldsymbol{\epsilon}^\top \boldsymbol{P}_{A^0}\boldsymbol{\epsilon}\right)^i}{i\|\boldsymbol{\epsilon}\|_2^{2(i-1)}} = \sum_{i=2}^{\infty} \frac{\boldsymbol{\epsilon}^\top(\boldsymbol{P}_{A^0 \cup B} - \boldsymbol{P}_{A^0})\boldsymbol{\epsilon}\left(\sum_{j=0}^{i-1}\left(\boldsymbol{\epsilon}^\top \boldsymbol{P}_{A^0 \cup B}\boldsymbol{\epsilon}\right)^j\left(\boldsymbol{\epsilon}^\top \boldsymbol{P}_{A^0}\boldsymbol{\epsilon}\right)^{i-j-1}\right)}{i\|\boldsymbol{\epsilon}\|_2^{2(i-1)}}.$$

Note that $\boldsymbol{P}_{A^0 \cup B} - \boldsymbol{P}_{A^0}$ is idempotent with the rank $|B|$. Moreover, $\boldsymbol{\epsilon}^\top \boldsymbol{P}_{A^0}\boldsymbol{\epsilon} \le \boldsymbol{\epsilon}^\top \boldsymbol{P}_{A^0 \cup B}\boldsymbol{\epsilon}$. Thus, $R(\boldsymbol{\epsilon})$ is no greater than

$$\boldsymbol{\epsilon}^\top(\boldsymbol{P}_{A^0 \cup B} - \boldsymbol{P}_{A^0})\boldsymbol{\epsilon}\sum_{i=2}^{\infty}\left(\frac{\boldsymbol{\epsilon}^\top \boldsymbol{P}_{A^0 \cup B}\boldsymbol{\epsilon}}{\|\boldsymbol{\epsilon}\|_2^2}\right)^{i-1} = \boldsymbol{\epsilon}^\top(\boldsymbol{P}_{A^0 \cup B} - \boldsymbol{P}_{A^0})\boldsymbol{\epsilon}\frac{\boldsymbol{\epsilon}^\top \boldsymbol{P}_{A^0 \cup B}\boldsymbol{\epsilon}}{\|\boldsymbol{\epsilon}\|_2^2}\left(1 - \frac{\boldsymbol{\epsilon}^\top \boldsymbol{P}_{A^0 \cup B}\boldsymbol{\epsilon}}{\|\boldsymbol{\epsilon}\|_2^2}\right)^{-1}$$

on the event that $\left\{\boldsymbol{\epsilon}^\top \boldsymbol{P}_{A^0 \cup B}\boldsymbol{\epsilon} < \|\boldsymbol{\epsilon}\|_2^2\right\}$. This, together with the facts that $n/\left\|\boldsymbol{\epsilon}\right\|_2^2 \xrightarrow{\mathbb{P}} 1$ and $\left|A^0\right|/n \to 0$, implies that $\Lambda_n(B) \xrightarrow{d} \chi^2(|B|)$ when $|B|$ is fixed, and $\frac{\Lambda_n(B) - |B|}{\sqrt{2|B|}} \xrightarrow{d} N(0,1)$ when $|B| \to \infty$ and $\frac{\sqrt{|B|}\left(\left|A^0\right| + |B|\right)}{n} \to 0$, because

$$R(\epsilon)/\sqrt{|B|} \leq \frac{\epsilon^\top (P_{A^0 \cup B} - P_{A^0})\epsilon}{\sqrt{|B|}} \frac{\epsilon^\top P_{A^0 \cup B}\epsilon}{\|\epsilon\|_2^2} \times \left(1 - \frac{\epsilon^\top P_{A^0 \cup B}\epsilon}{\|\epsilon\|_2^2}\right)^{-1} \xrightarrow{\mathbb{P}} 0$$

provided that $\frac{\sqrt{|B|}\left(\left|A^0\right| + |B|\right)}{n} \to 0$ and $|B| \to \infty$. This completes the proof. □

References

Alizadeh F, Haeberly JA, and Overton ML (1998), "Primal-Dual Interior-Point Methods for Semidefinite Programming: Convergence Rates, Stability and Numerical Results," SIAM Journal on Optimization, 8, 746–768. [220]

Alzheimer's Association (2016). "Changing the Trajectory of Alzheimer's Disease: How a Treatment by 2025 Saves Lives and Dollars," [225]

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. [218,221]

Brown LD (1986), Fundamentals of Statistical Exponential Families With Applications in Statistical Decision Theory (Lecture Notes-Monograph Series), Durham, NC: Duke University Press, pp. 1–279. [228]

Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, Buckner RL, Dale AM, Maguire RP, and Hyman BT (2006), "An Automated Labeling System for Subdividing the Human Cerebral Cortex on MRI Scans Into Gyral Based Regions of Interest," Neuroimage, 31, 968–980. [225] [PubMed: 16530430]

Fan J, Feng Y, and Wu Y (2009), "Network Exploration via the Adaptive LASSO and SCAD Penalties," The Annals of Applied Statistics, 3, 521–541. [217] [PubMed: 21643444]

Fan J, and Li R (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," Journal of the American Statistical Association, 96, 1348–1360. [217]

Friedman J, Hastie T, and Tibshirani R (2008), "Sparse Inverse Covariance Estimation With the Graphical Lasso," Biostatistics, 9, 432–441. [217] [PubMed: 18079126]

Greicius MD, Srivastava G, Reiss AL, and Menon V (2004), "Default-Mode Network Activity Distinguishes Alzheimer's Disease From Healthy Aging: Evidence From Functional MRI," Proceedings of the National Academy of Sciences of the United States of America, 101, 4637–4642. [225] [PubMed: 15070770]

He Y, Chen Z, and Evans A (2008), "Structural Insights Into Aberrant Topological Patterns of Large-Scale Cortical Networks in Alzheimer's Disease," The Journal of Neuroscience, 28, 4756–4766. [225,226] [PubMed: 18448652]

Janková J, and Van de Geer S (2017), "Honest Confidence Regions and Optimality in High-Dimensional Precision Matrix Estimation," TEST, 26, 143–162. [217,218,219,220,221,222,223,226]

Javanmard A, and Montanari A (2014), "Confidence Intervals and Hypothesis Testing for High-Dimensional Regression," Journal of Machine Learning Research, 15, 2869–2909. [217,218]

Li B, Chun H, and Zhao H (2012), "Sparse Estimation of Conditional Graphical Models With Application to Gene Networks," Journal of the American Statistical Association, 107, 152–167. [217] [PubMed: 24574574]

Lin Z, Wang T, Yang C, and Zhao H (2017), "On Joint Estimation of Gaussian Graphical Models for Spatial and Temporal Data," Biometrics, 73, 769–779. [217] [PubMed: 28099997]

Liu J, and Ye J (2009), "Efficient Euclidean Projections in Linear Time," in Proceedings of the 26th Annual International Conference on Machine Learning, pp. 657–664, ACM [221]

Meinshausen N, and Bühlmann P (2006), "High-Dimensional Graphs and Variable Selection With the Lasso," The Annals of Statistics, 34, 1436–1462. [217]

Montembeault M, Rouleau I, Provost JS, and Brambati SM (2015), "Altered Gray Matter Structural Covariance Networks in Early Stages of Alzheimer's Disease," Cerebral Cortex, 26, 2650–2662. [225,226] [PubMed: 25994962]

Portnoy S (1988), "Asymptotic Behavior of Likelihood Methods for Exponential Families When the Number of Parameters Tends to Infinity," The Annals of Statistics, 16, 356–366. [218,219]

Rothman A, Bickel P, Levina E, and Zhu J (2008), "Sparse Permutation Invariant Covariance Estimation," Electronic Journal of Statistics, 2, 494–515. [217]

Shen X (1997), "On Methods of Sieves and Penalization," The Annals of Statistics, 25, 2555–2591. [219]

Shen X, Pan W, and Zhu Y (2012), "Likelihood-Based Selection and Sharp Parameter Estimation," Journal of American Statistical Association, 107, 223–232. [217,218,219,221]

Shen X, Pan W, Zhu Y, and Zhou H(2013), "On Constrained and Regularized High-Dimensional Regression," Annals of the Institute of Statistical Mathematics, 65, 807–832. [218,220,228] [PubMed: 24465052]

Tibshirani R (1996), "Regression Shrinkage and Selection Via the Lasso," Journal of the Royal Statistical Society, Series B, 58, 267–288. [217]

Van de Geer S, Bühlmann P, Ritov Y, and Dezeure R (2014), "On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models," The Annals of Statistics, 42, 1166–1202. [217,218,223]

Yin J, and Li H (2013), "Adjusting for High-Dimensional Covariates in Sparse Precision Matrix Estimation by 1-Penalization," Journal of Multivariate Analysis, 116, 365–381. [217] [PubMed: 23687392]

Yuan M, and Lin Y (2007), "Model Selection and Estimation in the Gaussian Graphical Model," Biometrika, 94, 19–35. [217]

Zhang C (2010), "Nearly Unbiased Variable Selection Under Minimax Concave Penalty," The Annals of Statistics, 38, 894–942. [217]

Zhang C, and Zhang S (2014), "Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models," Journal of the Royal Statistical Society, Series B, 76, 217–242. [217,218,223,225]

Zhang X, and Cheng G (2017), "Simultaneous Inference for High-Dimensional Linear Models," Journal of the American Statistical Association, 112, 757–768. [223,225]

Zhu Y (2017), "An Augmented ADMM Algorithm With Application to the Generalized Lasso Problem," Journal of Computational and Graphical Statistics, 26, 195–204. [221]

Zhu Y, Shen X, and Pan W (2014), "Structural Pursuit Over Multiple Undirected Graphs," Journal of the American Statistical Association, 109, 1683–1696. [217] [PubMed: 25642006]

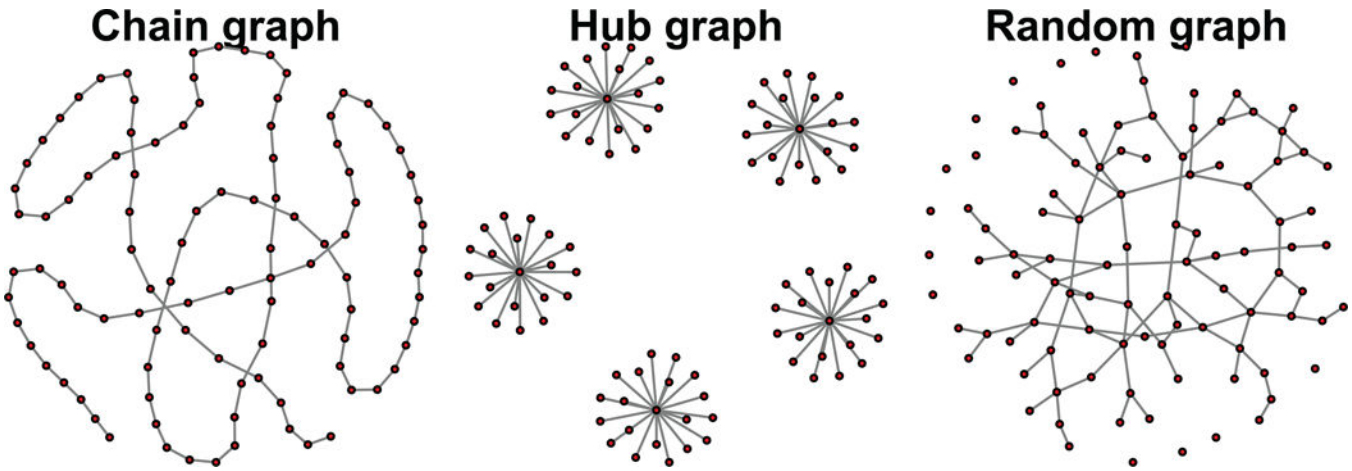

Figure 1.
Three types of graphs used in our simulations.

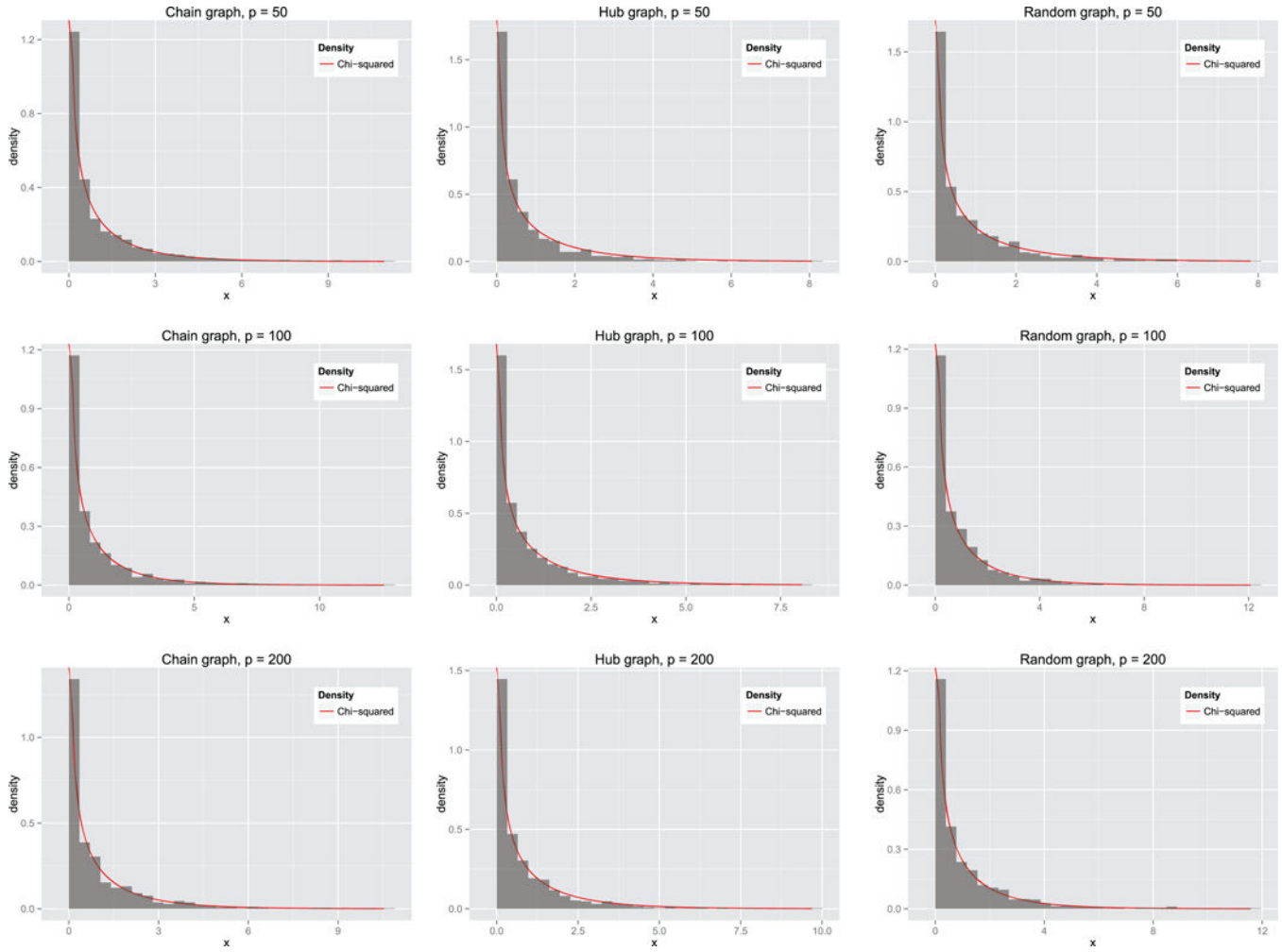

Figure 2.

Empirical null distribution of the proposed CMLR test based on the chi-square approximation with $n = 200$.

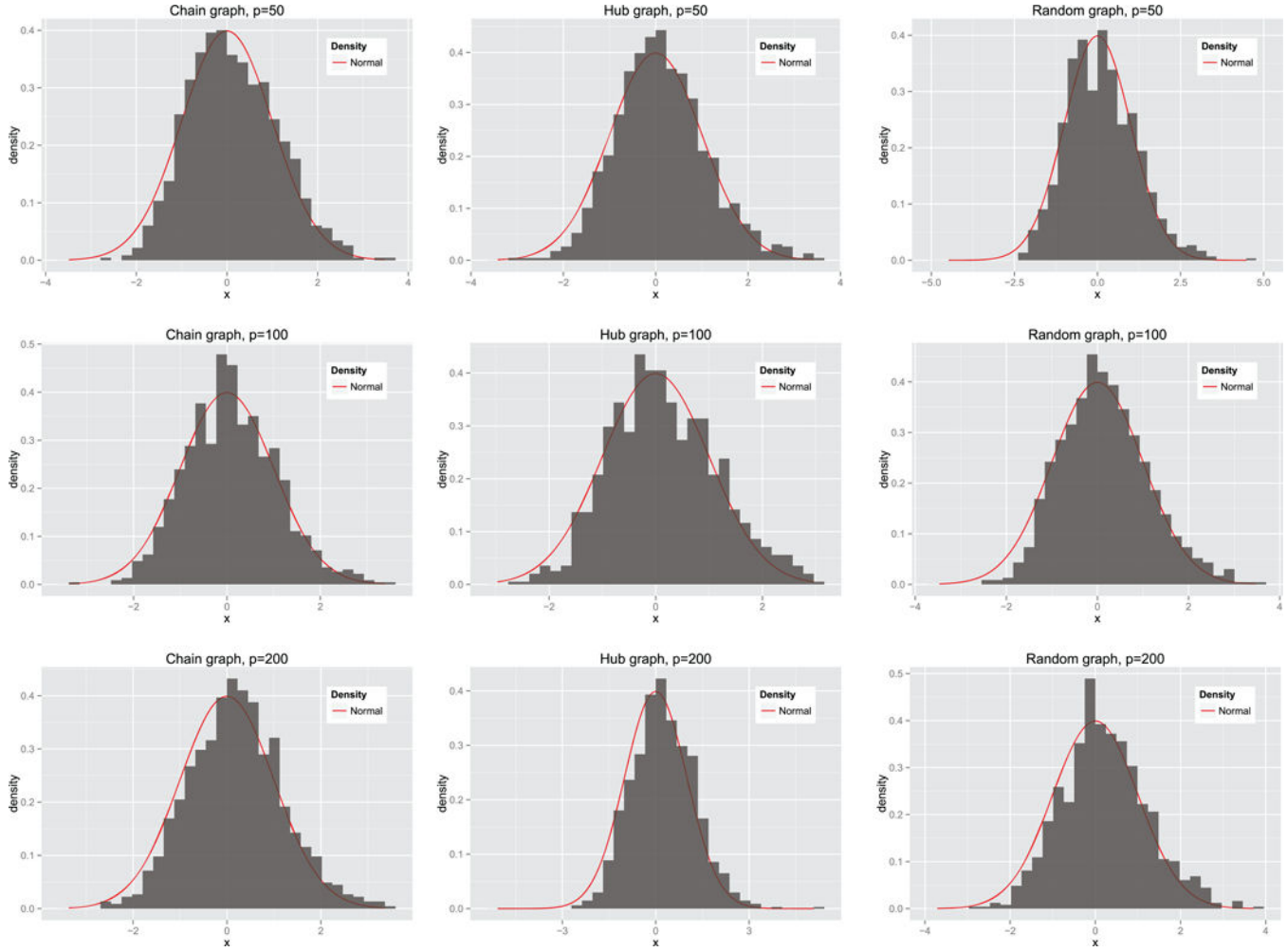

Figure 3.

Empirical null distribution of our likelihood ratio test based on the normal approximation for the second testing problem involving a single column/row.

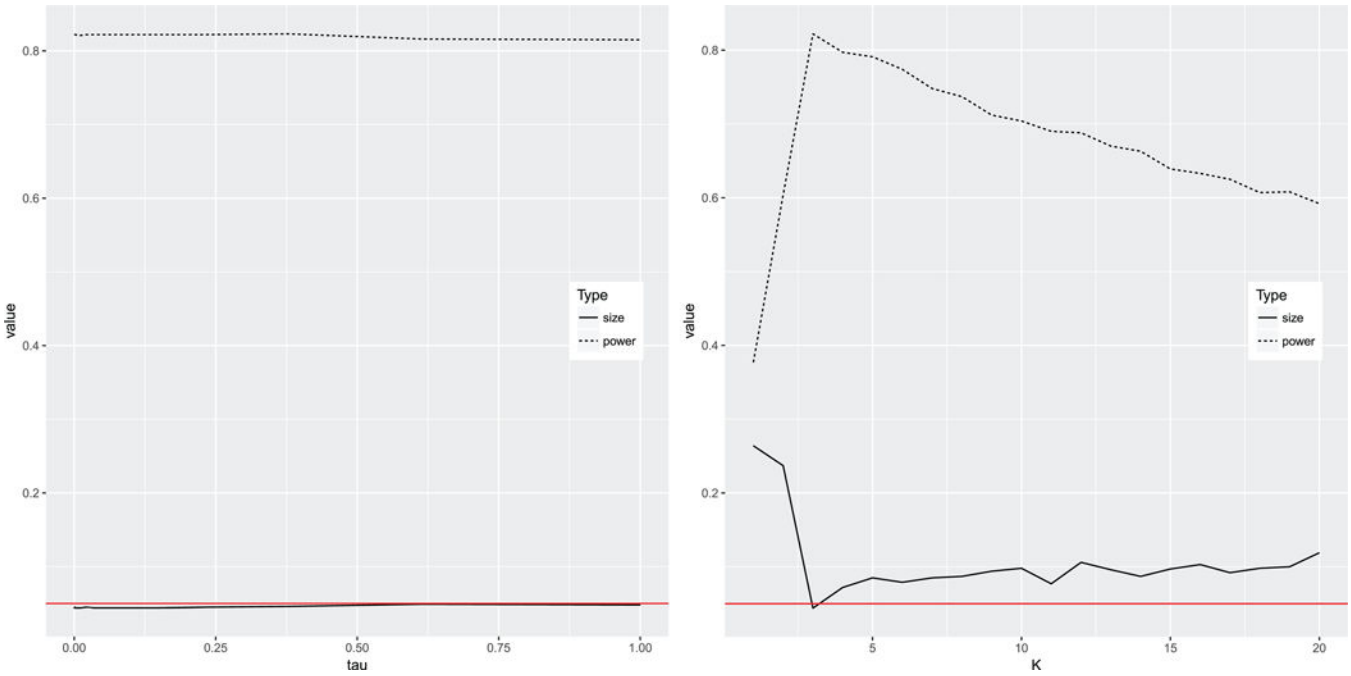

Figure 4.

Sensitivity study of power as a function of tuning parameters τ and K, when $n = 100$, $p = 100$, and $K_0 = 3$ in the linear regression problem based on 1000 simulations. Dotted and black lines represent empirical power and sizes of the proposed method, while red lines serve as a reference of the nominal size $\alpha = 0.05$.

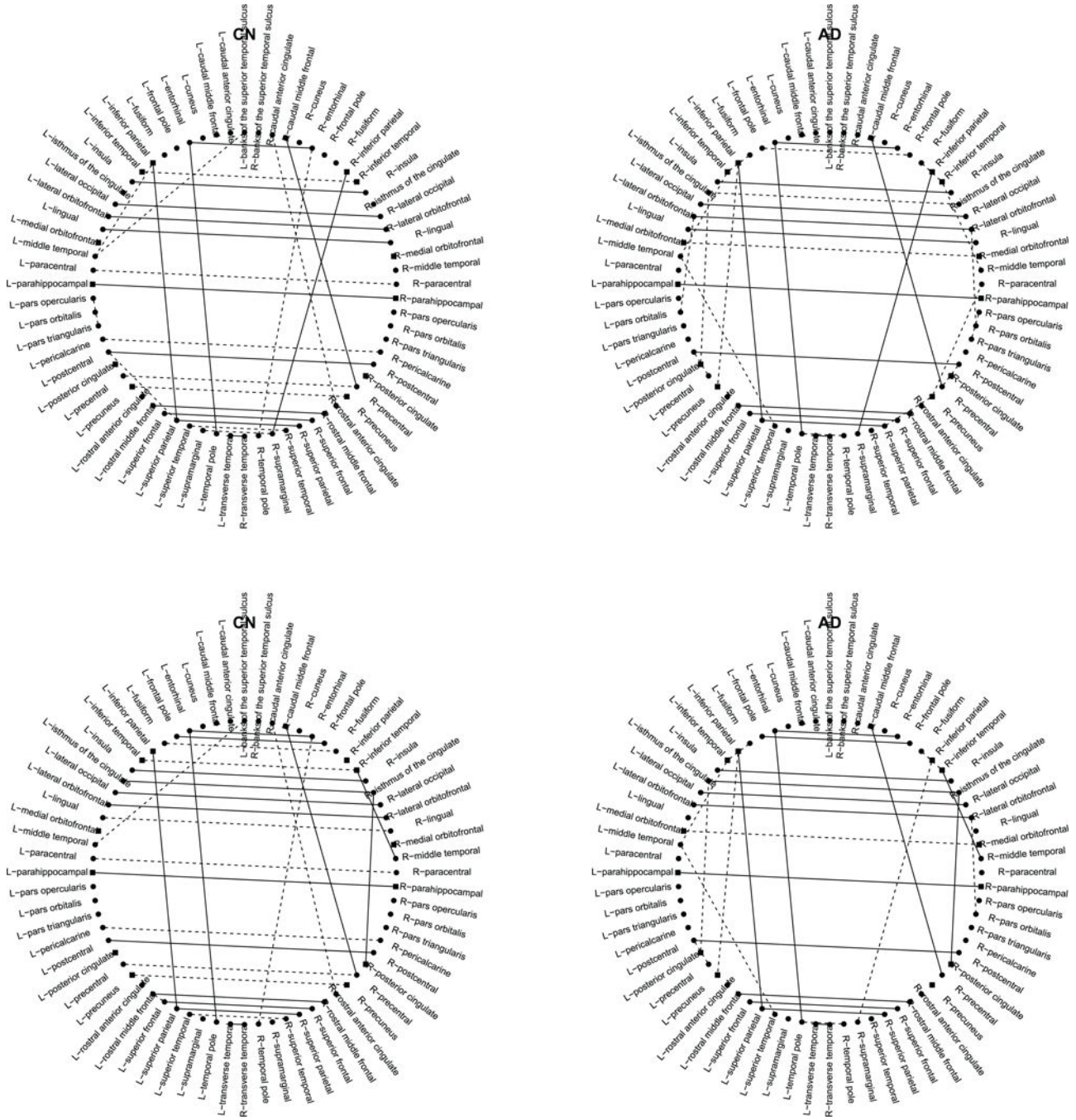

Figure 5.

Estimated networks by the proposed method (first row) and the method Janková and Van de Geer (2017) (second row) for the CN (left) and AD (right) groups, where reported edges are significant under a p-value of 0.05 after Bonferroni correction. Nodes with square shape belong to DMN. The solid edges denote those that are shared by the two groups, whereas the dashed edges denote those that are only present within one group.

Table 1.

Empirical size and power comparisons of the proposed CMLR test and test of Janková and Van de Geer (2017), denoted by *CMLR-chi-square* and *JG*, in the first testing problem for the Gaussian graphical model based on 1000 simulations.

		CMLR-chi-square		JG	
Graph	**(*n*, *p*)**	**Size**	**Power**	**Size**	**Power**
Band	(200,50)	0.054	(0.27, 0.78, 0.98, 1.0)	0.043	(0.24, 0.77, 0.99, 1.0)
	(200,100)	0.055	(0.30, 0.79, 0.98, 1.0)	0.042	(0.24, 0.75, 0.99, 1.0)
	(200,200)	0.048	(0.29, 0.80, 0.99, 1.0)	0.036	(0.23, 0.74, 0.98, 1.0)
Hub	(200,50)	0.019	(0.10, 0.36, 0.74, 0.95)	0.005	(0.06, 0.27, 0.66, 0.92)
	(200,100)	0.028	(0.12, 0.43, 0.81, 0.96)	0.005	(0.02, 0.17, 0.54, 0.86)
	(200,200)	0.031	(0.16, 0.55, 0.86, 0.98)	0.001	(0.02, 0.15, 0.50, 0.86)
Random	(200,50)	0.034	(0.15, 0.51, 0.86, 0.98)	0.025	(0.14, 0.49, 0.83, 0.98)
	(200,100)	0.041	(0.21, 0.68, 0.94, 1.0)	0.018	(0.11, 0.53, 0.92, 0.99)
	(200,200)	0.049	(0.15, 0.47, 0.81, 0.96)	0.034	(0.14, 0.41, 0.78, 0.95)

J Am Stat Assoc. Author manuscript; available in PMC 2020 August 11.

Table 2.

Empirical size and power comparisons in linear regression as well as estimated tuning parameter $\widehat{K}$ by a 5-fold cross-validation over 1000 simulations.

$\|B\|$	n	p	Method	Size	Power	$\widehat{K}$
1	100	50	**CMLR-chi-square**	0.057	(0.165, 0.489, 0.837, 0.972)	3.36 (1.08)
			CMLR-normal	0.061	(0.17, 0.495, 0.84, 0.972)	NA
			Zhang and Cheng	0.039	(0.109, 0.262, 0.579, 0.788)	NA
			DL	0.033	(0.132, 0.404, 0.724, 0.917)	NA
		200	**CMLR-chi-square**	0.055	(0.17, 0.524, 0.829, 0.974)	3.191 (0.591)
			CMLR-normal	0.058	(0.176, 0.532, 0.834, 0.975)	NA
			Zhang and Cheng	0.013	(0.042, 0.116, 0.306, 0.476)	NA
			DL	0.052	(0.144, 0.358, 0.694, 0.888)	NA
		500	**CMLR-chi-square**	0.051	(0.175, 0.509, 0.838, 0.963)	3.159 (0.583)
			CMLR-normal	0.051	(0.179, 0.513, 0.84, 0.963)	NA
			Zhang and Cheng	NA	NA	NA
			DL	NA	NA	NA
		1000	**CMLR-chi-square**	0.056	(0.165, 0.512, 0.828, 0.962)	3.115 (0.371)
			CMLR-normal	0.058	(0.17, 0.522, 0.83, 0.964)	NA
			Zhang and Cheng	NA	NA	NA
			DL	NA	NA	NA
5	100	50	**CMLR-chi-square**	0.058	(0.11, 0.328, 0.63, 0.865)	3.33 (0.94)
			CMLR-normal	0.052	(0.109, 0.322, 0.619, 0.862)	NA
			Zhang and Cheng	0.05	(0.063, 0.115, 0.226, 0.346)	NA
			DL	NA	NA	NA
		200	**CMLR-chi-square**	0.066	(0.114, 0.297, 0.601, 0.878)	3.188 (0.606)
			CMLR-normal	0.063	(0.112, 0.289, 0.592, 0.878)	NA
			Zhang and Cheng	0.037	(0.052, 0.111, 0.153, 0.253)	NA
			DL	NA	NA	NA
		500	**CMLR-chi-square**	0.064	(0.124, 0.321, 0.625, 0.895)	3.153 (0.56)

$\|B\|$	n	p	Method	Size	Power	$\widehat{K}$
			CMLR-normal	0.061	(0.118, 0.315, 0.618, 0.893)	NA
			Zhang and Cheng	NA	NA	NA
			DL	NA	NA	NA
		1000	**CMLR-chi-square**	0.059	(0.118, 0.304, 0.612, 0.872)	3.11 (0.355)
			CMLR-normal	0.057	(0.112, 0.3, 0.604, 0.869)	NA
			Zhang and Cheng	NA	NA	NA
			DL	NA	NA	NA
10	100	50	**CMLR-chi-square**	0.068	(0.094, 0.252, 0.528, 0.794)	3.41 (1.20)
			CMLR-normal	0.059	(0.085, 0.233, 0.503, 0.775)	NA
			Zhang and Cheng	0.054	(0.055, 0.085, 0.146, 0.21)	NA
			DL	NA	NA	NA
		200	**CMLR-chi-square**	0.086	(0.115, 0.253, 0.514, 0.786)	3.193 (0.618)
			CMLR-normal	0.079	(0.104, 0.238, 0.487, 0.767)	NA
			Zhang and Cheng	0.049	(0.055, 0.089, 0.106, 0.152)	NA
			DL	NA	NA	NA
		500	**CMLR-chi-square**	0.093	(0.123, 0.286, 0.54, 0.773)	3.159 (0.585)
			CMLR-normal	0.078	(0.113, 0.262, 0.516, 0.76)	NA
			Zhang and Cheng	NA	NA	NA
			DL	NA	NA	NA
		1000	**CMLR-chi-square**	0.073	(0.123, 0.252, 0.526, 0.779)	3.11 (0.355)
			CMLR-normal	0.066	(0.112, 0.23, 0.497, 0.766)	NA
			Zhang and Cheng	NA	NA	NA
			DL	NA	NA	NA

NOTES: Here “CMLR-chi-square,” “CMLR-normal,” “DL,” and “Zhang and Cheng” denote the proposed test based on a chi-square approximation, a normal approximation, the debias method of Zhang and Zhang (2014), and the method of Zhang and Cheng (2017). Note that the nominal size is 0.05, DL is a test converted from a CI, and NA means that a result is not applicable or the code fail to return a result after a code’s runtime exceeds one week.