On high-dimensional constrained maximum likelihood inference *

Yunzhang Zhu[1], Xiaotong Shen[2], and Wei Pan[3]

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 paper, 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.

Key words: Generalized Wilks phenomenon, Brain networks, High-dimensionality, L_0-regularization, (p, n)-asymptotics, similar tests.

*[1] Department of Statistics, The Ohio State University, Columbus, OH 43210; [2]School of Statistics, [3] Division of Biostatistics, University of Minnesota, Minneapolis, MN 55455. The authors thank the editors, the associate editor and anonymous referees for helpful comments and suggestions. Research supported in part by NSF grants DMS-1415500, DMS-1712564, DMS-1721216, DMS-1712580, DMS-1721445, and DMS-1721445, NIH grants 1R01GM081535-01, 1R01GM126002, HL65462 and R01HL105397. The authors would like to thank the editor, the associate editor and three anonymous referees for helpful comments and suggestions.

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 paper, 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 [23, 7, 27, 21], among others. For the Gaussian graphical model, methods include the regularized likelihood approach [19, 8, 26, 6, 21] and the nodewise regression approach [16], and their extensions such as conditional Gaussian graphical [13, 25] and multiple Gaussian graphical models [31, 14]. Despite progress, there is a paucity of inferential methods for high-dimensional models, although some have been recently proposed in [28, 24, 12, 11], where confidence intervals are constructed based on a bias-reduction method called "debias" [28]. 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 confidence interval.

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 unhypothesized 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 [22], which has been used in [18] 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 [28, 24, 12, 11]. 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, c.f., 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 [21]. For each convex relaxation, we employ the alternating direction method of multipliers algorithm [3], 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 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 Alzheimer's disease subjects versus healthy subjects in Section 6. Section 7 is devoted to technical proofs.

2 Constrained likelihood ratios

Given an iid sample $X_1, \ldots, X_n$ 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 $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_d) \in \mathbb{R}^d$, and $B \subseteq \{1, \ldots, d\}$. 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 $\boldsymbol{\theta}_{B^c}$. Specifically, define

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

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

where $L_n(\boldsymbol{\theta}) = \sum_{i=1}^n \log p_{\boldsymbol{\theta}}(X_i)$ is the log-likelihood, $p_\tau(x) = \min(x/\tau, 1)$ is the truncated L_1-function [21] as a surrogate of the L_0-function, and (K, τ) 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: $\Lambda_n(B) = 2\left(L_n\big(\widehat{\boldsymbol{\theta}}^{(1)}\big) - L_n\big(\widehat{\boldsymbol{\theta}}^{(0)}\big)\right)$. 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 τ in (1) and (2) are estimated using a cross-validation (CV) criterion based on the full model (1). Choosing the same values of (K, τ) 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 $\boldsymbol{X}_1 \cdots, \boldsymbol{X}_n$ 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 $\|\boldsymbol{a}\|_2 = \sqrt{a_1^2 + \ldots + a_m^2}$. Denote by $\bar{\boldsymbol{\Omega}}^0_{A \cup B} = \arg\min_{\boldsymbol{\Omega} \succ 0 : \boldsymbol{\Omega}_{(A \cup B)^c} = \mathbf{0}} K(\boldsymbol{\Omega}^0, \boldsymbol{\Omega})$ an approximating point in a space $\{\boldsymbol{\Omega} : \boldsymbol{\Omega} \succ 0, \boldsymbol{\Omega}_{(A \cup B)^c} = \mathbf{0}\}$ to the true $\boldsymbol{\Omega}^0$, where $K(\boldsymbol{\Omega}^0, \boldsymbol{\Omega}) = \frac{1}{2}\big(\operatorname{tr}(\boldsymbol{\Omega}\boldsymbol{\Sigma}^0) + \log \frac{\det(\boldsymbol{\Omega}^0)}{\det(\boldsymbol{\Omega})} - p\big)$ is the Kullback-Leibler information. Let $\|\boldsymbol{\Omega}^0 - \boldsymbol{\Omega}\| = \|\sqrt{\boldsymbol{\Sigma}^0}(\boldsymbol{\Omega} - \boldsymbol{\Omega}^0)\sqrt{\boldsymbol{\Sigma}^0}\|_F$ be the Fisher-norm between $\boldsymbol{\Omega}^0$ and $\boldsymbol{\Omega}$ [20]. Moreover, let $A^0 = \{i : \theta_i^0 \neq 0\}$ be the support of true parameter $\boldsymbol{\theta}^0$, $\kappa_0 = \lambda_{\max}(\boldsymbol{\Omega}^0)/\lambda_{\min}(\boldsymbol{\Omega}^0)$ be the condition number of $\boldsymbol{\Omega}^0$, and $\kappa_1 = \frac{\bar{\lambda}^2_{\max}}{\lambda^2_{\min}(\boldsymbol{\Omega}^0)}$, where $\bar{\lambda}_{\max} = \max_{A : |A| \leq |A^0|, A \cap B = \emptyset} \lambda_{\max}(\bar{\boldsymbol{\Omega}}^0_{A \cup B})$. Let $\bar{\lambda}_{\min} = \min_{A : |A| \leq |A^0|, A \cap B = \emptyset} \lambda_{\min}(\bar{\boldsymbol{\Omega}}^0_{A \cup B})$. Let $\gamma_{\min} = \min_{(i,j) \in A^0} |\omega^0_{ij}|$ be the minimum nonzero offdiagonals of $\boldsymbol{\Omega}^0$, representing the signal strength. The following technical conditions are made.

Assumption 1 (Degree of separation)

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

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

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

for some constant $C_2 > 0$.

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

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

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 [18] 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 $\tau \leq \frac{\bar{\lambda}_{min}\min(\sqrt{C_{min}}, C^2_{min})}{12|A^0|}$ such that under H_0*

(i) Wilks phenomenon: If $\omega^0_{ij} = 0$ 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 $\omega^0_{ij} = 0$ 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 can not be over-selection consistency when Assumption 1 breaks down in view of the result of [21]. 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 $\boldsymbol{\Omega}$ 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 [11]. 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

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

where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ and $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{ip})^T$ 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 $L_n(\boldsymbol{\theta}) = L_n(\boldsymbol{\beta}, \sigma) = -\frac{1}{2\sigma^2}\|Y - X\boldsymbol{\beta}\|_2^2 - n\log(\sqrt{2\pi}\sigma)$, and the constrained log-likelihood ratio is accordingly defined as $\Lambda_n(B) = 2(L_n(\widehat{\boldsymbol{\beta}}^{(1)}, \hat{\sigma}^{(1)}) - L_n(\widehat{\boldsymbol{\beta}}^{(0)}, \hat{\sigma}^{(0)}))$, where $\widehat{\boldsymbol{\beta}}^{(0)}$ and $\widehat{\boldsymbol{\beta}}^{(1)}$ 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 [22])

$$\min_{A:|A|\leq|A^0| \text{ and } A\neq A^0} \inf_{\boldsymbol{\beta}} \frac{\|X\boldsymbol{\beta}^0 - X_{A\cup B}\boldsymbol{\beta}_{A\cup B}\|_2^2}{n|A^0 \setminus A|} \geq C_0\sigma^2\frac{\log p}{n} \tag{6}$$

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 $0 < \tau \leq \sigma\sqrt{\frac{6}{(n+2)p\lambda_{max}(X^\top X)}}$ 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 $\frac{\sqrt{|B|}(|A^0|+|B|)}{n} \to 0$ in linear regression appears weaker than that $\frac{(|A^0|+|B|)\log p}{n^{1/2}} \to 0$ 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 $\boldsymbol{x} = (x_1, \ldots, x_p)$ are independently distributed from $N(\mu_i, 1)$ with $\mu_1 = 0$ and $\mu_j = 1$; $2 \leq j \leq p$, and ϵ is $N(0, 1 - n^{-1})$, independent of $\boldsymbol{x}$. Assume that $\beta_0 = 0$ and $\boldsymbol{\beta} = (n^{-1/2}, 0, \ldots, 0)$, 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 [11] in Gaussian graphical model. To that end, we first establish the asymptotic distribution of $\widehat{\boldsymbol{\theta}}_B$ under the null H_0 for fixed index set B for the Gaussian graphical 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 $\mathrm{vec}_B(\boldsymbol{C}) = \left(\sqrt{1 + \mathbb{I}(i \neq j)} c_{ij}\right)_{(i,j):(i,j) \text{ or } (j,i) \in B}$ is a sub-vector of $\mathrm{vec}(\boldsymbol{C})$ excluding components with indices not in B, $\mathrm{vec}(\boldsymbol{C}) = \left(\sqrt{1 + \mathbb{I}(i \neq j)} c_{ij}\right)_{i \leq j} \in \mathbb{R}^{\frac{p(p+1)}{2}}$ is a scaled vectorization of a $p \times p$ symmetric matrix $\boldsymbol{C}$ [1] and $\mathbb{I}(\cdot)$ is the indicator. For the Fisher information, we need the symmetric Kronecker product [1] 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 $\boldsymbol{C}$ $\boldsymbol{C} \otimes_s \boldsymbol{C} \in \mathbb{R}^{\frac{p(p+1)}{2} \times \frac{p(p+1)}{2}}$ as $(\boldsymbol{C} \otimes_s \boldsymbol{C}) \, \mathrm{vec}(\boldsymbol{\Delta}) = \mathrm{vec}\left(\boldsymbol{C} \boldsymbol{\Delta} \boldsymbol{C}\right)$ for any symmetric matrix $\boldsymbol{\Delta}$, and define the Fisher information matrix for the $\frac{p(p+1)}{2}$-dimensional vector $\mathrm{vec}(\boldsymbol{\Omega})$ as $\boldsymbol{I} = \nabla^2 \left(-\frac{1}{2} \log \det \boldsymbol{\Omega}^0\right) = \frac{1}{2} \boldsymbol{\Sigma}^0 \otimes_s \boldsymbol{\Sigma}^0$, 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 below gives the asymptotic distribution of $\mathrm{vec}_B(\widehat{\boldsymbol{\Omega}}^{(1)})$.

Proposition 1 *(Asymptotic distribution of CMLE $\widehat{\boldsymbol{\Omega}}^{(1)}$ for Gaussian graphical model) Under Assumptions 1-2, if $|B|$ is fixed, there exists a pair of tuning parameters (K, τ) with $K = |A^0|$ and $\tau \le \frac{\bar{\lambda}_{min} \min(\sqrt{C_{min}}, C_{min}^2)}{12|A^0|}$ such that $\widehat{\boldsymbol{\Omega}}^{(1)}$ satisfies:*

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

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 $\tau \le \sigma\sqrt{\frac{6}{(n+2)p\lambda_{max}(X^\top X)}}$ such that $\widehat{\boldsymbol{\theta}}^{(1)}_B$ satisfies:*

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

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 $\boldsymbol{\theta}^n = (\theta^n_1, \cdots, \theta^n_d)^T$. Subsequently, we study the behavior of the *local limiting power function* for the proposed CMLR test $\pi_{LR}(h, \theta_{B^c}) = \liminf_{n\to\infty} P_{H_a}(\Lambda_n(B) \ge \chi^2_{\alpha,|B|})$ if $|B|$ is fixed and $\liminf_{n\to\infty} P_{H_a}((2|B|)^{-1/2}\Lambda_n(B) - |B|) \ge z_\alpha)$ if $|B| \to \infty$. Let the corresponding $\pi_{\mathrm{debias}}(h, \theta_{B^c})$ of the debias test in [11] in the Gaussian graphical model as a result for linear regression is similar.

Theorem 3 *If for any $\theta^n = \Omega^n$ the Assumptions 1-2 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},*

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

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, $\lim_{h\to\infty} \pi_{LR}(h, \Omega_{B^c}) = 1$. Moreover, in the one-dimensional situation with $|B| = 1$, for any h and Ω_{B^c},

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

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 [11] for the one-dimensional case with $|B| = 1$. As indicated by Corollary 1 below, 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 $\left[\sqrt{n}(\widehat{\omega}_{ij} - \omega^0_{ij})\right]_{(i,j)\in B}$ 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 $\sqrt{n}(\widehat{\omega}_{ij} - \omega^0_{ij})$ is no larger than $\left[\omega^0_{ij}\right]^2 + \omega^0_{ii}\omega^0_{jj}$, the asymptotic variance of the regression estimator in [11].*

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

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

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

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}, $\lim_{h\to\infty} \pi_{LR}(h, \beta_{B^c}) = 1$.

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 [21] 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 [21] for convergence of the method.

For (1) and (2), at the m-th iteration, we solve

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

to yield $\widehat{\boldsymbol{\theta}}^{[m+1]}$, where $A_1 = B$ and $A_2 = \emptyset$ 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 [3], which amounts to the following iterative updating scheme

$$\boldsymbol{\theta}^{[k+1]} = \arg\min_{\boldsymbol{\theta}} \left(-L_n(\boldsymbol{\theta}) + (\rho/2) \cdot \left\|\boldsymbol{\theta} - \boldsymbol{\delta}^{[k]} + \boldsymbol{\gamma}^{[k]}\right\|_2^2\right), \tag{12}$$

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

where $\mathcal{F}^{[m]} = \left\{ \sum_{i\notin A_1} |\theta_i| \mathbb{I}(|\theta_i^{[m]}| \leq \tau) \leq \tau\big(K - \sum_{(i,j)\notin A_1} \mathbb{I}(|\theta_i^{[m]}| > \tau)\big), \boldsymbol{\theta}_{A_2} = 0 \right\}$,

$\mathcal{P}_{\mathcal{F}^{[m]}}(\cdot)$ denotes the projection onto the set $\mathcal{F}^{[m]}$, and $\rho > 0$ is fixed or can be adaptively updated using a strategy in [30]. 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 (cf. Section 6.5 of [3]) and solving a linear system for the linear model, while (13) is performed using the L_1-projection algorithm of [15] 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:

$$\min_{x \in \mathbb{R}^d} \tfrac{1}{2}\|x - y\|_2^2 \text{ subj to } \textstyle\sum_{i \notin A} c_i |x_i| \leq z \text{ and } x_i = 0, i \in A\,,$$

where $c_i \geq 0; i = 1, \cdots, d$ and A is a subset of $\{1, \cdots, d\}$. The solution of this problem is $x_i^\star = 0$ if $i \in A$; $x_i^\star = y_i$ if $\sum_{i \notin A} c_i |y_i| \leq z$; $x_i^\star = \operatorname{sgn}(y_i) \max(|y_i| - c_i \lambda^\star, 0)$ 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 $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, $\boldsymbol{\Omega}$ 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 $\omega_{ij} = .3$ if i and j are connected and diagonals equal to $.3 + c$ with c chosen so that the smallest eigenvalue of the resulting matrix equals to .2. Finally, a random sample of size $n = 200$ is drawn from $N(\mathbf{0}, \boldsymbol{\Omega}^{-1})$.

In what follows, we consider two hypothesis testing problems concerning conditional independence of components of a Gaussian random vector $\boldsymbol{X} = (X_1, \cdots, X_p)$. 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 $H_0 : \omega_{i_0 j} = 0$; $1 \leq j \neq i_0 \leq p$ 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 [11] 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 $\omega_{i_0 j_0}^{(l)} = \frac{\omega_{i_0 j_0} l}{4}$, $l = 1, \cdots, 4$. 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 $\tau = .001$ 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 five-fold CV. Specifically, we divide the dataset into five roughly equal parts denoted by $\mathcal{D}_1, \cdots, \mathcal{D}_5$. Define $\widehat{\boldsymbol{\Sigma}}_l$ and $\widehat{\boldsymbol{\Sigma}}_{-l}$ respectively as the sample covariance matrices calculated based on samples in $\mathcal{D}_l$ and $\{\mathcal{D}_1, \cdots, \mathcal{D}_5\} \setminus \mathcal{D}_l$; $l = 1, \cdots, 5$. Similarly, define $\widehat{\boldsymbol{\Omega}}_{-l}(K)$ to be the precision matrix calculated based on sample covariance matrix $\widehat{\boldsymbol{\Sigma}}_{-l}$; $l = 1, \cdots, 5$. The five-fold CV criterion is $\mathrm{CV}(K) = 5^{-1} \sum_{l=1}^{5} \Big(-\log\det\big(\widehat{\boldsymbol{\Omega}}_{-l}(K)\big) + \mathrm{tr}\big[\widehat{\boldsymbol{\Sigma}}_l \widehat{\boldsymbol{\Omega}}_{-l}(K)\big] - p \Big)$. Then the optimal tuning parameter is obtained by minimizing $\mathrm{CV}(K)$ 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 [11], denoted as **CMLR-chi-square** and **JG**, where the confidence interval in [11] 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 [28, 24] and the method of [29]. 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 $\beta^0 = (1, 2, 3, \beta^0_B, \mathbf{0})$ and $\|\boldsymbol{\beta}_B\|_2 = l/10$; $l = 0, 1, \ldots, 4$. 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 $|B| = 1, 5, 10$. 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 [29] 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 7, 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 $\hat{K}$ 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 [11, 28, 24, 29]. Moreover, the asymptotic approximation seems adequate in all the examples.

Figures 1-7 and Tables 1-2 about here

6 Brain network analysis

Alzheimer's disease (AD) 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 [2]. 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 [10]. 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 [5]. Since previous studies (e.g., [9, 17]) have identified the default mode network (DMN) to be associated with AD, we will pay particular attention to this subnetwork, which includes 12 ROIs in our dataset. As in [10], 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 [10, 17] 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 [11] or **JG**. As indicated in Figure 7, the CMLR test yields 28 and 33 significant 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 [17], 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 [17]. 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.

7 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 $\widehat{\boldsymbol{\Omega}}_A = \arg\max_{\boldsymbol{\Omega}\succ 0,\, \boldsymbol{\Omega}_{A^c}=0} L_n(\boldsymbol{\Omega})$, 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 $A^0 = \{(i,j) : i \leq j, \omega^0_{ij} \neq 0\}$ is the index set including all the indices corresponding to nonzero entries of the true precision matrix $\boldsymbol{\Omega}^0 = (\omega^0_{ij})_{p\times p}$

Lemma 2 *For any symmetric matrices $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$, $\operatorname{vec}(\boldsymbol{C}_1)^\top \operatorname{vec}(\boldsymbol{C}_2) = \operatorname{tr}(\boldsymbol{C}_1\boldsymbol{C}_2)$. Moreover, for any positive definite matrix $\boldsymbol{C} \succ 0$,*

$$\nabla\left(\log\det \boldsymbol{C}\right) = -\operatorname{vec}(\boldsymbol{C}^{-1}), \quad \nabla^2\left(-\log\det \boldsymbol{\Omega}^0\right) = \boldsymbol{C}^{-1} \otimes_s \boldsymbol{C}^{-1}, \tag{14}$$

$$\boldsymbol{I} = \tfrac{1}{2}\boldsymbol{\Sigma}^0 \otimes_s \boldsymbol{\Sigma}^0, \tag{15}$$

$$\operatorname{Var}\left(\operatorname{vec}(\boldsymbol{X}\boldsymbol{X}^\top)\right) = 4\boldsymbol{I} \text{ with } \boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}^0), \tag{16}$$

$$\operatorname{vec}(\boldsymbol{C})^\top \boldsymbol{I} \operatorname{vec}(\boldsymbol{C}) = \tfrac{1}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}^0\boldsymbol{C}\boldsymbol{\Sigma}^0\boldsymbol{C}\right). \tag{17}$$

Lemma 3 *For any symmetric matrix $\boldsymbol{T}$ and $\nu > 0$*

$$\mathbb{P}\left(|\operatorname{tr}((\boldsymbol{S}-\boldsymbol{\Sigma}^0)\boldsymbol{T})| \geq \nu\right) \leq 2\exp\left(-n\tfrac{\nu^2}{9\|\boldsymbol{T}\|^2 + 8\nu\|\boldsymbol{T}\|}\right), \tag{18}$$

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

$$\mathbb{P}\left(\max_{1\leq k\leq K}\left|\operatorname{tr}((\boldsymbol{S}-\boldsymbol{\Sigma}^0)\boldsymbol{T}_k)\right| \geq \nu\right) \leq 2\exp\left(-n\frac{\nu^2}{9c_0^2 + 8c_0\nu} + \log K\right), \tag{19}$$

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

$$\mathbb{P}\left(\|\operatorname{vec}_B(\boldsymbol{S}-\boldsymbol{\Sigma}^0)\|_\infty \geq \nu\right) \leq 2\exp\left(-n\frac{\nu^2}{9\lambda^2_{max}(\boldsymbol{\Sigma}^0) + 8\nu\lambda_{max}(\boldsymbol{\Sigma}^0)} + \log|B|\right), \tag{20}$$

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

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:*

$$K(\boldsymbol{\Omega}^0, \boldsymbol{\Omega}) \;\geq\; \min\left(\frac{1}{16\sqrt{2}}, \frac{\sqrt{K(\boldsymbol{\Omega}^0, \boldsymbol{\Omega})}}{2\sqrt{6}}\right) \|\boldsymbol{\Omega}^0 - \boldsymbol{\Omega}\|, \tag{21}$$

$$K(\boldsymbol{\Omega}^0, \boldsymbol{\Omega}) \;\geq\; \min\left(\frac{1}{16\sqrt{2}}, \frac{\|\boldsymbol{\Omega}^0 - \boldsymbol{\Omega}\|}{24}\right) \|\boldsymbol{\Omega}^0 - \boldsymbol{\Omega}\|\,. \tag{22}$$

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*

$$\|\widehat{\boldsymbol{\Omega}}_{\tilde{A}} - \boldsymbol{\Omega}^0\| \leq 12\|\boldsymbol{I}_{\tilde{A},\tilde{A}}^{-1/2}\operatorname{vec}(\boldsymbol{\Sigma}^0 - \boldsymbol{S})\|_2\,. \tag{23}$$

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

$$\|\widehat{\boldsymbol{\Omega}}_{\tilde{A}} - \boldsymbol{\Omega}^0\| = O_p\left(\sqrt{\frac{|\tilde{A}|\log p}{n}}\right)\,. \tag{24}$$

Lemma 6 *(Selection consistency) If* $K = |A^0|$*,* $\tau \leq \frac{\bar{\lambda}_{min}\min(\sqrt{C_{min}}, C_{min}^2)}{12|A^0|}$*, then*

$$\begin{aligned} &\max\left(P\left(\widehat{\boldsymbol{\Omega}}^{(0)} \neq \widehat{\boldsymbol{\Omega}}_{A^0}\right), P\left(\widehat{\boldsymbol{\Omega}}^{(1)} \neq \widehat{\boldsymbol{\Omega}}_{A^0 \cup B}\right)\right) \\ \leq\;& 2\exp\left(\frac{-nC_{min}}{2560 \times 512} + 2\log p\right) + \exp\left(\frac{-n}{2560} + |A^0|\log p\right) \\ &+2\exp\left(-n\frac{\min\left(\sqrt{\frac{\min(C_{min}/512\,, 3/32)}{48\lambda_{max}^2(|A^0|+|B|)}}, \lambda_{max}(\boldsymbol{\Sigma}^0)\right)^2}{18\lambda_{max}^2(\boldsymbol{\Sigma}^0)} + 2\log p\right) \longrightarrow 0 \end{aligned} \tag{25}$$

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

Lemma 7 *Let* $\boldsymbol{\gamma}_k = (\gamma_{k1}, \cdots \gamma_{km}) \in \mathbb{R}^m$*;* $k = 1, \cdots, n$ *be iid random vectors with* $\operatorname{Var}(\boldsymbol{\gamma}_1) = \boldsymbol{I}_{m\times m}$*. If* m *is fixed, then*

$$n^{-1}\|\sum_{k=1}^n \boldsymbol{\gamma}_k\|_2^2 \xrightarrow{d} \chi_m^2, \text{ as } n \to \infty\,. \tag{26}$$

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

$$\frac{\|\sum_{k=1}^n \boldsymbol{\gamma}_k\|_2^2 - nm}{n\sqrt{2m}} \xrightarrow{d} N(0,1), \text{ as } n \to \infty\,. \tag{27}$$

Lemma 8 *Let* $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}^0)$ *and* $\gamma = \operatorname{tr}(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{\Sigma}^0)\boldsymbol{T})$ *with* $\boldsymbol{T}$ *a symmetric matrix. Then*

$$\mathbb{E}(\gamma^{2m}) \leq (2m-1)!\,2^{m-1}\left(\mathbb{E}(\gamma^2)\right)^m \text{ for any integer } m \geq 1\,. \tag{28}$$

Lemma 9 *(Asymptotic distribution for log-likelihood ratios) The log-likelihood ratio statistic* $Lr = 2(L_n(\widehat{\boldsymbol{\Omega}}_{\tilde{A}}) - L_n(\widehat{\boldsymbol{\Omega}}_{A^0}))$*, where* $\widehat{\boldsymbol{\Omega}}_{\tilde{A}}$ *is the MLE over index set* $\tilde{A}$ *with* $\tilde{A} \supseteq A^0$*. Denote by* κ_0 *the condition number of* $\boldsymbol{\Sigma}^0$*. If* $\frac{\kappa_0|\tilde{A}|\log p}{\sqrt{n}} \to 0$ *with* $p \geq 2$*, then,*

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

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

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

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

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

where $R(\widehat{\mathbf{\Delta}}_{\tilde{A}}) = \mathbf{\Sigma}^0 \sum_{i=2}^{\infty}(-1)^i\left(\widehat{\mathbf{\Delta}}_A \mathbf{\Sigma}^0\right)^i$. Let $\boldsymbol{J} = \boldsymbol{I}^{-1}_{\tilde{A},\tilde{A}}$ 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

$$\sqrt{n}\boldsymbol{J}^{-1/2}_{B,B}\,\text{vec}_B\left(\widehat{\mathbf{\Omega}}_{A^0 \cup B} - \mathbf{\Omega}^0\right) = \frac{\sqrt{n}}{2}\boldsymbol{J}^{-1/2}_{B,B}\boldsymbol{J}_{B,\tilde{A}}\left(\text{vec}_{\tilde{A}}\left(\mathbf{\Lambda}\right) - \text{vec}_{\tilde{A}}\left(R(\widehat{\mathbf{\Delta}}_{\tilde{A}})\right)\right). \tag{29}$$

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 $\left\|\boldsymbol{J}^{-1/2}_{B,B}\boldsymbol{J}_{B,\tilde{A}}\boldsymbol{x}\right\|_2^2 = \boldsymbol{x}^\top\boldsymbol{J}\boldsymbol{x} - \boldsymbol{x}^\top\boldsymbol{I}^{-1}_{A^0,A^0}\boldsymbol{x} \le \boldsymbol{x}^\top\boldsymbol{J}\boldsymbol{x} \le \lambda^{-2}_{\min}(\mathbf{\Sigma}^0)\|\boldsymbol{x}\|_2^2$ for any $\boldsymbol{x} \in \mathbb{R}^{|A|}$. This, together with (B.37) of the online supplementary material, implies that

$$\begin{aligned}
&\left\|.5\sqrt{n}\boldsymbol{J}^{-1/2}_{B,B}\boldsymbol{J}_{B,\tilde{A}}\,\text{vec}_{\tilde{A}}\left(R(\widehat{\mathbf{\Delta}}_A)\right)\right\|_2 \le .5\sqrt{n}\left\|\boldsymbol{J}^{1/2}\,\text{vec}_{\tilde{A}}\left(R(\widehat{\mathbf{\Delta}}_{\tilde{A}})\right)\right\|_2 \\
\le\ & .5\sqrt{n}\lambda^{-1}_{\min}(\mathbf{\Sigma}^0)\left\|R(\widehat{\mathbf{\Delta}}_{\tilde{A}})\right\|_2 \le \sqrt{n}\kappa_0\|\mathbf{\Sigma}^0\widehat{\mathbf{\Delta}}_{\tilde{A}}\|_F^2 = O_p\left(\frac{\kappa_0|\tilde{A}|\log p}{\sqrt{n}}\right) = o_p(1)
\end{aligned} \tag{30}$$

under Assumption 2. For the first term, note that

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

where the second last equality uses the property of exponential family [4]. Hence, by the central limit theorem, $\mathrm{vec}_{\tilde{A}}(\mathbf{\Lambda}) \xrightarrow{d} N\left(0, [\boldsymbol{I}^{-1}_{\tilde{A},\tilde{A}}]_{B,B}\right)$. Finally, by Slutsky's Theorem, we obtain that $\sqrt{n}\,\mathrm{vec}_B\left(\widehat{\mathbf{\Omega}}_{A^0\cup B} - \mathbf{\Omega}^0\right) \xrightarrow{d} N\left(0, [\boldsymbol{I}^{-1}_{\tilde{A},\tilde{A}}]_{B,B}\right)$. This completes the proof.

Proof of Proposition 2: By Theorem 3 of [22], $\mathbb{P}\left(\{\widehat{\boldsymbol{\beta}}^{(1)} = \widehat{\boldsymbol{\beta}}^{ls}_{A^0\cup B}\}\right) \to 1$, as $n, p \to \infty$. Hence, with probability tending to 1,

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

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

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

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

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

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

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

Proof of Theorem 2: By Theorem 3 of [22], $\mathbb{P}\left(\{\widehat{\boldsymbol{\beta}}^{(1)} = \widehat{\boldsymbol{\beta}}^{ls}_{A^0\cup B}\} \cap \{\widehat{\boldsymbol{\beta}}^{(0)} = \widehat{\boldsymbol{\beta}}^{ls}_{A^0}\}\right) \to 1$, as $n, p \to \infty$, by Assumption 1, where $\widehat{\boldsymbol{\beta}}^{ls}_A$ is the least square estimate over A. Hence, in what follows, we focus our attention to event $\{\widehat{\boldsymbol{\beta}}^{(1)} = \widehat{\boldsymbol{\beta}}^{ls}_{A^0\cup B}\} \cap \{\widehat{\boldsymbol{\beta}}^{(0)} = \widehat{\boldsymbol{\beta}}^{ls}_{A^0}\}$.

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

$$n\left(\log(\|y - X\boldsymbol{\beta}\|_2^2) - \log(\|y - X\boldsymbol{\beta}^0\|_2^2)\right) = -n\sum_{i=1}^{\infty}\frac{(2\boldsymbol{\epsilon}^\top X\boldsymbol{\delta} - \|X\boldsymbol{\delta}\|_2^2)^i}{i\|\boldsymbol{\epsilon}\|_2^{2i}} \tag{31}$$

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

$$\widehat{\boldsymbol{\beta}}^{(1)} = \boldsymbol{\beta}^0 + (X_{A^0\cup B}^\top X_{A^0\cup B})^{-1}X_{A^0\cup B}^\top\boldsymbol{\epsilon} \text{ and } \widehat{\boldsymbol{\beta}}^{(0)} = \boldsymbol{\beta}^0 + (X_{A^0}^\top X_{A^0})^{-1}X_{A^0}^\top\boldsymbol{\epsilon}, \tag{32}$$

implying that $X(\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^0) = \boldsymbol{P}_{A^0\cup B}\boldsymbol{\epsilon}$ and $X(\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^0) = \boldsymbol{P}_{A^0}\boldsymbol{\epsilon}$. Consequently, replacing $\boldsymbol{\delta} = \widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^0$, the right-hand of (31) reduces to

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

Similarly, replacing $\boldsymbol{\delta}$ by $\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^0$, (31) 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 $\{\boldsymbol{\epsilon}^\top\boldsymbol{P}_{A^0\cup B}\boldsymbol{\epsilon} < \|\boldsymbol{\epsilon}\|_2^2\}$. This, together with the facts that $n/\|\boldsymbol{\epsilon}\|_2^2 \xrightarrow{\mathbb{P}} 1$ and

$|A^0|/n \to 0$, 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 $\frac{\sqrt{|B|}(|A^0|+|B|)}{n} \to 0$, because

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

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

References

[1] F. Alizadeh, J. A. Haeberly, and M. L. Overton. Primal-dual interior-point methods for semidefinite programming: convergence rates, stability and numerical results. *SIAM Journal on Optimization*, 8(3):746–768, 1998.

[2] Alzheimer's Association et al. Changing the trajectory of alzheimer's disease: how a treatment by 2025 saves lives and dollars, 2016.

[3] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn*, 3(1):1–122, 2011.

[4] L. D. Brown. Fundamentals of statistical exponential families with applications in statistical decision theory. *Lecture Notes-monograph series*, pages 1–279, 1986.

[5] R. S. Desikan, F. Ségonne, B. Fischl, B. T. Quinn, B. C. Dickerson, D. Blacker, R. L. Buckner, A. M. Dale, R. P. Maguire, B. T. Hyman, et al. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980, 2006.

[6] J. Fan, Y. Feng, and Y. Wu. Network exploration via the adaptive lasso and scad penalties. *The annals of applied statistics*, 3(2):521–541, 2009.

[7] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[8] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[9] M. D. Greicius, G. Srivastava, A. L. Reiss, and V. Menon. 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(13):4637–4642, 2004.

[10] Y. He, Z. Chen, and A. Evans. Structural insights into aberrant topological patterns of large-scale cortical networks in alzheimer's disease. *The Journal of neuroscience*, 28(18):4756–4766, 2008.

[11] J. Janková and S. Van de Geer. Honest confidence regions and optimality in high-dimensional precision matrix estimation. *TEST*, pages 1–20, 2016.

[12] A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

[13] B. Li, H. Chun, and H. Zhao. Sparse estimation of conditional graphical models with application to gene networks. *Journal of the American Statistical Association*, 107:152–167, 2012.

[14] Z. Lin, T. Wang, C. Yang, and H. Zhao. On joint estimation of gaussian graphical models for spatial and temporal data. *Biometrics*, 2017.

[15] J. Liu and J. Ye. Efficient euclidean projections in linear time. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 657–664. ACM, 2009.

[16] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, pages 1436–1462, 2006.

[17] M.E. Montembeault, I. Rouleau, J. S. Provost, and S. M. Brambati. Altered gray matter structural covariance networks in early stages of alzheimer's disease. *Cerebral Cortex*, page bhv105, 2015.

[18] S. Portnoy et al. Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *The Annals of Statistics*, 16(1):356–366, 1988.

[19] A.J. Rothman, P.J. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

[20] X. Shen. On methods of sieves and penalization. *The Annals of Statistics*, pages 2555–2591, 1997.

[21] X. Shen, W. Pan, and Y. Zhu. Likelihood-based selection and sharp parameter estimation. *Journal of American Statistical Association*, 107:223–232, 2012.

[22] X. Shen, W. Pan, Y. Zhu, and H. Zhou. On constrained and regularized high-dimensional regression. *Annals of the Institute of Statistical Mathematics*, 65(5):807–832, 2013.

[23] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[24] S. Van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

[25] J. Yin and H. Li. Adjusting for high-dimensional covariates in sparse precision matrix estimation by 1-penalization. *Journal of multivariate analysis*, 116:365–381, 2013.

[26] M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

[27] C.H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010.

[28] C.H. Zhang and S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

[29] X. Zhang and G. Cheng. Simultaneous inference for high-dimensional linear models. *Journal of the American Statistical Association*, 112(518):757–768, 2017.

[30] Y. Zhu. An augmented admm algorithm with application to the generalized lasso problem. *Journal of Computational and Graphical Statistics*, (1), 2017.

[31] Y. Zhu, X. Shen, and W. Pan. Structural pursuit over multiple undirected graphs. *Journal of the American Statistical Association*, 109(508):1683–1696, 2014.

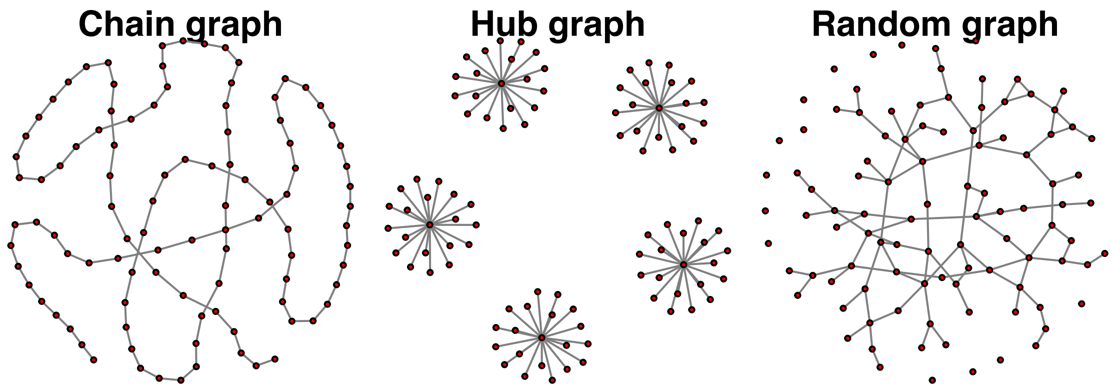

Figure 1: Three types of graphs used in our simulations.

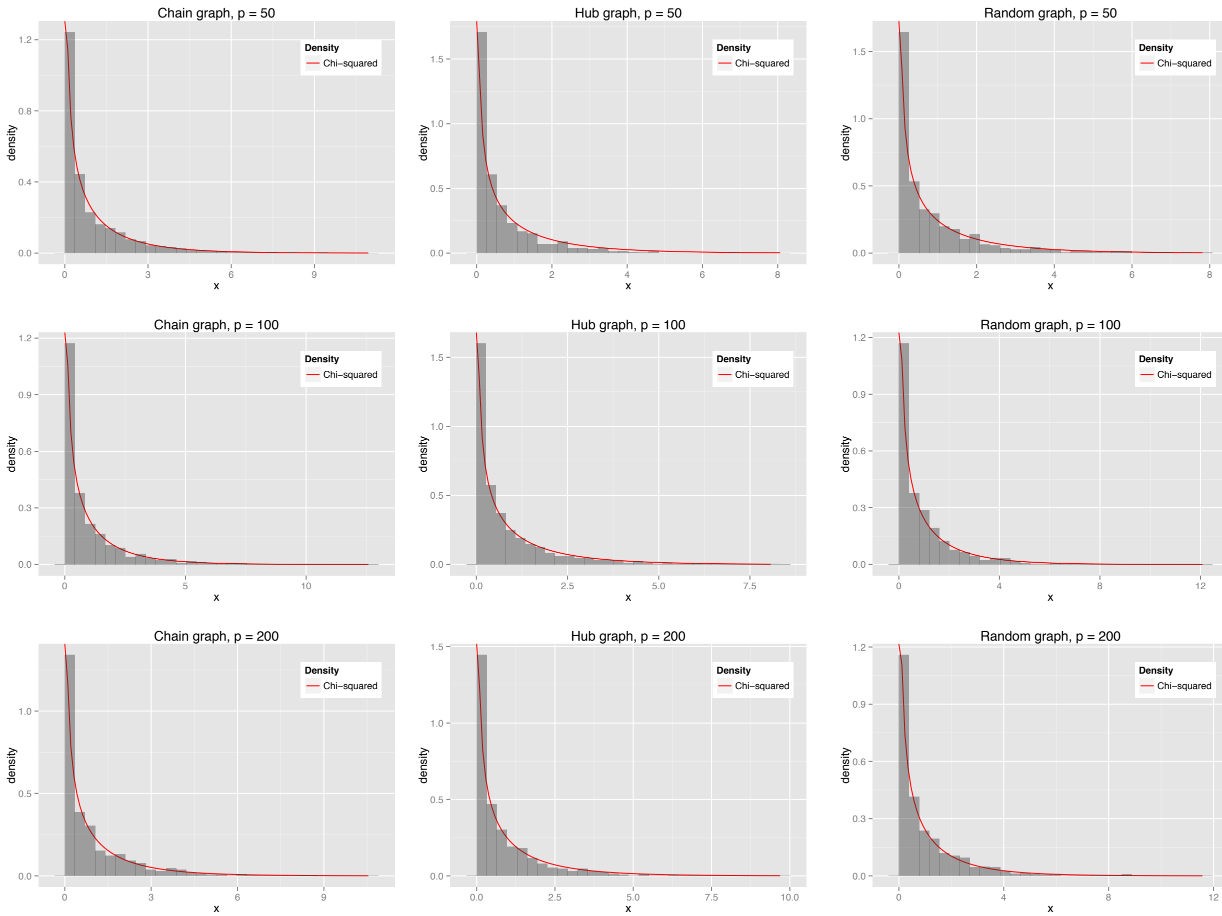

Figure 2: Empirical null distribution of the proposed CMLR test based on the chi-square approximation with $n = 200$.

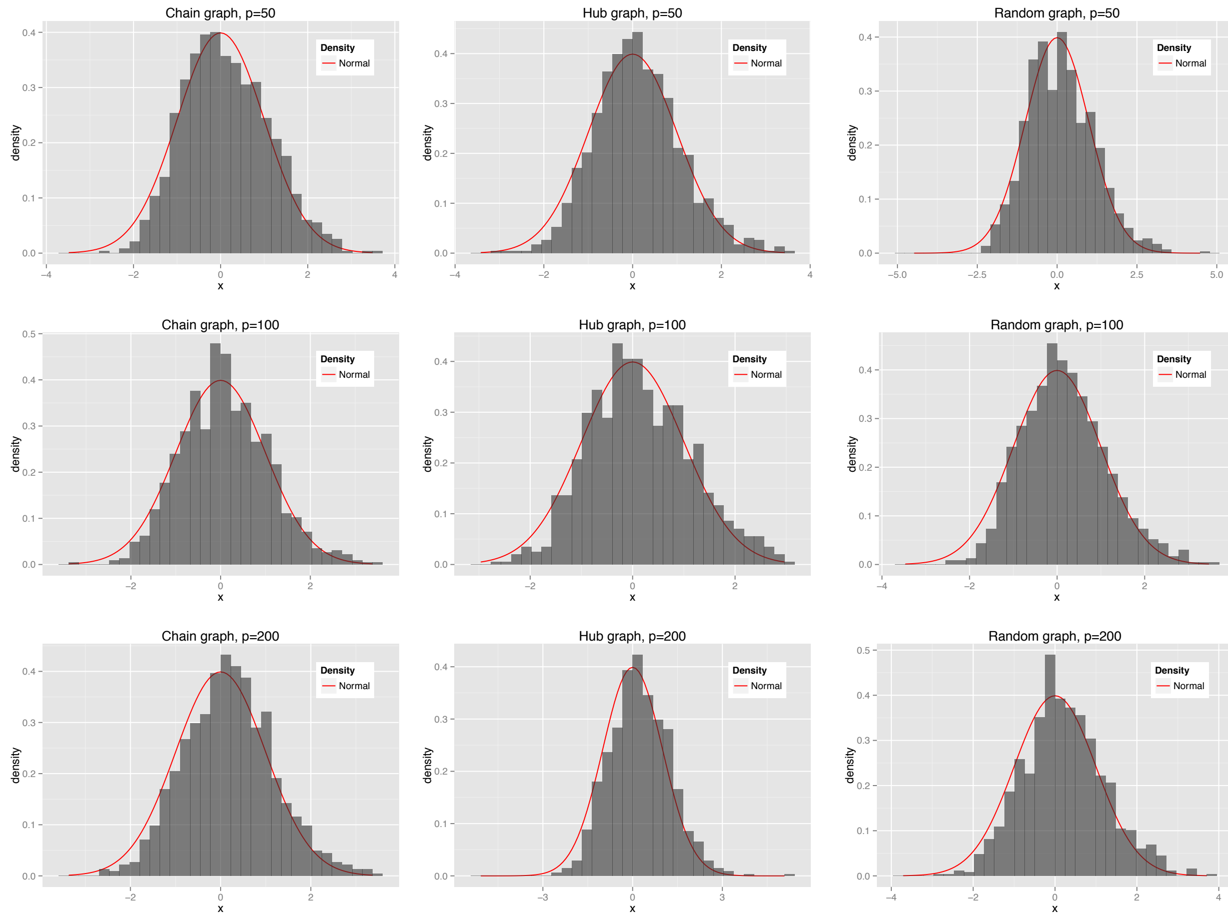

Figure 3: Empirical null distribution of our likelihood ratio test based on the normal approximation for the second testing problem involving a single column/row.

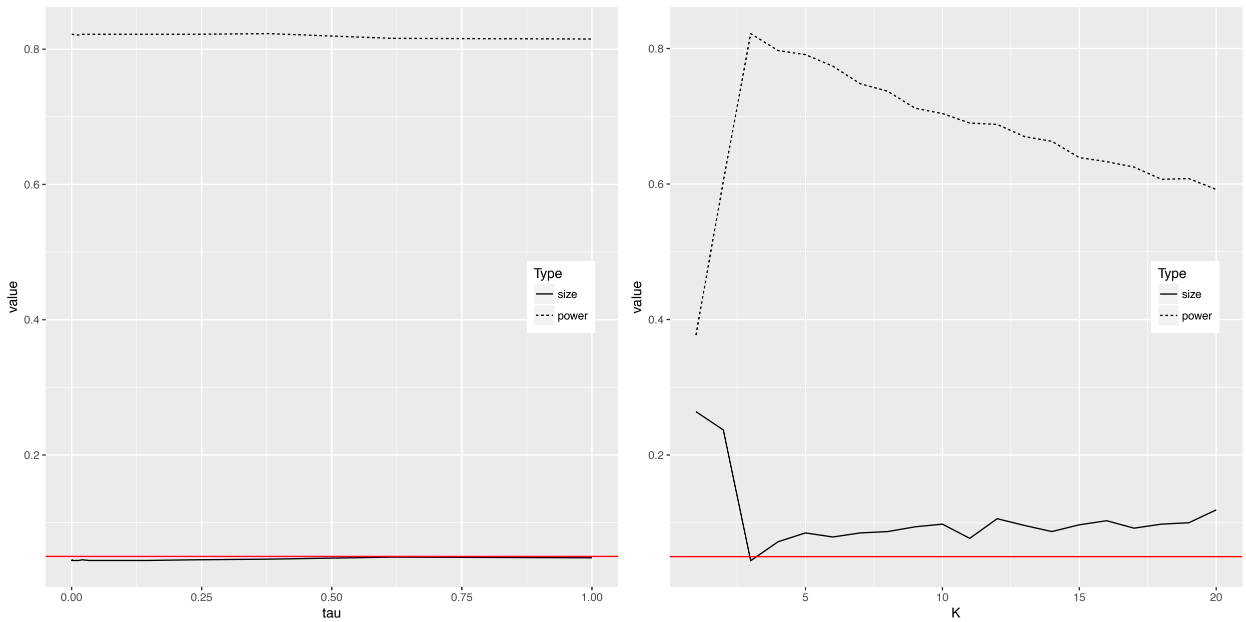

Figure 4: Sensitivity study of power as a function of tuning parameters τ and K when $n = 100$, $p = 100$, and $K_0 = 3$ in the linear regression problem based on 1000 simulations. Dotted and black lines represent empirical power and sizes of the proposed method, while red lines serve as a reference of the nominal size $\alpha = .05$.

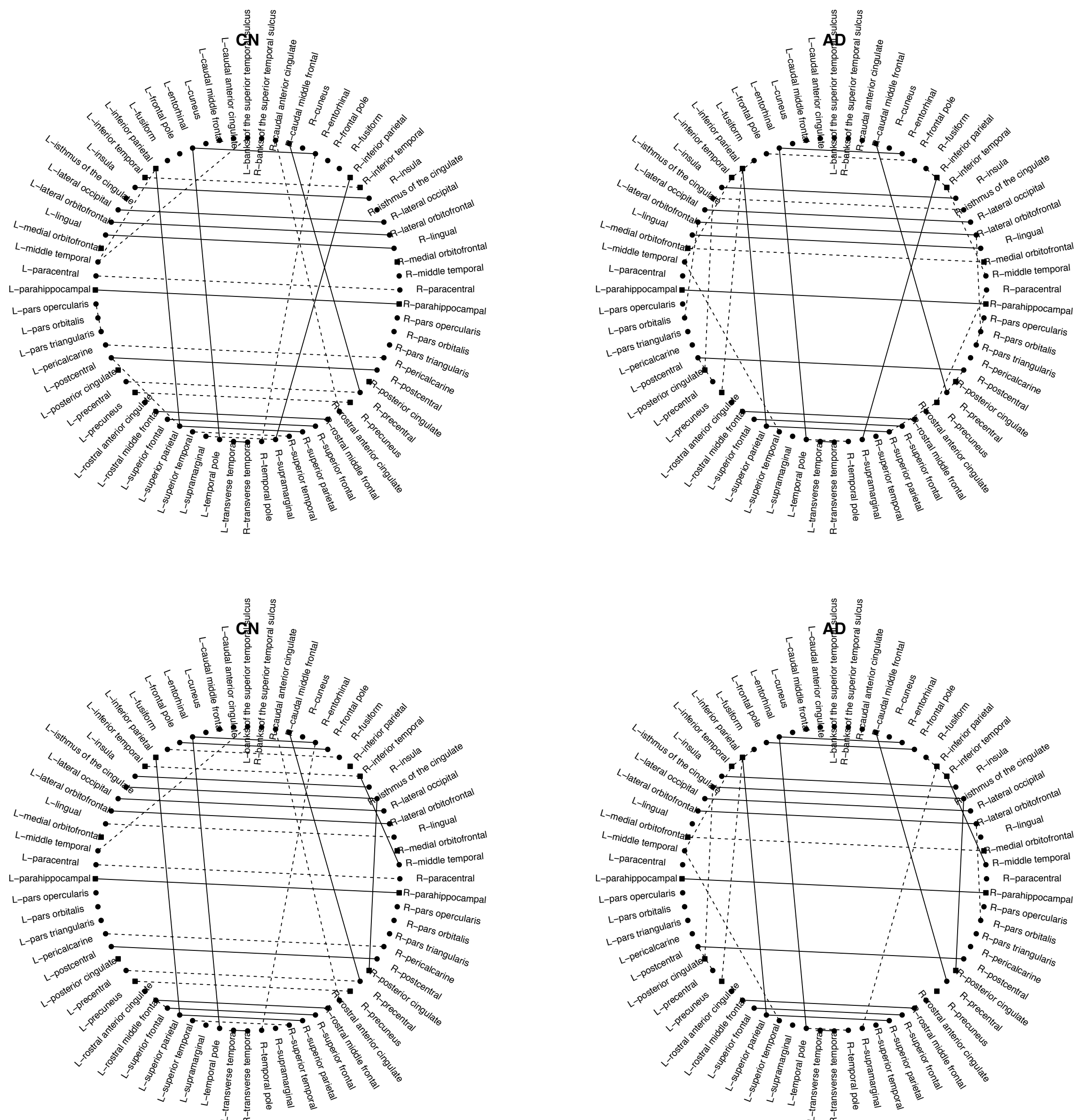

Figure 5: Estimated networks by the proposed method (first row) and the method [11] (second row) for the CN (Left) and AD (right) groups, where reported edges are significant under a p-value of .05 after Bonferroni correction. Nodes with square shape belong to DMN. The solid edges denote those that are shared by the two groups, whereas the dashed edges denote those that are only present within one group.

		CMLR-chi-square		JG	
Graph	(n,p)	Size	Power	Size	Power
band	(200,50)	.054	(.27, .78, .98, 1.0)	.043	(.24, .77, .99, 1.0)
	(200,100)	.055	(.30, .79, .98, 1.0)	.042	(.24, .75, .99, 1.0)
	(200,200)	.048	(.29, .80, .99, 1.0)	.036	(.23, .74, .98, 1.0)
hub	(200,50)	.019	(.10, .36, .74, .95)	.005	(.06, .27, .66, .92)
	(200,100)	.028	(.12, .43, .81, .96)	.005	(.02, .17, .54, .86)
	(200,200)	.031	(.16, .55, .86, .98)	.001	(.02, .15, .50, .86)
random	(200,50)	.034	(.15, .51, .86, .98)	.025	(.14, .49, .83, .98)
	(200,100)	.041	(.21, .68, .94, 1.0)	.018	(.11, .53, .92, .99)
	(200,200)	.049	(.15, .47, .81, .96)	.034	(.14, .41, .78, .95)

Table 1: Empirical size and power comparisons of the proposed CMLR test and test of [11], denoted by **CMLR-chi-square** and **JG**, in the first testing problem for the Gaussian graphical model based on 1000 simulations.

$\|B\|$	n	p	Method	Size	Power	$\hat{K}$
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 & 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 & 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 & 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 & 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 & 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 & 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)
			CMLR-normal	0.061	$(0.118, 0.315, 0.618, 0.893)$	NA
			Zhang & 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 & 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 & 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 & 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 & 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 & Cheng	NA	NA	NA
			DL	NA	NA	NA

Table 2: Empirical size and power comparisons in linear regression as well as estimated tuning parameter $\hat{K}$ by a 5-fold cross-validation over 1000 simulations. Here "CMLR-chi-square", "CMLR-normal", "DL", and "Zhang & Cheng" denote the proposed test based on a chi-square approximation, a normal approximation, the debias method of [28], and the method of [29]. Note that the nominal size is 0.05, DL is a test converted from a confidence interval, 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.

Appendix for "On high-dimensional constrained maximum likelihood inference"

A Technical details of the counter example

Lemma 1 *(A counter example) In (5) in the main text, we write* $y = \beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}$*, where* $\boldsymbol{x} = (x_1, \ldots, x_p)$ *are independently distributed from* $N(\mu_i, 1)$ *with* $\mu_1 = 0$ *and* $\mu_j = 1$*;* $2 \leq j \leq p$*, and* ϵ *is* $N(0, 1 - n^{-1})$*, independent of* $\boldsymbol{x}$*. Assume that* $\beta_0 = 0$ *and* $\boldsymbol{\beta} = n^{-1/2}, 0, \ldots, 0)$*, 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* $B = \{0\}$*.*

Proof of Lemma 1. Under the linear model, we have that

$$y_i = \beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}_i + \epsilon_i; i = 1, \ldots, n \,, \tag{A.1}$$

where $\boldsymbol{\beta} = (\beta_1, 0, \cdots, 0)$ and $\beta_0 = 0$, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip}) \sim N(\boldsymbol{\mu}, \boldsymbol{I}_{p \times p})$, and $\epsilon_i \sim N(0, 1 - \beta_1^2)$ and is independent of $\boldsymbol{x}_i$. Then, the constrained MLE for β_0 is

$$\hat{\beta}_0^{(1)} = \operatorname*{argmin}_{\sum_{i=1}^p \mathbb{I}(\beta_i \neq 0) \leq 1} \sum_{i=1}^n (y_i - \beta_0 - \boldsymbol{\beta}^\top \boldsymbol{x}_i)^2 = \bar{y} - \widehat{\mathrm{cor}}(x_{\cdot j^\star}, y) \frac{s_y}{s_{x_{\cdot j^\star}}} \bar{x}_{\cdot j^\star} \,, \tag{A.2}$$

where $x_{\cdot j}$ denotes a n-dimensional vector $(x_{1j}, \ldots, x_{nj})$, $\widehat{\mathrm{cor}}$ denotes the sample correlation between two vectors, $\bar{x}$ and s_x denote the sample mean and sample covariance of a vector x, respectively, and

$$j^\star = \underset{1\le j\le p}{\operatorname{argmax}}\ \widehat{\operatorname{cor}}(x_{\cdot j}, y) \tag{A.3}$$

denotes the index of which feature has the largest sample correlation between y. For each observation $(y_i, \boldsymbol{x}_i)$, it is easy to write out its joint distribution

$$(y_i, x_{i1}, \ldots, x_{ip}) \sim N\left((\beta_1\mu_1, \mu_1, \ldots, \mu_p)^\top, \begin{pmatrix} 1 & \beta_1 & 0 & \cdots & 0 \\ \beta_1 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}\right). \tag{A.4}$$

Hence, the conditional distribution of $\boldsymbol{x}_i$ given y_i is

$$\boldsymbol{x}_i|y_i \sim N\left((\beta_1(y_i - \beta_1\mu_1) + \mu_1, \mu_2, \ldots, \mu_p)^\top, \begin{pmatrix} 1-\beta_1^2 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}\right) \tag{A.5}$$

from which we can easily see that components of $\boldsymbol{x}_i$ are conditionally independent given y_i. Note that

$$\widehat{\operatorname{cor}}(x_{\cdot j}, y) = \frac{(n-1)^{-1}\sum_{i=1}^n x_{ij}(y_i - \bar{y})}{s_{\cdot j}s_y}, j = 1, \ldots, p \tag{A.6}$$

and $\operatorname{Var}(y) = \operatorname{Var}(x_{ij}) = 1$. Hence,

$$\sqrt{n}\widehat{\operatorname{cor}}(x_{\cdot j}, y)|y \stackrel{d}{=} Z_j + o_p(1)\,, \tag{A.7}$$

where $Z_j = \frac{\sum_{i=1}^n x_{ij}(y_i - \bar{y})}{(n-1)s_y}, j = 1, \ldots, p$, and Z_j's are independent and normally distributed conditioned on $\boldsymbol{y}$. By (A.5), we have that

$$Z_1 \sim N\left(\beta_1 s_y, 1-\beta_1^2\right) \text{ and } Z_j \sim N(0,1) \text{ for } j=2,\ldots,p\,. \tag{A.8}$$

Consequently, conditioned on $\boldsymbol{y}$,

$$\hat{\beta}_0^{(1)} = \bar{y} - \widehat{\text{cor}}(x_{\cdot j^\star}, y)\frac{s_y}{s_{x_{\cdot j^\star}}}\bar{x}_{\cdot j^\star} = \bar{y} - \beta_1\mu_1 + \beta_1\mu_1 - \widehat{\text{cor}}(x_{\cdot j^\star}, y) s_y \frac{\bar{x}_{\cdot j^\star} - \mu_{j^\star}}{s_{x_{\cdot j^\star}}} - \widehat{\text{cor}}(x_{\cdot j^\star}, y) s_y \frac{\mu_{j^\star}}{s_{x_{\cdot j^\star}}}$$

Now, we let $\mu_1 = 0$ and $\mu_2 = \cdots = \mu_p = 1$. Moreover, note that

$$\bar{y} - \beta_1\mu_1 = O_p\left(\frac{1}{\sqrt{n}}\right) \text{ and } \left|\frac{\bar{x}_{j^\star} - \mu_{j^\star}}{s_{x_{j^\star}}}\right| \leq \max_{1\leq j\leq p}\left|\frac{\bar{x}_j - \mu_j}{s_{x_j}}\right| \leq O\left(\sqrt{\frac{\log p}{n}}\right). \tag{A.9}$$

Hence, if $\sqrt{\frac{\log p}{n}} \leq O(1)$, then

$$\hat{\beta}_0^{(1)} = -\widehat{\text{cor}}(x_{\cdot j^\star}, y) s_y \frac{\mu_{j^\star}}{s_{x_{\cdot j^\star}}} + O_p\left(\frac{1}{\sqrt{n}}\right). \tag{A.10}$$

Now we choose β_1 to be small number so that with nonzero probability $\{j^\star \neq 1\}$, that is, we need $\mathbb{P}(Z_1 \leq \min_{2\leq j\leq p} Z_j)$ to be nonzero, which is easy to achieve when β_1 is chosen to be close to 0. Under the event $\{j^\star \geq 2\}$

$$\begin{aligned}
\hat{\beta}_0^{(1)} &= -\widehat{\text{cor}}(x_{\cdot j^\star}, y) s_y \frac{\mu_{j^\star}}{s_{x_{\cdot j^\star}}} + O_p\left(\frac{1}{\sqrt{n}}\right) = -\max_{2\leq j\leq p}\widehat{\text{cor}}(x_{\cdot j}, y)\frac{s_y}{s_{x_{\cdot j^\star}}} + O_p\left(\frac{1}{\sqrt{n}}\right) \\
&= O_p\left(\sqrt{\frac{\log p}{n}}\right) + O_p\left(\frac{1}{\sqrt{n}}\right),
\end{aligned}$$

because $\max_{2\leq j\leq p}\widehat{\text{cor}}(x_{\cdot j}, y) = O_p\left(\sqrt{\frac{\log p}{n}}\right)$ and $s_y \to 1$ in probability and $s_{x_{\cdot j^\star}} \to 1$ in probability. Hence, $n\left(\hat{\beta}_0^{(1)}\right)^2 \to \infty$ if $p \to \infty$ as $n \to \infty$. Next, we show that under this model, the log-likelihood ratio test statistic is of the same order as $n\hat{\beta}_0^2$ under the null model.

Toward this end, denote by $f(\beta_0) = \sup_{\|\boldsymbol{\beta}\|_0 \leq 1, \sigma > 0} n^{-1} L_n(\beta_0, \boldsymbol{\beta}, \sigma)$. By definition of $\hat{\beta}_0^{(1)}$, it must maximizes $f(\beta_0)$ as a function of β_0 and hence must satisfies $f'(\hat{\beta}_0^{(1)}) = 0$. Moreover, we note that the log-likelihood ratio can be rewritten in terms of $f(\cdot)$

$$\Lambda_n(B) = 2n(f(\hat{\beta}_0^{(1)}) - f(0)) \tag{A.11}$$

Applying a Taylor expansion around $\hat{\beta}_0^{(1)}$, we obtain

$$\Lambda_n(B) = -n(\hat{\beta}_0^{(1)})^2 f''(\beta^\star) \tag{A.12}$$

where $\beta^\star$ is some number between 0 and $\hat{\beta}_0^{(1)}$. Under $\log p/n \to 0$, it is easy to show that $\hat{\beta}_0^{(1)}$ is consistent, hence converges to 0 in probability. Hence, $\Lambda_n(B) = -n(\hat{\beta}_0^{(1)})^2 (f''(0) + o_p(1)) \xrightarrow{\mathbb{P}} \infty$, which completes the proof.

B Proofs of Lemmas 2-9

This section provides detailed proofs of Lemmas 2-9 to be used in "On high-dimensional constrained maximum likelihood inference".

Lemma 2 *For any symmetric matrices $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$, $\mathrm{vec}(\boldsymbol{C}_1)^\top \mathrm{vec}(\boldsymbol{C}_2) = \mathrm{tr}(\boldsymbol{C}_1 \boldsymbol{C}_2)$. Moreover, for any positive definite matrix $\boldsymbol{C} \succ 0$,*

$$\nabla \left(\log\det \boldsymbol{C}\right) = -\,\mathrm{vec}(\boldsymbol{C}^{-1})\,, \quad \nabla^2\left(-\log\det \boldsymbol{\Omega}^0\right) = \boldsymbol{C}^{-1} \otimes_s \boldsymbol{C}^{-1}\,, \tag{B.1}$$

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

$$\mathrm{Var}\left(\mathrm{vec}(\boldsymbol{X}\boldsymbol{X}^\top)\right) = 4\boldsymbol{I} \text{ with } \boldsymbol{X} \sim N(0, \boldsymbol{\Sigma}^0)\,, \tag{B.3}$$

$$\mathrm{vec}(\boldsymbol{C})^\top \boldsymbol{I}\,\mathrm{vec}(\boldsymbol{C}) = \tfrac{1}{2}\,\mathrm{tr}\left(\boldsymbol{\Sigma}^0 \boldsymbol{C} \boldsymbol{\Sigma}^0 \boldsymbol{C}\right)\,. \tag{B.4}$$

Proof of Lemma 2: By the definition, (B.1) follows from an identity:

$$\mathrm{vec}(\boldsymbol{C}_1)^\top \mathrm{vec}(\boldsymbol{C}_2) = \sum_{i\le j}(1+\mathbb{I}(i\neq j))\boldsymbol{S}_1(i,j)\boldsymbol{S}_2(i,j) = \sum_{i,j}\boldsymbol{S}_1(i,j)\boldsymbol{S}_2(i,j) = \mathrm{tr}(\boldsymbol{S}_1\boldsymbol{S}_2)\,.$$

Moreover, it follows from Taylor's expansion of the log det function that

$$\begin{aligned}
&\log\det(\boldsymbol{C}+\boldsymbol{\Delta}) - \log\det(\boldsymbol{C}) = \mathrm{tr}(\boldsymbol{C}^{-1}\boldsymbol{\Delta}) - \frac{1}{2}\mathrm{tr}\left((\boldsymbol{C}^{-1}\boldsymbol{\Delta})^2\right) + o(\|\boldsymbol{C}^{-1/2}\boldsymbol{\Delta}\boldsymbol{C}^{-1/2}\|_F^2)\\
= \;& \mathrm{vec}(\boldsymbol{C}^{-1})^\top \mathrm{vec}(\boldsymbol{\Delta}) - \frac{1}{2}\mathrm{vec}(\boldsymbol{\Delta})^\top \mathrm{vec}(\boldsymbol{C}^{-1}\boldsymbol{\Delta}\boldsymbol{C}^{-1}) + o(\|\boldsymbol{C}^{-1/2}\boldsymbol{\Delta}\boldsymbol{C}^{-1/2}\|_F^2)\\
= \;& \mathrm{vec}(\boldsymbol{C}^{-1})^\top \mathrm{vec}(\boldsymbol{\Delta}) - \frac{1}{2}\mathrm{vec}(\boldsymbol{\Delta})^\top \left(\boldsymbol{C}^{-1}\otimes_s \boldsymbol{C}^{-1}\right)\mathrm{vec}(\boldsymbol{\Delta}) + o(\|\boldsymbol{C}^{-1/2}\boldsymbol{\Delta}\boldsymbol{C}^{-1/2}\|_F^2)\,,
\end{aligned}$$

where the definition of $\otimes_s$ and (B.1) have been used. This yields (B.2).

For (B.3), the log-likelihood for $\boldsymbol{X}\sim N(0,\boldsymbol{\Sigma}^0)$ is $-\frac{1}{2}\mathrm{vec}(\boldsymbol{\Omega}^0)^\top\mathrm{vec}(\boldsymbol{X}\boldsymbol{X}^\top)+\frac{1}{2}\log\det(\boldsymbol{\Omega}^0)$. Using properties of the exponential family [2], $\mathrm{Var}\left(\frac{1}{2}\mathrm{vec}(\boldsymbol{X}\boldsymbol{X}^\top)\right) = \nabla^2\left(-\frac{1}{2}\log\det\boldsymbol{\Omega}^0\right) = \boldsymbol{I}$, implying (B.3). Finally, for any symmetric matrix $\boldsymbol{C}$, note that

$$\begin{aligned}
\mathrm{vec}(\boldsymbol{C})^\top\boldsymbol{I}\,\mathrm{vec}(\boldsymbol{C}) &= \frac{1}{2}\mathrm{vec}(\boldsymbol{C})^\top\left(\boldsymbol{\Sigma}^0\otimes_s\boldsymbol{\Sigma}^0\right)\mathrm{vec}(\boldsymbol{C})\\
&= \frac{1}{2}\mathrm{vec}(\boldsymbol{C})^\top\mathrm{vec}(\boldsymbol{\Sigma}^0\boldsymbol{C}\boldsymbol{\Sigma}^0) = \frac{1}{2}\mathrm{tr}(\boldsymbol{C}\boldsymbol{\Sigma}^0\boldsymbol{C}\boldsymbol{\Sigma}^0),
\end{aligned}$$

leading to (B.4). This completes the proof.

Lemma 3 *For any symmetric matrix $\boldsymbol{T}$ and $\nu>0$*

$$\mathbb{P}\left(|\,\mathrm{tr}\left((\boldsymbol{S}-\boldsymbol{\Sigma}^0)\boldsymbol{T}\right)|\ge\nu\right) \le 2\exp\left(-n\frac{\nu^2}{9\|\boldsymbol{T}\|^2+8\nu\|\boldsymbol{T}\|}\right)\,, \tag{B.5}$$

where $\|\boldsymbol{T}\|^2=\frac{n}{2}\mathrm{Var}\left(\mathrm{tr}\left((\boldsymbol{S}-\boldsymbol{\Sigma}^0)\boldsymbol{T}\right)\right)$. *Furthermore, for* $\boldsymbol{T}_1,\cdots,\boldsymbol{T}_K$ *such that* $\|\boldsymbol{T}_k\|\le c_0; k=$

$1, \cdots, K$ *with* $c_0 > 0$ *and any* $\nu > 0$*, we have that*

$$\mathbb{P}\left(\max_{1 \le k \le K} \left|\operatorname{tr}((\boldsymbol{S} - \boldsymbol{\Sigma}^0)\boldsymbol{T}_k)\right| \ge \nu\right) \le 2\exp\left(-n\frac{\nu^2}{9c_0^2 + 8c_0\nu} + \log K\right), \tag{B.6}$$

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

$$\mathbb{P}\left(\|\operatorname{vec}_B(\boldsymbol{S} - \boldsymbol{\Sigma}^0)\|_\infty \ge \nu\right) \le 2\exp\left(-n\frac{\nu^2}{9\lambda_{max}^2(\boldsymbol{\Sigma}^0) + 8\nu\lambda_{max}(\boldsymbol{\Sigma}^0)} + \log|B|\right), \tag{B.7}$$

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

Proof of Lemma 3: By Markov's inequality, for any $\nu > 0$,

$$\begin{aligned} P\left(\operatorname{tr}\left((\boldsymbol{S} - \boldsymbol{\Sigma}^0)\boldsymbol{T}\right) \ge \nu\right) &\le \exp\left(-\frac{\gamma\sqrt{n}\nu}{2}\right)\mathbb{E}\exp\left(\frac{\gamma\sqrt{n}}{2}\operatorname{tr}\left((\boldsymbol{S} - \boldsymbol{\Sigma}^0)\boldsymbol{T}\right)\right) \\ &\le \exp\Bigg(\underbrace{\log\mathbb{E}\exp\left(\frac{\gamma\sqrt{n}}{2}tr\left((\boldsymbol{S} - \boldsymbol{\Sigma}^0)\boldsymbol{T}\right)\right) - \frac{\gamma\sqrt{n}\nu}{2}}_{I_1}\Bigg), \end{aligned}$$

where γ is chosen such that $\gamma \in \left[0, \frac{M_0\sqrt{n}}{\|\sqrt{\boldsymbol{\Sigma}^0}\boldsymbol{T}\sqrt{\boldsymbol{\Sigma}^0}\|_F}\right]$ for some constant $0 < M_0 < 1$, which is to be determined later. Moreover, after some calculations, we have that

$$\begin{aligned} \mathbb{E}\exp\left(\frac{\gamma\sqrt{n}}{2}\operatorname{tr}\left((\boldsymbol{S} - \boldsymbol{\Sigma}^0)\boldsymbol{T}\right)\right) &= \left(\mathbb{E}\exp\left(\frac{\gamma\sqrt{n}}{2}\operatorname{tr}\left((\boldsymbol{X}\boldsymbol{X}^T - \boldsymbol{\Sigma}^0)\boldsymbol{T}\right)\right)\right)^n \\ &= \exp\left(-\frac{\gamma\sqrt{n}}{2}\operatorname{tr}(\boldsymbol{\Sigma}^0\boldsymbol{T})\right)\det\left(\boldsymbol{I} - \frac{\gamma}{\sqrt{n}}\boldsymbol{\Sigma}^0\boldsymbol{T}\right)^{-n/2}, \end{aligned} \tag{B.8}$$

where $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}^0)$ and the last equality requires that $\sqrt{n}\boldsymbol{\Omega}^0 \succeq \gamma\boldsymbol{T}$, which is ensured by the fact that $\gamma \le \frac{M_0\sqrt{n}}{\|\sqrt{\boldsymbol{\Sigma}^0}\boldsymbol{T}\sqrt{\boldsymbol{\Sigma}^0}\|_F} < \frac{\sqrt{n}}{\|\sqrt{\boldsymbol{\Sigma}^0}\boldsymbol{T}\sqrt{\boldsymbol{\Sigma}^0}\|_F}$. Consequently,

$$\log\mathbb{E}\exp\left(\frac{\gamma\sqrt{n}}{2}\operatorname{tr}\left((\boldsymbol{S} - \boldsymbol{\Sigma}^0)\boldsymbol{T}\right)\right) = \log\det\left(\boldsymbol{I} - \frac{\gamma}{\sqrt{n}}\boldsymbol{\Sigma}^0\boldsymbol{T}\right)^{-n/2} - \frac{\gamma\sqrt{n}}{2}\operatorname{tr}(\boldsymbol{\Sigma}^0\boldsymbol{T}). \tag{B.9}$$

An expansion of the log det function gives

$$\begin{aligned}
&\log\det(\boldsymbol{I} - \frac{\gamma}{\sqrt{n}}\boldsymbol{\Sigma}^0\boldsymbol{T})^{-n/2} \\
= \; & \frac{\gamma\sqrt{n}}{2}\operatorname{tr}(\boldsymbol{\Sigma}^0\boldsymbol{T}) + \frac{\gamma^2}{4}\operatorname{tr}((\boldsymbol{\Sigma}^0\boldsymbol{T})^2) + \underbrace{\frac{n}{2}\sum_{l=3}^{\infty} l^{-1}\operatorname{tr}\big((\frac{\gamma\boldsymbol{\Sigma}^0\boldsymbol{T}}{\sqrt{n}})^l\big)}_{I_2}.
\end{aligned} \tag{B.10}$$

For I_2, note that $I_2 \leq \frac{n}{2}\sum_{l=3}^{\infty} l^{-1}\left(\frac{\gamma\|\boldsymbol{T}\|}{\sqrt{n}}\right)^l \leq \gamma^2\|\boldsymbol{T}\|^2\frac{3-M_0}{12(1-M_0)}$. Similarly, $I_1 \leq \frac{M_1+1}{4}\gamma^2\|\boldsymbol{T}\|^2 - \frac{\gamma\sqrt{n}\nu}{2}$, where $M_1 = \frac{3-M_0}{3(1-M_0)}$. Minimizing this upper bound of I_1 as a function of γ over the interval $\left[0, \frac{M_0\sqrt{n}}{\|\boldsymbol{T}\|}\right]$, we obtain that

$$\begin{array}{ll}
I_1 \leq -\frac{n\nu^2}{4(1+M_1)\|\boldsymbol{T}\|^2} & \text{if } \nu \leq M_0(1+M_1)\|\boldsymbol{T}\| \\
I_1 \leq -\frac{nM_0}{2\|\boldsymbol{T}\|}\left(\nu - \frac{M_0(1+M_1)}{2}\|\boldsymbol{T}\|\right) & \text{otherwise.}
\end{array}$$

A combination of these two cases yields that $I_1 \leq -\frac{nM_0\nu^2}{4M_0(M_1+1)\|\boldsymbol{T}\|^2+2\nu\|\boldsymbol{T}\|}$. Set $M_0 = 4^{-1}$, and then $M_1 = 11/9$, we obtain the desired results

$$P\Big(\operatorname{tr}\big((\boldsymbol{S}-\boldsymbol{\Sigma}^0)\boldsymbol{T}\big) \geq \nu\Big) \leq \exp\Big(-n\frac{\nu^2}{9\|\boldsymbol{T}\|^2 + 8\nu\|\boldsymbol{T}\|}\Big),$$

for any $\nu > 0$. The other direction follows exactly the same argument, and thus is omitted.

Finally, (B.7) follows by letting $\{\boldsymbol{T}_1, \cdots, \boldsymbol{T}_k\} = \big\{(\boldsymbol{e}_i^\top\boldsymbol{e}_j + \boldsymbol{e}_j^\top\boldsymbol{e}_i)/2\big\}_{(i,j)\in B}$ then applying an inequality $\|\sqrt{\boldsymbol{\Sigma}^0}(\boldsymbol{e}_i^\top\boldsymbol{e}_j + \boldsymbol{e}_j^\top\boldsymbol{e}_i)\sqrt{\boldsymbol{\Sigma}^0}/2\|_F^2 \leq \lambda_{\max}(\boldsymbol{\Sigma}^0)$ and a union bound. This completes the proof.

Lemma 4 *(The Kullback-Leibler divergence and Fisher-norm) For a positive definite matrix*

$\mathbf{\Omega}$ *the following connection holds:*

$$K(\mathbf{\Omega}^0, \mathbf{\Omega}) \geq \min\left(\frac{1}{16\sqrt{2}}, \frac{\sqrt{K(\mathbf{\Omega}^0, \mathbf{\Omega})}}{2\sqrt{6}}\right) \|\mathbf{\Omega} - \mathbf{\Omega}^0\|, \tag{B.11}$$

$$K(\mathbf{\Omega}^0, \mathbf{\Omega}) \geq \min\left(\frac{1}{16\sqrt{2}}, \frac{\|\mathbf{\Omega} - \mathbf{\Omega}^0\|}{24}\right) \|\mathbf{\Omega} - \mathbf{\Omega}^0\|. \tag{B.12}$$

Proof of Lemma 4: Let $\mathbf{\Delta} = \mathbf{\Omega} - \mathbf{\Omega}^0$ and $\lambda_1, \cdots, \lambda_p$ be the eigenvalues of $\sqrt{\mathbf{\Sigma}^0}\mathbf{\Delta}\sqrt{\mathbf{\Sigma}^0}$. Then $\lambda_j > -1$; $j = 1, \cdots, p$, because $\boldsymbol{I}_{p\times p} + \sqrt{\mathbf{\Sigma}^0}\mathbf{\Delta}\sqrt{\mathbf{\Sigma}^0} = \sqrt{\mathbf{\Sigma}^0}\mathbf{\Omega}\sqrt{\mathbf{\Sigma}^0}$ is positive definite. Moreover, let $B_1 = \sum_{i=1}^p \lambda_i^2 \mathbb{I}(\lambda_i \leq 1/3)$, $B_2 = \sum_{i=1}^p \lambda_i^2 \mathbb{I}(\lambda_i > 1/3)$, and $B_3 = \sum_{i=1}^p \lambda_i \mathbb{I}(\lambda_i > 1/3)$. Easily, $\|\mathbf{\Omega} - \mathbf{\Omega}^0\| = \sqrt{B_1 + B_2}$. Using the inequality $x - \log(1+x) \geq 6^{-1}x^2\mathbb{I}(x \leq 1/3) + 8^{-1}x\mathbb{I}(x > 1/3)$ for $x > -1$, we have that

$$\begin{aligned} K(\mathbf{\Omega}^0, \mathbf{\Omega}) &= \frac{1}{2}\left(\mathrm{tr}(\sqrt{\mathbf{\Sigma}^0}\mathbf{\Delta}\sqrt{\mathbf{\Sigma}^0}) - \log\det(\boldsymbol{I}_{p\times p} + \sqrt{\mathbf{\Sigma}^0}\mathbf{\Delta}\sqrt{\mathbf{\Sigma}^0})\right) \\ &= \frac{1}{2}\sum_{i=1}^p \lambda_i - \frac{1}{2}\sum_{i=1}^p \log(1+\lambda_i) \\ &\geq 12^{-1}\sum_{i=1}^p \lambda_i^2 \mathbb{I}(\lambda_i \leq 1/3) + 16^{-1}\sum_{i=1}^p \lambda_i \mathbb{I}(\lambda_i > 1/3) = 12^{-1}B_1 + 16^{-1}B_3\,. \end{aligned}$$

Next we examine two cases. First, if $B_1 < B_2$, then $\frac{K(\mathbf{\Omega}^0,\mathbf{\Omega})}{\|\mathbf{\Omega}-\mathbf{\Omega}^0\|} \geq \frac{12^{-1}B_1 + 16^{-1}B_3}{\sqrt{B_1+B_2}} \geq \frac{B_3}{16\sqrt{2B_2}} \geq \frac{1}{16\sqrt{2}}$ because $B_3^2 \geq B_2$. If $B_1 \geq B_2$, then

$$\frac{K(\mathbf{\Omega}^0, \mathbf{\Omega})}{\|\mathbf{\Omega} - \mathbf{\Omega}^0\|} \geq \frac{12^{-1}B_1 + 16^{-1}B_3}{\sqrt{B_1+B_2}} \geq \frac{B_1}{12\sqrt{B_1+B_2}} \geq \frac{B_1+B_2}{24\sqrt{B_1+B_2}} \geq \frac{\sqrt{B_1+B_2}}{24} = \frac{\|\mathbf{\Omega} - \mathbf{\Omega}^0\|}{24}.$$

Similarly,

$$\frac{K(\mathbf{\Omega}^0, \mathbf{\Omega})}{\|\mathbf{\Omega} - \mathbf{\Omega}^0\|} \geq \sqrt{K(\mathbf{\Omega}^0, \mathbf{\Omega})}\frac{\sqrt{12^{-1}B_1 + 16^{-1}B_3}}{\sqrt{B_1+B_2}} \geq \sqrt{K(\mathbf{\Omega}^0, \mathbf{\Omega})}\frac{\sqrt{24^{-1}(B_1+B_2)}}{\sqrt{B_1+B_2}} = \frac{\sqrt{K(\mathbf{\Omega}^0, \mathbf{\Omega})}}{2\sqrt{6}}.$$

This leads to (B.12) and (B.11).

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*

$$\|\widehat{\boldsymbol{\Omega}}_{\tilde{A}} - \boldsymbol{\Omega}^0\| \leq 12\|\boldsymbol{I}_{\tilde{A},\tilde{A}}^{-1/2}\operatorname{vec}(\boldsymbol{\Sigma}^0 - \boldsymbol{S})\|_2 \,. \tag{B.13}$$

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

$$\|\widehat{\boldsymbol{\Omega}}_{\tilde{A}} - \boldsymbol{\Omega}^0\| = O_p\left(\sqrt{\frac{|\tilde{A}|\log p}{n}}\right) \,. \tag{B.14}$$

Proof of Lemma 5: By definition of the CMLE, $L_n(\widehat{\boldsymbol{\Omega}}_{\tilde{A}}) - L_n(\boldsymbol{\Omega}^0) \geq 0$, or $-\log\det\widehat{\boldsymbol{\Omega}}_{\tilde{A}} + \log\det\boldsymbol{\Omega}^0 \leq -\operatorname{tr}((\widehat{\boldsymbol{\Omega}}_{\tilde{A}} - \boldsymbol{\Omega}^0)\boldsymbol{S})$. By the Cauchy-Schwarz inequality, this inequality becomes

$$\begin{aligned} 2K(\boldsymbol{\Omega}^0, \widehat{\boldsymbol{\Omega}}_{\tilde{A}}) &\leq \operatorname{tr}((\widehat{\boldsymbol{\Omega}}_{\tilde{A}} - \boldsymbol{\Omega}^0)(\boldsymbol{\Sigma}^0 - \boldsymbol{S})) \leq \|\sqrt{\boldsymbol{\Sigma}^0}(\widehat{\boldsymbol{\Omega}}_{\tilde{A}} - \boldsymbol{\Omega}^0)\sqrt{\boldsymbol{\Sigma}^0}\|_F\|\boldsymbol{I}_{\tilde{A},\tilde{A}}^{-1/2}\operatorname{vec}_{\tilde{A}}(\boldsymbol{\Sigma}^0 - \boldsymbol{S})\|_2 \\ &= \|\widehat{\boldsymbol{\Omega}}_{\tilde{A}} - \boldsymbol{\Omega}^0\|\|\boldsymbol{I}_{\tilde{A},\tilde{A}}^{-1/2}\operatorname{vec}_{\tilde{A}}(\boldsymbol{\Sigma}^0 - \boldsymbol{S})\|_2 \end{aligned} \tag{B.15}$$

On the other hand, by (B.12) $\frac{K(\boldsymbol{\Omega}^0, \widehat{\boldsymbol{\Omega}}_{\tilde{A}})}{\|\widehat{\boldsymbol{\Omega}}_{\tilde{A}} - \boldsymbol{\Omega}^0\|} \geq \min\left(\frac{1}{16\sqrt{2}}, \frac{\|\widehat{\boldsymbol{\Omega}}_{\tilde{A}} - \boldsymbol{\Omega}^0\|}{24}\right)$, which, together with (B.15), implies that $\min\left(\frac{1}{8\sqrt{2}}, \frac{\|\widehat{\boldsymbol{\Omega}}_{\tilde{A}} - \boldsymbol{\Omega}^0\|}{12}\right) \leq \|\boldsymbol{I}_{\tilde{A},\tilde{A}}^{-1/2}\operatorname{vec}_{\tilde{A}}(\boldsymbol{\Sigma}^0 - \boldsymbol{S})\|_2$. If $\frac{\|\widehat{\boldsymbol{\Omega}}_{\tilde{A}} - \boldsymbol{\Omega}^0\|}{12} \leq \frac{1}{8\sqrt{2}}$, then it follows immediately that $\|\widehat{\boldsymbol{\Omega}}_{\tilde{A}} - \boldsymbol{\Omega}^0\| \leq 12\|\boldsymbol{I}_{\tilde{A},\tilde{A}}^{-1/2}\operatorname{vec}_{\tilde{A}}(\boldsymbol{\Sigma}^0 - \boldsymbol{S})\|_2$. If $\frac{\|\widehat{\boldsymbol{\Omega}}_{\tilde{A}} - \boldsymbol{\Omega}^0\|}{12} > \frac{1}{8\sqrt{2}}$, then $\frac{1}{8\sqrt{2}} \leq \|\boldsymbol{I}_{\tilde{A},\tilde{A}}^{-1/2}\operatorname{vec}(\boldsymbol{\Sigma}^0 - \boldsymbol{S})\|_2$, which does not happen on the event $\{\|\boldsymbol{I}_{\tilde{A},\tilde{A}}^{-1/2}\operatorname{vec}_{\tilde{A}}(\boldsymbol{\Sigma}^0 - \boldsymbol{S})\|_2 < \frac{1}{8\sqrt{2}}\}$.

Moreover, by property of exponential family [2], $\operatorname{Var}(\operatorname{vec}_{\tilde{A}}(\boldsymbol{\Sigma}^0 - \boldsymbol{S})) = 4n^{-1}\boldsymbol{I}_{\tilde{A},\tilde{A}}$. Thus, $\operatorname{Var}(\boldsymbol{I}_{\tilde{A},\tilde{A}}^{-1/2}\operatorname{vec}_{\tilde{A}}(\boldsymbol{\Sigma}^0 - \boldsymbol{S})) = 4n^{-1}\boldsymbol{I}_{|\tilde{A}|\times|\tilde{A}|}$. This, combined with Lemma 3, implies that

$$\|\boldsymbol{I}_{\tilde{A},\tilde{A}}^{-1/2}\operatorname{vec}_{\tilde{A}}(\boldsymbol{\Sigma}^0 - \boldsymbol{S})\|_2 \leq \sqrt{|\tilde{A}|}\|\boldsymbol{I}_{\tilde{A},\tilde{A}}^{-1/2}\operatorname{vec}_{\tilde{A}}(\boldsymbol{\Sigma}^0 - \boldsymbol{S})\|_\infty = O_p\left(\sqrt{\frac{|\tilde{A}|\log p}{n}}\right) \tag{B.16}$$

on the event that $\{\|\boldsymbol{I}_{\tilde{A},\tilde{A}}^{-1/2}\operatorname{vec}_{\tilde{A}}(\boldsymbol{\Sigma}^0 - \boldsymbol{S})\|_2 < \frac{1}{8\sqrt{2}}\}$. This event, on the other hand, happens with probability tending to 1 by the assumption that $\frac{|\tilde{A}|\log p}{n} \to 0$. This completes the proof.

Lemma 6 *(Selection consistency) If* $K = |A^0|$, $\tau \le \frac{\bar{\lambda}_{min} \min(\sqrt{C_{min}}, C_{min}^2)}{12|A^0|}$, *then*

$$\begin{aligned} & \max\left(P\left(\widehat{\mathbf{\Omega}}^{(0)} \neq \widehat{\mathbf{\Omega}}_{A^0}\right), P\left(\widehat{\mathbf{\Omega}}^{(1)} \neq \widehat{\mathbf{\Omega}}_{A^0 \cup B}\right)\right) \\ \le \;& 2\exp\left(\frac{-nC_{min}}{2560} + 2\log p\right) + \exp\left(\frac{-n}{2560 \times 512} + |A^0|\log p\right) \\ & + 2\exp\left(-n\frac{\min\left(\sqrt{\frac{\min(C_{min}/512, 3/32)}{48\lambda_{max}^2(|A^0|+|B|)}}, \lambda_{max}(\mathbf{\Sigma}^0)\right)^2}{18\lambda_{max}^2(\mathbf{\Sigma}^0)} + 2\log p\right) \longrightarrow 0 \end{aligned} \quad \text{(B.17)}$$

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

Proof of Lemma 6: Let $\hat{A} = \{(i,j) : |\widehat{\omega}_{ij}^{(1)}| \ge \tau, (i,j) \notin B\}$. By definition, $|\hat{A}| \le |A^0|$, $\hat{A} \cap B = \emptyset$ and $\sum_{(i,j)\notin \hat{A}\cup B} |\widehat{\omega}_{ij}^{(1)}| \le \tau(|A^0| - |\hat{A}|)$. Hence, if $\hat{A} = A^0$, then $\widehat{\mathbf{\Omega}}^{(1)} = \widehat{\mathbf{\Omega}}_{A^0\cup B}$. Suppose $\hat{A} \neq A^0$. On event $\{\hat{A} = A\}$; with fixed $A \neq A^0$, $|A| \le |A^0|$, and $A \cap B = \emptyset$, we bound the Fisher-norm between $\widehat{\mathbf{\Omega}}_{A\cup B}^{(1)}$ and an approximating point of $\mathbf{\Omega}^0$, $\bar{\mathbf{\Omega}}_{A\cup B}^0 = \operatorname{argmin}_{\mathbf{\Omega}:\mathbf{\Omega}_{(A\cup B)^c}=\mathbf{0}} K(\mathbf{\Omega}^0, \mathbf{\Omega})$. Let $\bar{\mathbf{\Sigma}}_{A\cup B}^0 = (\bar{\mathbf{\Omega}}_{A\cup B}^0)^{-1}$. By the Karush-Kuhn-Tucker conditions, $\operatorname{vec}_{A\cup B}(\bar{\mathbf{\Sigma}}_{A\cup B}^0) = \operatorname{vec}_{A\cup B}(\mathbf{\Sigma}^0)$. Moreover, let $\bar{\lambda}_{\max} = \max_{A:|A|\le K, A\cap B=\emptyset} \lambda_{\max}(\bar{\mathbf{\Omega}}_{A\cup B}^0)$ and $\bar{\lambda}_{\min} = \min_{A:|A|\le K, A\cap B=\emptyset} \lambda_{\min}(\bar{\mathbf{\Omega}}_{A\cup B}^0)$. We also define

$$\mathcal{G} = \left\{\|\boldsymbol{S} - \mathbf{\Sigma}^0\|_\infty \le \min\left(\frac{1}{16\sqrt{2}\bar{\lambda}_{\max}\sqrt{|A^0|+|B|}}, \sqrt{\frac{\tilde{C}_{\min}}{48\bar{\lambda}_{\max}^2|A^0\cup B|}}, \lambda_{\max}(\mathbf{\Sigma}^0)\right)\right\},$$

where

$$\tilde{C}_{\min} = \min_{A:A\neq A^0,\, |A|=|A^0|,\, A\cap B=\emptyset} \min\left(\frac{\max(K(\mathbf{\Omega}^0, \bar{\mathbf{\Omega}}_{A\cup B}^0), K^2(\mathbf{\Omega}^0, \bar{\mathbf{\Omega}}_{A\cup B}^0))}{|A^0 \setminus A|}, 1\right). \quad \text{(B.18)}$$

By definition of the CMLE, $L_n(\widehat{\mathbf{\Omega}}^{(1)}) - L_n(\bar{\mathbf{\Omega}}_{A\cup B}^0) \ge 0$, or $-\log\det\widehat{\mathbf{\Omega}}^{(1)} + \log\det\bar{\mathbf{\Omega}}_{A\cup B}^0 \le -\operatorname{tr}((\widehat{\mathbf{\Omega}}^{(1)} - \bar{\mathbf{\Omega}}_{A\cup B}^0)\boldsymbol{S})$. Now let $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{\Omega}}_{A\cup B}^{(1)} - \bar{\mathbf{\Omega}}_{A\cup B}^0$ and $\mathbf{\Phi} = \widehat{\mathbf{\Omega}}^{(1)} - \widehat{\mathbf{\Omega}}_{A\cup B}^{(1)}$, where $\|\mathbf{\Phi}\|_1 = \sum_{(i,j)\notin \hat{A}\cup B} |\widehat{\omega}_{ij}^{(1)}| \le (|A^0| - |A|)\tau$. By the Cauchy-Schwarz inequality, the forgoing inequality becomes

$$
\begin{aligned}
& -\log\det(\boldsymbol{I}_{p\times p}+\sqrt{\bar{\boldsymbol{\Sigma}}^0_{A\cup B}}(\widehat{\boldsymbol{\Delta}}+\boldsymbol{\Phi})\sqrt{\bar{\boldsymbol{\Sigma}}^0_{A\cup B}})+\mathrm{tr}(\sqrt{\bar{\boldsymbol{\Sigma}}^0_{A\cup B}}(\widehat{\boldsymbol{\Delta}}+\boldsymbol{\Phi})\sqrt{\bar{\boldsymbol{\Sigma}}^0_{A\cup B}}) \\
\leq\ & \mathrm{tr}((\widehat{\boldsymbol{\Delta}}+\boldsymbol{\Phi})(\bar{\boldsymbol{\Sigma}}^0_{A\cup B}-\boldsymbol{S}))=\mathrm{vec}_A(\widehat{\boldsymbol{\Delta}})^\top\,\mathrm{vec}_A(\bar{\boldsymbol{\Sigma}}^0_{A\cup B}-\boldsymbol{S})+\mathrm{tr}(\boldsymbol{\Phi}(\bar{\boldsymbol{\Sigma}}^0_{A\cup B}-\boldsymbol{S})) \\
=\ & (\bar{\boldsymbol{I}}^{1/2}_{A\cup B,A\cup B}\,\mathrm{vec}_{A\cup B}(\widehat{\boldsymbol{\Delta}}))^\top\bar{\boldsymbol{I}}^{-1/2}_{A\cup B,A\cup B}\,\mathrm{vec}_{A\cup B}(\boldsymbol{\Sigma}^0_{A\cup B}-\boldsymbol{S})+\mathrm{tr}(\boldsymbol{\Phi}(\bar{\boldsymbol{\Sigma}}^0_{A\cup B}-\boldsymbol{S})) \\
\leq\ & \left\|\sqrt{\bar{\boldsymbol{\Sigma}}^0_{A\cup B}}\widehat{\boldsymbol{\Delta}}\sqrt{\bar{\boldsymbol{\Sigma}}^0_{A\cup B}}\right\|_F\left\|\bar{\boldsymbol{I}}^{-1/2}_{A\cup B,A\cup B}\,\mathrm{vec}_{A\cup B}(\boldsymbol{\Sigma}^0_{A\cup B}-\boldsymbol{S})\right\|_2+\tau(|A^0|-|A|)\|\bar{\boldsymbol{\Sigma}}^0_{A\cup B}-\boldsymbol{S}\|_\infty \\
\leq\ & \left\|\sqrt{\bar{\boldsymbol{\Sigma}}^0_{A\cup B}}\widehat{\boldsymbol{\Delta}}\sqrt{\bar{\boldsymbol{\Sigma}}^0_{A\cup B}}\right\|_F\lambda_{\max}(\bar{\boldsymbol{\Omega}}^0_{A\cup B})\sqrt{|A\cup B|}\|\boldsymbol{\Sigma}^0-\boldsymbol{S}\|_\infty \\
& \quad +(2\lambda_{\max}(\boldsymbol{\Sigma}^0)+\lambda_{\max}(\bar{\boldsymbol{\Sigma}}^0_{A\cup B}))\tau K \\
\leq\ & \bar{\lambda}_{\max}\sqrt{|A^0\cup B|}\left\|\sqrt{\bar{\boldsymbol{\Sigma}}^0_{A\cup B}}\widehat{\boldsymbol{\Delta}}\sqrt{\bar{\boldsymbol{\Sigma}}^0_{A\cup B}}\right\|_F\|\boldsymbol{\Sigma}^0-\boldsymbol{S}\|_\infty+3\bar{\lambda}^{-1}_{\min}\tau K
\end{aligned}
\tag{B.19}
$$

on the event $\mathcal{G}$, where $\bar{\boldsymbol{I}}_{A\cup B,A\cup B}=\left[\bar{\boldsymbol{\Sigma}}^0_{A\cup B,A\cup B}\otimes_s\bar{\boldsymbol{\Sigma}}^0_{A\cup B,A\cup B}\right]_{A\cup B,A\cup B}$. On the other hand, by Lemma 4,

$$
\begin{aligned}
& -\log\det(\boldsymbol{I}_{p\times p}+\sqrt{\bar{\boldsymbol{\Sigma}}^0_{A\cup B}}(\widehat{\boldsymbol{\Delta}}+\boldsymbol{\Phi})\sqrt{\bar{\boldsymbol{\Sigma}}^0_{A\cup B}})+\mathrm{tr}(\sqrt{\bar{\boldsymbol{\Sigma}}^0_{A\cup B}}(\widehat{\boldsymbol{\Delta}}+\boldsymbol{\Phi})\sqrt{\bar{\boldsymbol{\Sigma}}^0_{A\cup B}}) \\
\geq\ & \min\left(\frac{\|\sqrt{\bar{\boldsymbol{\Sigma}}^0_{A\cup B}}(\widehat{\boldsymbol{\Delta}}+\boldsymbol{\Phi})\sqrt{\bar{\boldsymbol{\Sigma}}^0_{A\cup B}}\|_F}{8\sqrt{2}},\frac{\|\sqrt{\bar{\boldsymbol{\Sigma}}^0_{A\cup B}}(\widehat{\boldsymbol{\Delta}}+\boldsymbol{\Phi})\sqrt{\bar{\boldsymbol{\Sigma}}^0_{A\cup B}}\|_F^2}{12}\right) \\
\geq\ & \min\left(\frac{\|\sqrt{\bar{\boldsymbol{\Sigma}}^0_{A\cup B}}\widehat{\boldsymbol{\Delta}}\sqrt{\bar{\boldsymbol{\Sigma}}^0_{A\cup B}}\|_F}{8\sqrt{2}},\frac{\|\sqrt{\bar{\boldsymbol{\Sigma}}^0_{A\cup B}}\widehat{\boldsymbol{\Delta}}\sqrt{\bar{\boldsymbol{\Sigma}}^0_{A\cup B}}\|_F^2}{24}\right) \\
& -\max\left(\frac{(|A^0|-|A|)\lambda_{\max}(\bar{\boldsymbol{\Sigma}}^0_{A\cup B})\tau}{8\sqrt{2}},\frac{(|A^0|-|A|)^2\lambda^2_{\max}(\bar{\boldsymbol{\Sigma}}^0_{A\cup B})\tau^2}{12}\right) \\
\geq\ & \min\left(\frac{\|\sqrt{\bar{\boldsymbol{\Sigma}}^0_{A\cup B}}\widehat{\boldsymbol{\Delta}}\sqrt{\bar{\boldsymbol{\Sigma}}^0_{A\cup B}}\|_F}{8\sqrt{2}},\frac{\|\sqrt{\bar{\boldsymbol{\Sigma}}^0_{A\cup B}}\widehat{\boldsymbol{\Delta}}\sqrt{\bar{\boldsymbol{\Sigma}}^0_{A\cup B}}\|_F^2}{24}\right)-\frac{\lambda_{\max}(\bar{\boldsymbol{\Sigma}}^0_{A\cup B})K\tau}{8}
\end{aligned}
$$

where the last two inequalities use that $\|\boldsymbol{M}_1+\boldsymbol{M}_2\|_F^2\geq 2^{-1}\|\boldsymbol{M}_1\|_F^2-\|\boldsymbol{M}_2\|_F^2$, $\|\sqrt{\bar{\boldsymbol{\Sigma}}^0_{A\cup B}}\boldsymbol{\Phi}\sqrt{\bar{\boldsymbol{\Sigma}}^0_{A\cup B}}\|_F^2\leq \lambda^2_{\max}(\bar{\boldsymbol{\Sigma}}^0_{A\cup B})\|\boldsymbol{\Phi}\|_F^2\leq\lambda^2_{\max}(\bar{\boldsymbol{\Sigma}}^0_{A\cup B})\|\boldsymbol{\Phi}\|_1^2\leq\lambda^2_{\max}(\bar{\boldsymbol{\Sigma}}^0_{A\cup B})(|A^0|-|A|)^2\tau^2$, and $\min(a-b,c-d)\geq$

$\min(a,c)-\max(b,d)$. Combining this with (B.19), we obtain

$$\begin{aligned}
&\bar{\lambda}_{\max}\sqrt{|A^0\cup B|}\left\|\sqrt{\bar{\boldsymbol{\Sigma}}^0_{A\cup B}}\widehat{\boldsymbol{\Delta}}\sqrt{\bar{\boldsymbol{\Sigma}}^0_{A\cup B}}\right\|_F\|\boldsymbol{\Sigma}^0-\boldsymbol{S}\|_\infty+4\bar{\lambda}_{\min}^{-1}\tau K\\
\geq\ &\min\left(\frac{\|\sqrt{\bar{\boldsymbol{\Sigma}}^0_{A\cup B}}\widehat{\boldsymbol{\Delta}}\sqrt{\bar{\boldsymbol{\Sigma}}^0_{A\cup B}}\|_F}{8\sqrt{2}},\frac{\|\sqrt{\bar{\boldsymbol{\Sigma}}^0_{A\cup B}}\widehat{\boldsymbol{\Delta}}\sqrt{\bar{\boldsymbol{\Sigma}}^0_{A\cup B}}\|_F^2}{24}\right),
\end{aligned}$$

which implies that

$$\left\|\sqrt{\bar{\boldsymbol{\Sigma}}^0_{A\cup B}}\widehat{\boldsymbol{\Delta}}\sqrt{\bar{\boldsymbol{\Sigma}}^0_{A\cup B}}\right\|_F\leq 24\bar{\lambda}_{\max}\sqrt{|A^0\cup B|}\|\boldsymbol{S}-\boldsymbol{\Sigma}^0\|_\infty+4\sqrt{6\bar{\lambda}_{\min}^{-1}\tau K},$$

on the event $\{\hat{A}=A\}\cap\mathcal{G}$. Next, note that

$$\begin{aligned}
&\frac{2}{n}\left(L_n(\widehat{\boldsymbol{\Omega}}^{(1)})-L_n(\boldsymbol{\Omega}^0)\right)+2\left(L(\boldsymbol{\Omega}^0)-L(\bar{\boldsymbol{\Omega}}^0_{A\cup B})\right)\\
=\ &\frac{2}{n}\left(L_n(\widehat{\boldsymbol{\Omega}}^{(1)})-L_n(\bar{\boldsymbol{\Omega}}^0_{A\cup B})\right)+\mathrm{tr}\left((\boldsymbol{\Omega}^0-\bar{\boldsymbol{\Omega}}^0_{A\cup B})(\boldsymbol{S}-\boldsymbol{\Sigma}^0)\right)\\
=\ &2\left(L(\widehat{\boldsymbol{\Omega}}^{(1)})-L(\bar{\boldsymbol{\Omega}}^0_{A\cup B})\right)+\mathrm{tr}((\boldsymbol{S}-\bar{\boldsymbol{\Sigma}}^0_{A\cup B})(\widehat{\boldsymbol{\Omega}}^{(1)}-\bar{\boldsymbol{\Omega}}^0_{A\cup B}))+\mathrm{tr}\left((\boldsymbol{\Omega}^0-\bar{\boldsymbol{\Omega}}^0_{A\cup B})(\boldsymbol{S}-\boldsymbol{\Sigma}^0)\right)\\
\leq\ &\mathrm{tr}((\boldsymbol{S}-\bar{\boldsymbol{\Sigma}}^0_{A\cup B})(\widehat{\boldsymbol{\Omega}}^{(1)}-\widehat{\boldsymbol{\Omega}}^{(1)}_{A\cup B})+\mathrm{tr}((\boldsymbol{S}-\boldsymbol{\Sigma}^0)(\widehat{\boldsymbol{\Omega}}^{(1)}_{A\cup B}-\bar{\boldsymbol{\Omega}}^0_{A\cup B}))\\
&+\mathrm{tr}\left((\boldsymbol{\Omega}^0-\bar{\boldsymbol{\Omega}}^0_{A\cup B})(\boldsymbol{S}-\boldsymbol{\Sigma}^0)\right) \qquad \text{(B.20)}
\end{aligned}$$

For the first two terms, using $\tau\leq\frac{\bar{\lambda}_{\min}\min(\sqrt{\tilde{C}_{\min}},\tilde{C}_{\min}^2)}{12|A^0|}$ and $\|\boldsymbol{S}-\boldsymbol{\Sigma}^0\|_\infty\leq\sqrt{\frac{\tilde{C}_{\min}}{48\bar{\lambda}_{\max}^2(|A^0|+|B|)}}$, we have that on the event $\mathcal{G}$

$$\begin{aligned}
&\mathrm{tr}((\boldsymbol{S}-\boldsymbol{\Sigma}^0)(\widehat{\boldsymbol{\Omega}}^{(1)}_{A\cup B}-\bar{\boldsymbol{\Omega}}^0_{A\cup B}))+\mathrm{tr}((\boldsymbol{S}-\bar{\boldsymbol{\Sigma}}^0_{A\cup B})(\widehat{\boldsymbol{\Omega}}^{(1)}-\widehat{\boldsymbol{\Omega}}^{(1)}_{A\cup B})\\
\leq\ &\left\|\sqrt{\bar{\boldsymbol{\Sigma}}^0_{A\cup B}}\widehat{\boldsymbol{\Delta}}\sqrt{\bar{\boldsymbol{\Sigma}}^0_{A\cup B}}\right\|_F\left\|\bar{\boldsymbol{I}}^{-1/2}_{A\cup B,A\cup B}\,\mathrm{vec}_{A\cup B}(\boldsymbol{S}-\boldsymbol{\Sigma}^0)\right\|_2+\tau K\|\boldsymbol{S}-\bar{\boldsymbol{\Sigma}}^0_{A\cup B}\|_\infty\\
\leq\ &24\min\left(\bar{\lambda}_{\max}^2|A^0\cup B|\|\boldsymbol{S}-\boldsymbol{\Sigma}^0\|_\infty^2,\frac{\bar{\lambda}_{\max}\sqrt{|A^0\cup B|}\|\boldsymbol{S}-\boldsymbol{\Sigma}^0\|_\infty}{16\sqrt{2}}\right)\\
&+\frac{\sqrt{3\bar{\lambda}_{\min}^{-1}\tau K}}{4}+3\bar{\lambda}_{\min}^{-1}\tau K\\
\leq\ &2^{-1}K(\boldsymbol{\Omega}^0,\bar{\boldsymbol{\Omega}}^0_{A\cup B})+2^{-1}K(\boldsymbol{\Omega}^0,\bar{\boldsymbol{\Omega}}^0_{A\cup B}))=L(\boldsymbol{\Omega}^0)-L(\bar{\boldsymbol{\Omega}}^0_{A\cup B}),
\end{aligned}$$

which, together with (B.20), implies that for any $A \neq A^0, |A| \leq K, A \cap B = \emptyset$, we have that

$$\Big\{L_n(\widehat{\mathbf{\Omega}}^{(1)}) - L_n(\mathbf{\Omega}^0) \geq 0; \hat{A} = A; \mathcal{G}\Big\} \subseteq \big\{\mathrm{tr}\left((\mathbf{\Omega}^0 - \bar{\mathbf{\Omega}}^0_{A\cup B})(\boldsymbol{S} - \mathbf{\Sigma}^0)\right) \geq L(\mathbf{\Omega}^0) - L(\bar{\mathbf{\Omega}}^0_{A\cup B})\big\}$$

Hence,

$$\begin{aligned}
\mathbb{P}\left(\widehat{\mathbf{\Omega}}^{(1)} \neq \widehat{\mathbf{\Omega}}_{A^0\cup B}\right) &\leq \sum_{A:A\neq A^0,|A|\leq K, A\cap B=\emptyset} \mathbb{P}\left(L_n(\widehat{\mathbf{\Omega}}^{(1)}) - L_n(\mathbf{\Omega}^0) \geq 0; \hat{A} = A; \mathcal{G}\right) + \mathbb{P}(\mathcal{G}^c) \\
&\leq \sum_{A:A\neq A^0,|A|\leq K, A\cap B=\emptyset} \mathbb{P}\left(\mathrm{tr}\left((\mathbf{\Omega}^0 - \bar{\mathbf{\Omega}}^0_{A\cup B})(\boldsymbol{S} - \mathbf{\Sigma}^0)\right) \geq L(\mathbf{\Omega}^0) - L(\bar{\mathbf{\Omega}}^0_{A\cup B})\right) + \mathbb{P}(\mathcal{G}^c)\,,
\end{aligned}$$

where the first probability can be further bounded by applying Lemmas 3 and 4.

$$\begin{aligned}
&\sum_{A:A\neq A^0,|A|\leq K, A\cap B=\emptyset} \mathbb{P}\left(\mathrm{tr}\left((\mathbf{\Omega}^0 - \bar{\mathbf{\Omega}}^0_{A\cup B})(\boldsymbol{S} - \mathbf{\Sigma}^0)\right) \geq L(\mathbf{\Omega}^0) - L(\bar{\mathbf{\Omega}}^0_{A\cup B})\right) \\
\leq\ & \sum_{A:A\neq A^0,|A|\leq K, A\cap B=\emptyset} \exp\left(\frac{-n10^{-1}K^2(\mathbf{\Omega}^0, \bar{\mathbf{\Omega}}^0_{A\cup B})}{\|\bar{\mathbf{\Omega}}^0_{A\cup B} - \mathbf{\Omega}^0\|^2 + K(\mathbf{\Omega}^0, \bar{\mathbf{\Omega}}^0_{A\cup B})\|\bar{\mathbf{\Omega}}^0_{A\cup B} - \mathbf{\Omega}^0\|}\right) \\
\leq\ & \sum_{A:A\neq A^0,|A|\leq K, A\cap B=\emptyset} \exp\left(\frac{-n\min\left(128^{-1}, K(\mathbf{\Omega}^0, \bar{\mathbf{\Omega}}^0_{A\cup B})\right)}{20}\right) \\
\leq\ & \sum_{A:A\neq A^0,|A|\leq K, A\cap B=\emptyset, K(\mathbf{\Omega}^0, \bar{\mathbf{\Omega}}^0_{A\cup B})\leq 1} \exp\left(\frac{-nK(\mathbf{\Omega}^0, \bar{\mathbf{\Omega}}^0_{A\cup B})}{2560}\right) \\
&+ \sum_{A:A\neq A^0,|A|\leq K, A\cap B=\emptyset, K(\mathbf{\Omega}^0, \bar{\mathbf{\Omega}}^0_{A\cup B})> 1} \exp\left(\frac{-n}{2560}\right) \\
\leq\ & \sum_{j=1}^{|A^0|}\sum_{i=1}^{|A^0|-j} \binom{|A^0|}{j}\binom{p-|A^0|}{i} \exp\left(\frac{-nj\tilde{C}_{\min}}{2560}\right) + \exp\left(\frac{-n}{2560} + |A^0|\log p\right) \\
\leq\ & \sum_{j=1}^{|A^0|} \exp\left(\frac{-nj\tilde{C}_{\min}}{2560} + 2j\log p\right) + \exp\left(\frac{-n}{2560} + |A^0|\log p\right) \\
\leq\ & 2\exp\left(\frac{-n\tilde{C}_{\min}}{2560} + 2\log p\right) + \exp\left(\frac{-n}{2560} + |A^0|\log p\right) \longrightarrow 0
\end{aligned}$$

as $n \to \infty$, provided that $\frac{|A^0|\log p}{n} \leq 3000^{-1}$ and $\tilde{C}_{\min} \geq 3000\frac{\log p}{n}$.

To bound $\mathbb{P}(\mathcal{G}^c)$, we apply Lemma 3 with $\nu = \min\left(\frac{1}{16\sqrt{2}\bar{\lambda}_{\max}\sqrt{|A^0|+|B|}}, \sqrt{\frac{\tilde{C}_{\min}}{48\bar{\lambda}_{\max}^2|A^0\cup B|}}, \lambda_{\max}(\mathbf{\Sigma}^0)\right)$ and get

$$
\begin{aligned}
\mathbb{P}(\mathcal{G}^c) &\leq \mathbb{P}\left(\|\boldsymbol{S}-\mathbf{\Sigma}^0\|_\infty \geq \nu\right) \leq 2\exp\left(-n\frac{\nu^2}{9\lambda_{\max}^2(\mathbf{\Sigma}^0)+8\nu\lambda_{\max}(\mathbf{\Sigma}^0)}+2\log p\right) \\
&\leq 2\exp\left(-n\frac{\nu^2}{18\lambda_{\max}^2(\mathbf{\Sigma}^0)}+2\log p\right) \longrightarrow 0\,,
\end{aligned}
$$

provided that $\tilde{C}_{\min} \geq 2000\frac{\bar{\lambda}_{\max}^2}{\lambda_{\min}^2(\mathbf{\Omega}^0)}\frac{(|A^0|+|B|)\log p}{n}$ and $\frac{\bar{\lambda}_{\max}^2}{\lambda_{\min}^2(\mathbf{\Omega}^0)}\frac{(|A^0|+|B|)\log p}{n} \leq 18000$. Combining, we obtain

$$
\begin{aligned}
P\left(\widehat{\mathbf{\Omega}}^{(1)} \neq \widehat{\mathbf{\Omega}}_{A^0\cup B}\right) &\leq \exp\left(\frac{-n\tilde{C}_{\min}}{2560}+2\log p\right)+\exp\left(\frac{-n}{2560}+|A^0|\log p\right) \\
&+\exp\left(-n\frac{\min\left(\sqrt{\frac{\min\left(\tilde{C}_{\min},3/32\right)}{48\bar{\lambda}_{\max}^2(|A^0|+|B|)}}, \lambda_{\max}(\mathbf{\Sigma}^0)\right)^2}{18\lambda_{\max}^2(\mathbf{\Sigma}^0)}+2\log p\right)
\end{aligned}
$$

For $\mathbb{P}\left(\widehat{\mathbf{\Omega}}^{(0)} \neq \widehat{\mathbf{\Omega}}_{A^0}\right)$, we let $B=\emptyset$ and a similar bound can be established. Moreover, by Lemma 4, it is easy to see that $\max(K(\mathbf{\Omega}^0,\mathbf{\Omega}), K^2(\mathbf{\Omega}^0,\mathbf{\Omega})) \geq \frac{\|\mathbf{\Omega}^0-\mathbf{\Omega}\|^2}{512}$ for any $\mathbf{\Omega}$. Consequently, $\tilde{C}_{\min} \geq \frac{C_{\min}}{512}$. Thus, the bound in (B.17) is established. This completes the proof.

Lemma 7 *Let* $\mathbf{\Gamma}_k = (\gamma_{k1}, \cdots \gamma_{km}) \in \mathbb{R}^m$*;* $k=1,\cdots,n$ *be iid random vectors with* $\mathrm{Var}(\boldsymbol{\gamma}_1) = \boldsymbol{I}_{m\times m}$*. If* m *is fixed, then*

$$
n^{-1}\|\sum_{k=1}^n \boldsymbol{\gamma}_k\|_2^2 \xrightarrow{d} \chi_m^2, \;\; \textit{as } n\to\infty\,. \tag{B.21}
$$

Otherwise, if $\max\left(m, m_2m/n, m_3/n, m_3m^{3/2}/n^2\right) \to 0$*, where* $m_j = \max_{1\leq i\leq m}\mathbb{E}\gamma_{1i}^{2j}$*;* $j=2,3,$ *then*

$$
\frac{\|\sum_{k=1}^n \boldsymbol{\gamma}_k\|_2^2 - nm}{n\sqrt{2m}} \xrightarrow{d} N(0,1), \;\; \textit{as } n\to\infty\,. \tag{B.22}
$$

Proof of Lemma 7: If m is fixed, then (B.21) follows from the central limit theorem and the continuous mapping theorem.

For (B.22), let $\boldsymbol{\Gamma}_k = \sum_{j=1}^k \boldsymbol{\gamma}_j$; $k = 1, \cdots, n$ be a partial sum of k iid m-dimensional vectors $\boldsymbol{\gamma}_j$'s. Next we apply Theorem 18.1 of [1] to show that $\frac{\|\boldsymbol{\Gamma}_n\|_2^2 - nm}{n\sqrt{2m}} \to N(0,1)$ for triangular arrays of martingale differences $\{\eta_{n,k} = \frac{\|\boldsymbol{\Gamma}_k\|_2^2 - \|\boldsymbol{\Gamma}_{k-1}\|_2^2 - m}{n\sqrt{2m}} = \frac{\|\boldsymbol{\gamma}_k\|_2^2 - m + 2\boldsymbol{\gamma}_k^\top \boldsymbol{\Gamma}_{k-1}}{n\sqrt{2m}}\}$. Towards this end, we verify that

$$\sum_{k=1}^n \mathbb{E}\big(\eta_{n,k}^2 \,|\, \boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_{k-1}\big) \xrightarrow{P} 1, \quad \sum_{k=1}^n \mathbb{E}|\eta_{n,k}|^3 \to 0. \tag{B.23}$$

For the first condition of (B.23), we compute $\mathbb{E}$ and Var of $\mathbb{E}\big(\eta_{n,k}^2 \,|\, \boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_{k-1}\big)$. Note that $\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_m$ are iid vectors with $\text{Var}(\boldsymbol{\gamma}_m) = \boldsymbol{I}_{m\times m}$, $\mathbb{E}\boldsymbol{\Gamma}_{k-1} = 0$, and $\mathbb{E}\|\boldsymbol{\Gamma}_{k-1}\|_2^2 = (k-1)m$. Then, for each $k = 1, \cdots, n$, $\mathbb{E}\mathbb{E}\big(\eta_{n,k}^2 \,|\, \boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_{k-1}\big)$ becomes

$$\begin{aligned}
&(2mn^2)^{-1}\Big(\mathbb{E}\big(\|\boldsymbol{\gamma}_k\|_2^2 - m\big)^2 + 4\mathbb{E}\big((\|\boldsymbol{\gamma}_k\|_2^2 - m)\boldsymbol{\gamma}_k\big)^\top \mathbb{E}\boldsymbol{\Gamma}_{k-1} + 4\mathbb{E}\mathbb{E}\big((\boldsymbol{\gamma}_k^\top \boldsymbol{\Gamma}_{k-1})^2 \,|\, \boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_{k-1}\big)\Big) \\
&= (2mn^2)^{-1}\Big(\text{Var}(\|\boldsymbol{\gamma}_k\|_2^2) + 4\mathbb{E}\|\boldsymbol{\Gamma}_{k-1}\|_2^2\Big) = (2mn^2)^{-1}\Big(\text{Var}(\|\boldsymbol{\gamma}_k\|_2^2) + 4(k-1)m\Big),
\end{aligned}$$

which, after summing over $k = 1, \cdots, n$, leads to

$$\sum_{k=1}^n \frac{2(k-1)}{n^2} \le \mathbb{E}\Big(\sum_{k=1}^n \mathbb{E}\big(\eta_{n,k}^2 \,|\, \boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_{k-1}\big)\Big) \le \frac{mm_2}{2n} + \sum_{k=1}^n \frac{2(k-1)}{n^2},$$

where $\text{Var}(\|\boldsymbol{\gamma}_k\|_2) \le m^2 m_2$; $k = 1, \cdots, n$. Consequently, $\Big|\mathbb{E}\Big(\sum_{k=1}^n \mathbb{E}\big(\eta_{n,k}^2 \,|\, \boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_{k-1}\big)\Big) - 1\Big| \le \frac{2}{n} + \frac{mm_2}{2n}$. Let $\boldsymbol{a} = \mathbb{E}\big((\|\boldsymbol{\gamma}_1\|_2^2 - m)\boldsymbol{\gamma}_1\big)$. Similarly, using an inequality $(a_1 + a_2 + a_3)^2 \le$

$3(a_1^2 + a_2^2 + a_3^2)$ for real numbers a_j; $j = 1, \cdots, 3$.

$$
\begin{aligned}
&\mathrm{Var}\Big(\sum_{k=1}^{n} \mathbb{E}\big(\eta_{n,k}^2 \,|\, \boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_{k-1}\big)\Big) = \frac{4}{m^2 n^4}\,\mathrm{Var}\left(\sum_{k=1}^{n} \big(\boldsymbol{a}^\top \boldsymbol{\Gamma}_{k-1} + \|\boldsymbol{\Gamma}_{k-1}\|_2^2\big)\right) \\
&= \frac{4}{m^2 n^4}\,\mathrm{Var}\left(\sum_{k=1}^{n} (n-k)\Big(\boldsymbol{a}^\top \boldsymbol{\gamma}_k + \|\boldsymbol{\gamma}_k\|_2^2\Big) + 2\sum_{k<k'} (n-(k\vee k'))\boldsymbol{\gamma}_k^\top \boldsymbol{\gamma}_{k'}\right) \\
&\le \frac{12}{m^2 n^4}\left[\mathrm{Var}\left(\sum_{k=1}^{n} (n-k)\boldsymbol{a}^\top \boldsymbol{\gamma}_k\right) + \mathrm{Var}\left(\sum_{k=1}^{n} (n-k)\|\boldsymbol{\gamma}_k\|_2^2\right)\right. \\
&\quad \left. + \mathrm{Var}\left(\sum_{k<k'} (n-(k\vee k'))\boldsymbol{\gamma}_k^\top \boldsymbol{\gamma}_{k'}\right)\right] \equiv \frac{12}{m^2 n^4}\Big[T_1 + T_2 + T_3\Big]. \qquad \text{(B.24)}
\end{aligned}
$$

For T_1, note that $\|\boldsymbol{a}\|_2^2 \le \sum_{k=1}^m \mathbb{E}^2\big((\|\boldsymbol{\gamma}_1\|_2^2 - m)\gamma_{1k}\big) \le \sum_{k=1}^m \mathbb{E}\big((\|\boldsymbol{\gamma}_1\|_2^2 - m)^2\big)\mathbb{E}\gamma_{1k}^2 \le m^3 m_2$. Then

$$
\begin{aligned}
\mathrm{Var}\left(\sum_{k=1}^{n} (n-k)\boldsymbol{a}^\top \boldsymbol{\gamma}_k\right) &= \sum_{k=1}^{n} (n-k)^2 \mathbb{E}\left(\boldsymbol{a}^\top \boldsymbol{\gamma}_k\right)^2 = \sum_{k=1}^{n} (n-k)^2 \sum_{j=1}^{m} a_j^2 \mathbb{E}\gamma_{kj}^2 \\
&= \frac{\|\boldsymbol{a}\|_2^2}{6}(n-1)n(2n-1) \le n^3 m^3 m_2\,.
\end{aligned}
$$

For T_2, note that $\mathrm{Var}\,(\sum_{k=1}^n (n-k)\|\boldsymbol{\gamma}_k\|_2^2) \le \sum_{k=1}^n (n-k)^2 m^2 m_2 = \frac{1}{6}(n-1)n(2n-1)m^2 m_2$. To bound T_3, note that, for $k \neq k'$ and $j \neq j'$, $\mathbb{E}\big(\boldsymbol{\gamma}_k^\top \boldsymbol{\gamma}_{k'} \boldsymbol{\gamma}_j^\top \boldsymbol{\gamma}_{j'}\big) = \mathbb{I}(\{j, j'\} = \{k, k'\})\mathbb{E}\left(\boldsymbol{\gamma}_k^\top \boldsymbol{\gamma}_{k'}\right)^2 = \mathbb{I}(\{j, j'\} = \{k, k'\})m$, yielding that

$$
\mathrm{Var}\left(\sum_{k<k'} (n-(k\vee k'))\boldsymbol{\gamma}_k^\top \boldsymbol{\gamma}_{k'}\right) = \sum_{k<k'} (n-(k\vee k'))^2 \mathbb{E}\left(\boldsymbol{\gamma}_k^\top \boldsymbol{\gamma}_{k'}\right)^2 \le n^4 m\,.
$$

Combining (B.24) with the bounds of $T_1 - T_3$, we obtain

$$
\mathrm{Var}\Big(\sum_{k=1}^{n} \mathbb{E}\big(\eta_{n,k}^2 \,|\, \boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_{k-1}\big)\Big) \le \frac{12\,(n^3 m^3 m_2 + n^3 m^2 m_2 + n^4 m)}{m^2 n^4}.
$$

Hence the first condition of (B.23) is implied by the assumption that $m m_2/n \to 0$ and

$m \to \infty$.

For the second condition of (B.23), note that $\mathbb{E}|\eta_{n,k}|^3 = \mathbb{E}\left(\left|\|\boldsymbol{\gamma}_k\|_2^2 - m + 2\boldsymbol{\gamma}_k^\top \boldsymbol{\Gamma}_{k-1}\right|^3\right)$ is bounded by

$$\begin{aligned} & 4\mathbb{E}\left(\left|\|\boldsymbol{\gamma}_k\|_2^2 - m\right|^3\right) + 16\mathbb{E}\left(\left|\boldsymbol{\gamma}_k^\top \boldsymbol{\Gamma}_{k-1}\right|^3\right) \le \mathbb{E}\left(\|\boldsymbol{\gamma}_k\|_2^6\right) + \sqrt{\mathbb{E}\big((\boldsymbol{\gamma}_k^\top \boldsymbol{\Gamma}_{k-1})^6\big)} \\ \le\ & m^3 m_3 + \sqrt{(k-1)^3 m^3 m_3 + (k-1)^2 m^3 m_2 m_3 + (k-1) m^3 m_3^2} \\ \le\ & m^3 m_3 + k^{3/2} m^{3/2} m_3^{1/2} + k m^{3/2} m_2^{1/2} m_3^{1/2} + k^{1/2} m^{3/2} m_3 . \end{aligned}$$

Summing over k, $\frac{\sum_{k=1}^n \mathbb{E}\left(\left|\|\boldsymbol{\gamma}_k\|_2^2 - m + 2\boldsymbol{\gamma}_k^\top \boldsymbol{\Gamma}_{k-1}\right|^3\right)}{n^3 m^{3/2}}$ is upper bounded by

$$\begin{aligned} & \frac{\left(n m^3 m_3 + n^{5/2} m^{3/2} m_3^{1/2} + n^2 m^{3/2} m_2^{1/2} m_3^{1/2} + n^{3/2} m^{3/2} m_3\right)}{n^3 m^{3/2}} \\ = & \frac{m^{3/2} m_3}{n^2} + \frac{m_3^{1/2}}{n^{1/2}} + \frac{m_2^{1/2} m_3^{1/2}}{n} + \frac{m_3}{n^{3/2}} \to 0 , \end{aligned}$$

provided that $\max\left(m_2 m/n, m_3/n, m_3 m^{3/2}/n^2\right) \to 0$. Thus the second condition in (B.23) is met. As a consequence of Theorem 18.1 of [1], the desired asymptotic normality is established. This completes the proof.

Lemma 8 *Let* $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma}^0)$ *and* $\gamma = \operatorname{tr}(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{\Sigma}^0)\boldsymbol{T})$ *with* $\boldsymbol{T}$ *a symmetric matrix. Then*

$$\mathbb{E}(\gamma^{2m}) \le (2m-1)!\, 2^{m-1} \left(\mathbb{E}(\gamma^2)\right)^m \ \textit{for any integer } m \ge 1 . \tag{B.25}$$

Proof of Lemma 8: As in (B.8) and (B.10), we expand the moment generating function of γ: $M_\gamma(\lambda) = \mathbb{E}\exp(\lambda\gamma) = \lambda^2 \|\sqrt{\boldsymbol{\Sigma}^0}\boldsymbol{T}\sqrt{\boldsymbol{\Sigma}^0}\|_F^2 + (1/2)\sum_{l=3}^\infty l^{-1}\lambda^l \operatorname{tr}\left[(2\boldsymbol{T}\boldsymbol{\Sigma}^0)^l\right]$ for any $|\lambda| < \|\sqrt{\boldsymbol{\Sigma}^0}\boldsymbol{T}\sqrt{\boldsymbol{\Sigma}^0}\|_F/2$. Direct computation of high-order derivatives of $M_\gamma(\lambda)$ in λ yields that $\mathbb{E}(\gamma^{2m}) = (2m-1)!\, 2^{2m-1} \operatorname{tr}\left((\boldsymbol{T}\boldsymbol{\Sigma}^0)^{2m}\right)$ for any integer $m \ge 1$. An application of $\operatorname{tr}\left((\boldsymbol{T}\boldsymbol{\Sigma}^0)^{2m}\right) \le \|\sqrt{\boldsymbol{\Sigma}^0}\boldsymbol{T}\sqrt{\boldsymbol{\Sigma}^0}\|_F^{2m}$ yields that $\mathbb{E}(\gamma^{2m}) \le (2m-1)!\, 2^{2m-1}\|\sqrt{\boldsymbol{\Sigma}^0}\boldsymbol{T}\sqrt{\boldsymbol{\Sigma}^0}\|_F^{2m} = (2m-1)!\, 2^{m-1}\left(\mathbb{E}(\gamma^2)\right)^m$. This completes the proof.

Proof of Lemma 9: Let $\widehat{\boldsymbol{\Delta}}_{\tilde{A}} = \widehat{\boldsymbol{\Omega}}_{\tilde{A}} - \boldsymbol{\Omega}^0$ for any $\tilde{A} \supseteq A^0$. Applying Lemma 5 to $\widehat{\boldsymbol{\Delta}}_{\tilde{A}}$ and $\widehat{\boldsymbol{\Delta}}_{A^0}$, we have that both $\|\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\|$ and $\|\widehat{\boldsymbol{\Delta}}_{A^0}\|$ tend to zero in probability as n goes to infinity. Hence, we could assume throughout the proof that $\max\left(\|\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\|, \|\widehat{\boldsymbol{\Delta}}_{A^0}\|\right) \leq 1/2$ holds with probability tending to one. Note that $\boldsymbol{\Omega}^0 = (\boldsymbol{\Sigma}^0)^{-1}$, and $\log\det(\widehat{\boldsymbol{\Omega}}_{\tilde{A}}) = \log\det(\boldsymbol{I}_{p\times p} + \widehat{\boldsymbol{\Delta}}_{\tilde{A}}\boldsymbol{\Sigma}^0) + \log\det(\boldsymbol{\Omega}^0)$. Then

$$\begin{aligned}
& \log\det(\boldsymbol{I}_{p\times p} + \widehat{\boldsymbol{\Delta}}_{\tilde{A}}\boldsymbol{\Sigma}^0) \\
= \; & \log\det(\boldsymbol{I}_{p\times p} + [\boldsymbol{\Sigma}^0]^{1/2}\widehat{\boldsymbol{\Delta}}_{\tilde{A}}[\boldsymbol{\Sigma}^0]^{1/2}) = \mathrm{tr}(\log(\boldsymbol{I}_{p\times p} + [\boldsymbol{\Sigma}^0]^{1/2}\widehat{\boldsymbol{\Delta}}_{\tilde{A}}[\boldsymbol{\Sigma}^0]^{1/2})) \\
= \; & \mathrm{tr}\left(\sum_{i=1}^{\infty}(-1)^{i+1}\frac{\left([\boldsymbol{\Sigma}^0]^{1/2}\widehat{\boldsymbol{\Delta}}_{\tilde{A}}[\boldsymbol{\Sigma}^0]^{1/2}\right)^i}{i}\right), \\
= \; & \mathrm{tr}\left(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\boldsymbol{\Sigma}^0\right) - \frac{1}{2}\mathrm{tr}\left(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\boldsymbol{\Sigma}^0\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\boldsymbol{\Sigma}^0\right) + R_1(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}),
\end{aligned} \tag{B.26}$$

where $R_1(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}) = \sum_{i=3}^{\infty}\frac{(-1)^{i+1}}{i}\mathrm{tr}\left(\left(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\boldsymbol{\Sigma}^0\right)^i\right)$ and the expansion is valid since $\|\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\| \leq 1/2 < 1$. As a result,

$$\begin{aligned}
& n^{-1}\left(L_n(\widehat{\boldsymbol{\Omega}}_{\tilde{A}}) - L_n(\boldsymbol{\Omega}^0)\right) \\
= \; & \frac{1}{2}\mathrm{tr}\left(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\boldsymbol{\Sigma}^0\right) - \frac{1}{4}\mathrm{tr}\left(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\boldsymbol{\Sigma}^0\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\boldsymbol{\Sigma}^0\right) - \frac{1}{2}\mathrm{tr}(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\boldsymbol{S}) + \frac{1}{2}R_1(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}) \\
= \; & \frac{1}{2}\mathrm{tr}\left(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}(\boldsymbol{\Sigma}^0 - \boldsymbol{S})\right) - \frac{1}{4}\|\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\|^2 + \frac{1}{2}R_1(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}).
\end{aligned} \tag{B.27}$$

Moreover, using the property of the CMLE, $\widehat{\boldsymbol{\Delta}}_{\tilde{A}}$ satisfies a score equation: $[-(\widehat{\boldsymbol{\Delta}}_{\tilde{A}} + \boldsymbol{\Omega}^0)^{-1} + \boldsymbol{S}]_{\tilde{A}} = 0$. This, in turn, yields that

$$\left[\boldsymbol{\Sigma}^0\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\boldsymbol{\Sigma}^0\right]_{\tilde{A}} = \left[R_2(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}) + \boldsymbol{\Sigma}^0 - \boldsymbol{S}\right]_{\tilde{A}}, \tag{B.28}$$

where $(\widehat{\boldsymbol{\Delta}}_{\tilde{A}} + \boldsymbol{\Omega}^0)^{-1} = \boldsymbol{\Sigma}^0 - \boldsymbol{\Sigma}^0\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\boldsymbol{\Sigma}^0 + R_2(\widehat{\boldsymbol{\Delta}}_{\tilde{A}})$ is used, and $R_2(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}) = \boldsymbol{\Sigma}^0\sum_{i=2}^{\infty}(-1)^i\left(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\boldsymbol{\Sigma}^0\right)^i$. By the definition of $\otimes$ and (B.2), (B.28) can be rewritten in a vector form as

$$2\boldsymbol{I}_{\tilde{A},\tilde{A}}\,\mathrm{vec}_{\tilde{A}}(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}) = \mathrm{vec}\left(R_2(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}) + \boldsymbol{\Sigma}^0 - \boldsymbol{S}\right). \tag{B.29}$$

Moreover, after taking the inner product with $\widehat{\boldsymbol{\Delta}}_{\tilde{A}}$ for both sides of (B.28), we obtain

$$\mathrm{tr}\left(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\boldsymbol{\Sigma}^0\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\boldsymbol{\Sigma}^0\right) = \mathrm{tr}\left(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}R_2(\widehat{\boldsymbol{\Delta}}_{\tilde{A}})\right) + \mathrm{tr}\left(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}(\boldsymbol{\Lambda})\right), \tag{B.30}$$

where $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^0 - \boldsymbol{S}$. Hence, combining (B.29) and (B.30) with (B.27) yields that

$$2n^{-1}\left(L_n(\widehat{\boldsymbol{\Omega}}_{\tilde{A}}) - L_n(\boldsymbol{\Omega}^0)\right) = \frac{1}{2}\operatorname{tr}\left(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\boldsymbol{\Lambda}\right) - \frac{1}{2}\operatorname{tr}\left(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}R_2(\widehat{\boldsymbol{\Delta}}_{\tilde{A}})\right) + R_1(\widehat{\boldsymbol{\Delta}}_{\tilde{A}})$$
$$= \frac{1}{2}\left(\operatorname{vec}_{\tilde{A}}(\widehat{\boldsymbol{\Delta}})\right)^\top \operatorname{vec}_{\tilde{A}}\left(\boldsymbol{\Lambda} - R_2(\widehat{\boldsymbol{\Delta}}_{\tilde{A}})\right) + R_1(\widehat{\boldsymbol{\Delta}}_{\tilde{A}})$$
$$= \frac{1}{4}\operatorname{vec}_{\tilde{A}}\left(\boldsymbol{\Lambda} + R_2(\widehat{\boldsymbol{\Delta}}_{\tilde{A}})\right)^\top \boldsymbol{I}^{-1}_{\tilde{A},\tilde{A}}\operatorname{vec}_{\tilde{A}}\left(\boldsymbol{\Lambda} - R_2(\widehat{\boldsymbol{\Delta}}_{\tilde{A}})\right) + R_1(\widehat{\boldsymbol{\Delta}}_{\tilde{A}})$$
$$= \frac{1}{4}\operatorname{vec}_{\tilde{A}}(\boldsymbol{\Lambda})^\top \boldsymbol{I}^{-1}_{\tilde{A},\tilde{A}}\operatorname{vec}_{\tilde{A}}(\boldsymbol{\Lambda}) - \frac{1}{4}\operatorname{vec}_{\tilde{A}}\left(R_2(\widehat{\boldsymbol{\Delta}}_{\tilde{A}})\right)^\top \boldsymbol{I}^{-1}_{\tilde{A},\tilde{A}}\operatorname{vec}_{\tilde{A}}\left(R_2(\widehat{\boldsymbol{\Delta}}_{\tilde{A}})\right) + R_1(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}).$$

Similarly,

$$2n^{-1}\left(L_n(\widehat{\boldsymbol{\Omega}}_{A^0}) - L_n(\boldsymbol{\Omega}^0)\right)$$
$$= \frac{1}{4}\operatorname{vec}_{A^0}(\boldsymbol{\Lambda})^\top \boldsymbol{I}^{-1}_{A^0,A^0}\operatorname{vec}_{A^0}(\boldsymbol{\Lambda}) - \frac{1}{4}\operatorname{vec}_{A^0}\left(R_2(\widehat{\boldsymbol{\Delta}}_{A^0})\right)\boldsymbol{I}^{-1}_{A^0,A^0}\operatorname{vec}_{A^0}\left(R_2(\widehat{\boldsymbol{\Delta}}_{A^0})\right) + R_1(\widehat{\boldsymbol{\Delta}}_{A^0}).$$

Combining, we obtain that

$$2\left(L_n(\widehat{\boldsymbol{\Omega}}_{\tilde{A}}) - L_n(\widehat{\boldsymbol{\Omega}}_{A^0})\right) = \frac{n}{4}\operatorname{vec}_{\tilde{A}}\left(\boldsymbol{\Lambda}\right)^\top \boldsymbol{I}^{-1}_{B,B}\operatorname{vec}_{\tilde{A}}\left(\boldsymbol{\Lambda}\right) - \frac{n}{4}\operatorname{vec}_{A^0}\left(\boldsymbol{\Lambda}\right)^\top \boldsymbol{I}^{-1}_{A^0,A^0}\operatorname{vec}_{A^0}\left(\boldsymbol{\Lambda}\right) + R(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}, \widehat{\boldsymbol{\Delta}}_{A^0}) \quad \text{(B.31)}$$

where

$$R(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}, \widehat{\boldsymbol{\Delta}}_{A^0}) = nR_1(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}) - \frac{n}{4}\operatorname{vec}_{\tilde{A}}\left(R_2(\widehat{\boldsymbol{\Delta}}_{\tilde{A}})\right)^\top \boldsymbol{I}^{-1}_{\tilde{A},\tilde{A}}\operatorname{vec}_{\tilde{A}}\left(R_2(\widehat{\boldsymbol{\Delta}}_{\tilde{A}})\right) - nR_1(\widehat{\boldsymbol{\Delta}}_{A^0}) + \frac{n}{4}\operatorname{vec}_{A^0}\left(R_2(\widehat{\boldsymbol{\Delta}}_{A^0})\right)^\top \boldsymbol{I}^{-1}_{A^0,A^0}\operatorname{vec}_{A^0}\left(R_2(\widehat{\boldsymbol{\Delta}}_{A^0})\right) \quad \text{(B.32)}$$

is the remainder to be bounded subsequently. For now, we focus on the leading term in the likelihood ratio expansion. Let $\boldsymbol{\lambda} = \sqrt{n}\operatorname{vec}_{\tilde{A}}\left(\boldsymbol{\Sigma}^0 - \boldsymbol{S}\right)$. Now write $\boldsymbol{I}^{-1}_{\tilde{A},\tilde{A}}$ as

$$\boldsymbol{I}^{-1}_{\tilde{A},\tilde{A}} = \begin{pmatrix} \boldsymbol{J}_{A^0,A^0} & \boldsymbol{J}_{A^0,B} \\ \boldsymbol{J}_{B,A^0} & \boldsymbol{J}_{B,B} \end{pmatrix}. \quad \text{(B.33)}$$

Note that $\boldsymbol{I}_{A^0,A^0} = [\boldsymbol{J}^{-1}]_{A^0,A^0} = \left(\boldsymbol{J}_{A^0,A^0} - \boldsymbol{J}_{A^0,B}\boldsymbol{J}^{-1}_{B,B}\boldsymbol{J}_{B,A^0}\right)^{-1}$. Thus,

$$\frac{n}{4}\operatorname{vec}_{\tilde{A}}\left(\boldsymbol{\Lambda}\right)^\top \boldsymbol{I}^{-1}_{\tilde{A},\tilde{A}}\operatorname{vec}_{\tilde{A}}\left(\boldsymbol{\Lambda}\right) - \frac{n}{4}\operatorname{vec}_{A^0}\left(\boldsymbol{\Lambda}\right)^\top \boldsymbol{I}^{-1}_{A^0,A^0}\operatorname{vec}_{A^0}\left(\boldsymbol{\Lambda}\right)$$
$$= \frac{1}{4}\boldsymbol{\lambda}^\top_{\tilde{A}}\boldsymbol{I}^{-1}_{\tilde{A},\tilde{A}}\boldsymbol{\lambda}_{\tilde{A}} - \frac{1}{4}\boldsymbol{\lambda}^\top_{A^0}\boldsymbol{I}^{-1}_{A^0,A^0}\boldsymbol{\lambda}_{A^0}$$
$$= \frac{1}{4}\boldsymbol{\lambda}^\top_{\tilde{A}}\boldsymbol{J}\boldsymbol{\lambda}_{\tilde{A}} - \frac{1}{4}\boldsymbol{\lambda}^\top_{A^0}\left(\boldsymbol{J}_{A^0,A^0} - \boldsymbol{J}_{A^0,B}\boldsymbol{J}^{-1}_{B,B}\boldsymbol{J}_{B,A^0}\right)\boldsymbol{\lambda}_{A^0}$$
$$= \frac{1}{4}\left(\boldsymbol{J}_{B,A^0}\boldsymbol{\lambda}_{A^0} + \boldsymbol{J}_{\tilde{A}\setminus A^0,B\setminus A^0}\boldsymbol{\lambda}_B\right)^\top \boldsymbol{J}^{-1}_{A\setminus A^0,B}\left(\boldsymbol{J}_{\tilde{A}\setminus A^0,A^0}\boldsymbol{\lambda}_{A^0} + \boldsymbol{J}_{B\setminus A^0,B}\boldsymbol{\lambda}_B\right)$$
$$= \frac{1}{4}\boldsymbol{\lambda}^\top_{\tilde{A}}\boldsymbol{J}_{\tilde{A},B}\boldsymbol{J}^{-1}_{B,B}\boldsymbol{J}_{\tilde{A}\setminus A^0,A}\boldsymbol{\lambda}_{\tilde{A}} = \left\|\frac{1}{2}\boldsymbol{J}^{-1/2}_{B,B}\boldsymbol{J}_{B,\tilde{A}}\sqrt{n}\operatorname{vec}_{\tilde{A}}(\boldsymbol{\Lambda})\right\|_2^2. \quad \text{(B.34)}$$

This, together with (B.31), implies that

$$2\left(L_n(\widehat{\boldsymbol{\Omega}}_{\tilde{A}}) - L_n(\widehat{\boldsymbol{\Omega}}_{A^0})\right) = \left\|\frac{1}{2}\boldsymbol{J}^{-1/2}_{B,B}\boldsymbol{J}_{B,\tilde{A}}\sqrt{n}\operatorname{vec}_{\tilde{A}}(\boldsymbol{\Lambda})\right\|_2^2 + R(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}, \widehat{\boldsymbol{\Delta}}_{A^0}), \quad \text{(B.35)}$$

Recall from (B.47) that $\mathrm{Var}\left(\frac{1}{2}\boldsymbol{J}_{B,B}^{-1/2}\boldsymbol{J}_{B,\tilde{A}}\sqrt{n}\,\mathrm{vec}_A(\boldsymbol{\Lambda})\right) = \boldsymbol{I}_{|B|\times|B|}$, thus by Lemma 7 and Lemma 8, if $|B|$ is a fixed constant, $2\Big(L_n(\widehat{\boldsymbol{\Omega}}_{\tilde{A}}) - L_n(\widehat{\boldsymbol{\Omega}}_{A^0})\Big) \xrightarrow{P_0} W_{|\tilde{A}\setminus A^0|}$ provided that $R(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}, \widehat{\boldsymbol{\Delta}}_{A^0}) = o_p(1)$; if $|\tilde{A}\setminus A^0| \to \infty$, $(2|\tilde{A}\setminus A^0|)^{-1/2}\Big(2\big(L_n(\widehat{\boldsymbol{\Omega}}_{\tilde{A}}) - L_n(\widehat{\boldsymbol{\Omega}}_{A^0})\big) - |\tilde{A}\setminus A^0|\Big) \xrightarrow{P_0} N(0,1)$ provided that $R(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}, \widehat{\boldsymbol{\Delta}}_{A^0})/\sqrt{|B|} = o_p(1)$. Next it remains to prove that the remainder term $R(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}, \widehat{\boldsymbol{\Delta}}_{A^0})$ satisfies the aforementioned conditions. Toward this end, we bound $R_1(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}) - R_1(\widehat{\boldsymbol{\Delta}}_{A^0})$ and $\mathrm{vec}_{\tilde{A}}\left(R_2(\widehat{\boldsymbol{\Delta}}_{\tilde{A}})\right)\boldsymbol{I}_{\tilde{A},\tilde{A}}^{-1}\mathrm{vec}_{\tilde{A}}\left(R_2(\widehat{\boldsymbol{\Delta}}_{\tilde{A}})\right) - \mathrm{vec}_{A^0}\left(R_2(\widehat{\boldsymbol{\Delta}}_{A^0})\right)\boldsymbol{I}_{A^0,A^0}^{-1}\mathrm{vec}_{A^0}\left(R_2(\widehat{\boldsymbol{\Delta}}_{A^0})\right)$ respectively.

For $\mathrm{vec}_{\tilde{A}}\left(R_2(\widehat{\boldsymbol{\Delta}}_{\tilde{A}})\right)\boldsymbol{I}_{\tilde{A},\tilde{A}}^{-1}\mathrm{vec}_{\tilde{A}}\left(R_2(\widehat{\boldsymbol{\Delta}}_{\tilde{A}})\right)$, recursively applying $\|C_1C_2\|_F \le \|C_1\|_F\|C_2\|_F$ and using the fact that $\|C_1C_2\|_F \le \lambda_{\max}(C_2)\|C_1\|_F$ and $\|C_1C_2\|_F \le \lambda_{\max}(C_1)\|C_2\|_F$, we obtain

$$
\begin{aligned}
\left\|\mathrm{vec}_{\tilde{A}}\left(\boldsymbol{\Sigma}^0\left(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\boldsymbol{\Sigma}^0\right)^i\right)\right\|_2 &\le \left\|\sqrt{\boldsymbol{\Sigma}^0}\left(\sqrt{\boldsymbol{\Sigma}^0}\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\sqrt{\boldsymbol{\Sigma}^0}\right)^i\sqrt{\boldsymbol{\Sigma}^0}\right\|_F \\
&\le \lambda_{\max}(\boldsymbol{\Sigma}^0)\left\|\sqrt{\boldsymbol{\Sigma}^0}\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\sqrt{\boldsymbol{\Sigma}^0}\right\|_F^i = \lambda_{\max}(\boldsymbol{\Sigma}^0)\|\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\|^i \qquad \text{(B.36)}
\end{aligned}
$$

Summing over i yields that

$$
\begin{aligned}
\left\|\mathrm{vec}_{\tilde{A}}\left(R_2(\widehat{\boldsymbol{\Delta}}_{\tilde{A}})\right)\right\|_2 &\le \sum_{i=2}^{\infty}\left\|\mathrm{vec}_{\tilde{A}}\left(\boldsymbol{\Sigma}^0\left(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\boldsymbol{\Sigma}^0\right)^i\right)\right\|_2 \\
&\le \lambda_{\max}(\boldsymbol{\Sigma}^0)\sum_{i=2}^{\infty}\|\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\|^i \le 2\lambda_{\max}(\boldsymbol{\Sigma}^0)\|\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\|^2. \qquad \text{(B.37)}
\end{aligned}
$$

Consequently,

$$
\begin{aligned}
&\mathrm{vec}_B\left(R_2(\widehat{\boldsymbol{\Delta}}_{\tilde{A}})\right)\boldsymbol{I}_{B,B}^{-1}\mathrm{vec}_{\tilde{A}}\left(R_2(\widehat{\boldsymbol{\Delta}}_{\tilde{A}})\right) \le \left\|\boldsymbol{I}_{B,B}^{-1}\right\|_{\mathrm{opt}}\left\|\mathrm{vec}_{\tilde{A}}\left(R_2(\widehat{\boldsymbol{\Delta}}_{\tilde{A}})\right)\right\|_2^2 \\
\le\ & \lambda_{\min}^{-2}(\boldsymbol{\Sigma}^0)\left\|\mathrm{vec}_{\tilde{A}}\left(R_2(\widehat{\boldsymbol{\Delta}}_{\tilde{A}})\right)\right\|_2^2 \le 4\kappa_0^2\|\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\|^4. \qquad \text{(B.38)}
\end{aligned}
$$

Similarly, $\mathrm{vec}_{A^0}\left(R_2(\widehat{\boldsymbol{\Delta}}_{A^0})\right)\boldsymbol{I}_{A^0,A^0}^{-1}\mathrm{vec}_{A^0}\left(R_2(\widehat{\boldsymbol{\Delta}}_{A^0})\right) \le 4\kappa_0^2\|\widehat{\boldsymbol{\Delta}}_{A^0}\|^4$. Hence,

$$
\begin{aligned}
&\frac{1}{4}\mathrm{vec}_{\tilde{A}}\left(R_2(\widehat{\boldsymbol{\Delta}}_{\tilde{A}})\right)\boldsymbol{I}_{\tilde{A},\tilde{A}}^{-1}\mathrm{vec}_{\tilde{A}}\left(R_2(\widehat{\boldsymbol{\Delta}}_{\tilde{A}})\right) - \frac{1}{4}\mathrm{vec}_{A^0}\left(R_2(\widehat{\boldsymbol{\Delta}}_{A^0})\right)\boldsymbol{I}_{A^0,A^0}^{-1}\mathrm{vec}_{A^0}\left(R_2(\widehat{\boldsymbol{\Delta}}_{A^0})\right) \\
\le\ & \kappa_0^2\|\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\|^4 + \kappa_0^2\|\widehat{\boldsymbol{\Delta}}_{A^0}\|^4 \qquad \text{(B.39)}
\end{aligned}
$$

For $R_1(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}) - R_1(\widehat{\boldsymbol{\Delta}}_{A^0})$, by Cauchy-Schwartz inequality, we have that $\mathrm{tr}(\left(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\boldsymbol{\Sigma}^0\right)^i) \le \|\sqrt{\boldsymbol{\Sigma}^0}\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\sqrt{\boldsymbol{\Sigma}^0}\|_F\left\|(\sqrt{\boldsymbol{\Sigma}^0}\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\sqrt{\boldsymbol{\Sigma}^0})^{i-1}\right\|_F \le \|\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\|^i$; $i = 2, \cdots,$. Hence,

$$
\Big|\sum_{i=4}^{\infty}\frac{(-1)^{i+1}}{i}\mathrm{tr}(\left(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\boldsymbol{\Sigma}^0\right)^i)\Big| \le \sum_{i=4}^{\infty} i^{-1}\|\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\|^i \le \frac{\|\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\|^4}{4(1-\|\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\|)} \le \frac{1}{2}\|\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\|^4. \qquad \text{(B.40)}
$$

Similarly, $\left|\sum_{i=4}^{\infty}\frac{(-1)^{i+1}}{i}\mathrm{tr}(\left(\widehat{\boldsymbol{\Delta}}_{A^0}\boldsymbol{\Sigma}^0\right)^i)\right| \le \frac{1}{2}\|\widehat{\boldsymbol{\Delta}}_{A^0}\|^4$. Combining, we have that

$$\left|R_1(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}) - R_1(\widehat{\boldsymbol{\Delta}}_{A^0})\right| \leq \frac{\left|\mathrm{tr}\left(\left(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\boldsymbol{\Sigma}^0\right)^3\right) - \mathrm{tr}\left(\left(\widehat{\boldsymbol{\Delta}}_{A^0}\boldsymbol{\Sigma}^0\right)^3\right)\right|}{3} + \frac{\|\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\|^4 + \|\widehat{\boldsymbol{\Delta}}_{A^0}\|^4}{2} \tag{B.41}$$

Let $f_{\tilde{A}}(\mathrm{vec}_{\tilde{A}}(\boldsymbol{\Delta})) = \mathrm{tr}\left(\left(\boldsymbol{\Delta}\boldsymbol{\Sigma}^0\right)^3\right)$ with $\mathrm{vec}_{A^c}(\boldsymbol{\Delta}) = \mathbf{0}$. A Taylor expansion of $f_{\tilde{A}}(\mathrm{vec}_{\tilde{A}}(\boldsymbol{\Delta}))$ at $\mathrm{vec}_{A^0}(\boldsymbol{\Delta}))$ yields that

$$\begin{aligned}
\frac{1}{3}\left|\mathrm{tr}\left(\left(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\boldsymbol{\Sigma}^0\right)^3\right) - \mathrm{tr}\left(\left(\widehat{\boldsymbol{\Delta}}_{A^0}\boldsymbol{\Sigma}^0\right)^3\right)\right| &= \frac{1}{3}\left(\mathrm{vec}_{\tilde{A}}(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}) - \mathrm{vec}_{\tilde{A}}(\widehat{\boldsymbol{\Delta}}_{A^0})\right)^\top \nabla f(\mathrm{vec}_{\tilde{A}}(\widehat{\boldsymbol{\Delta}}^*)) \\
&= \left(\mathrm{vec}_{\tilde{A}}(\widehat{\boldsymbol{\Delta}}_{\tilde{A}} - \widehat{\boldsymbol{\Delta}}_{A^0})\right)^\top \mathrm{vec}_{\tilde{A}}\left(\boldsymbol{\Sigma}^0(\widehat{\boldsymbol{\Delta}}^*\boldsymbol{\Sigma}^0)^2\right) = \mathrm{tr}\left(\boldsymbol{\Sigma}^0(\widehat{\boldsymbol{\Delta}}_{\tilde{A}} - \widehat{\boldsymbol{\Delta}}_{A^0})(\boldsymbol{\Sigma}^0\widehat{\boldsymbol{\Delta}}^*)^2\right) \\
&\leq 2\left\|\sqrt{\boldsymbol{\Sigma}^0}(\widehat{\boldsymbol{\Delta}}_{\tilde{A}} - \widehat{\boldsymbol{\Delta}}_{A^0})\sqrt{\boldsymbol{\Sigma}^0}\right\|_F \max\left(\|\sqrt{\boldsymbol{\Sigma}^0}\widehat{\boldsymbol{\Delta}}_{A^0}\sqrt{\boldsymbol{\Sigma}^0}\|_F^2, \|\sqrt{\boldsymbol{\Sigma}^0}\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\sqrt{\boldsymbol{\Sigma}^0}\|_F^2\right)
\end{aligned} \tag{B.42}$$

where $\widehat{\boldsymbol{\Delta}}^*$ is some convex combination of $\widehat{\boldsymbol{\Delta}}_{\tilde{A}}$ and $\widehat{\boldsymbol{\Delta}}_{A^0}$ and the last equality uses (B.36). Lastly, we bound $\left\|\sqrt{\boldsymbol{\Sigma}^0}(\widehat{\boldsymbol{\Delta}}_{\tilde{A}} - \widehat{\boldsymbol{\Delta}}_{A^0})\sqrt{\boldsymbol{\Sigma}^0}\right\|_F = \left\|\boldsymbol{I}_{\tilde{A},\tilde{A}}^{1/2}\,\mathrm{vec}_{\tilde{A}}(\widehat{\boldsymbol{\Delta}}_{\tilde{A}} - \widehat{\boldsymbol{\Delta}}_{A^0})\right\|_2$. By (B.29), we have that

$$\begin{aligned}
\boldsymbol{I}_{\tilde{A},\tilde{A}}^{1/2}\,\mathrm{vec}_{\tilde{A}}(\widehat{\boldsymbol{\Delta}}_{\tilde{A}} - \widehat{\boldsymbol{\Delta}}_{A^0}) &= \boldsymbol{I}_{\tilde{A},\tilde{A}}^{1/2}\left(\mathrm{vec}_{\tilde{A}}(\widehat{\boldsymbol{\Omega}}_{\tilde{A}} - \boldsymbol{\Omega}^0) - \mathrm{vec}_{\tilde{A}}(\widehat{\boldsymbol{\Omega}}_{A^0} - \boldsymbol{\Omega}^0)\right) \\
&= \frac{1}{2}\boldsymbol{I}_{\tilde{A},\tilde{A}}^{-1/2}\,\mathrm{vec}_{\tilde{A}}(\boldsymbol{\Lambda} + R_2(\widehat{\boldsymbol{\Delta}}_{\tilde{A}})) - \frac{1}{2}\boldsymbol{I}_{\tilde{A},\tilde{A}}^{1/2}\begin{bmatrix} \boldsymbol{I}_{A^0,A^0}^{-1}\,\mathrm{vec}_{A^0}(\boldsymbol{\Lambda} + R_2(\widehat{\boldsymbol{\Delta}}_{A^0})) \\ \mathbf{0} \end{bmatrix} \\
&= \frac{1}{2}\boldsymbol{I}_{\tilde{A},\tilde{A}}^{-1/2}\left(\mathrm{vec}_{\tilde{A}}(\boldsymbol{\Lambda} + R_2(\widehat{\boldsymbol{\Delta}}_{\tilde{A}})) - \begin{bmatrix} \mathrm{vec}_{A^0}(\boldsymbol{\Lambda} + R_2(\widehat{\boldsymbol{\Delta}}_{A^0})) \\ \boldsymbol{I}_{B,A^0}\boldsymbol{I}_{A^0,A^0}^{-1}\,\mathrm{vec}_{A^0}(\boldsymbol{\Lambda} + R_2(\widehat{\boldsymbol{\Delta}}_{A^0})) \end{bmatrix}\right) \\
&= \frac{1}{2}\boldsymbol{I}_{\tilde{A},\tilde{A}}^{-1/2}\begin{bmatrix} \mathrm{vec}_{A^0}(R_2(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}) - R_2(\widehat{\boldsymbol{\Delta}}_{A^0})) \\ \mathrm{vec}_B(\boldsymbol{\Lambda} + R_2(\widehat{\boldsymbol{\Delta}}_{\tilde{A}})) - \boldsymbol{I}_{B,A^0}\boldsymbol{I}_{A^0,A^0}^{-1}\,\mathrm{vec}_{A^0}(\boldsymbol{\Lambda} + R_2(\widehat{\boldsymbol{\Delta}}_{A^0})) \end{bmatrix},
\end{aligned} \tag{B.43}$$

where $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^0 - \boldsymbol{S}$. Let $\boldsymbol{J} = \boldsymbol{I}_{\tilde{A},\tilde{A}}^{-1}$. An application of an inequality $\|\boldsymbol{I}_{\tilde{A},\tilde{A}}^{-1/2}\boldsymbol{x}\|_2^2 = \boldsymbol{x}^\top\boldsymbol{J}\boldsymbol{x} \leq 2\boldsymbol{x}_{A^0}^\top\boldsymbol{J}_{A^0,A^0}\boldsymbol{x}_{A^0} + 2\boldsymbol{x}_B^\top\boldsymbol{J}_{B,B}\boldsymbol{x}_B$ yields that

$$\begin{aligned}
&\left\|\boldsymbol{I}_{\tilde{A},\tilde{A}}^{-1/2}\begin{bmatrix} \mathrm{vec}_{A^0}(R_2(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}) - R_2(\widehat{\boldsymbol{\Delta}}_{A^0})) \\ \mathrm{vec}_B(\boldsymbol{\Lambda} + R_2(\widehat{\boldsymbol{\Delta}}_{\tilde{A}})) - \boldsymbol{I}_{B,A^0}\boldsymbol{I}_{A^0,A^0}^{-1}\,\mathrm{vec}_{A^0}(\boldsymbol{\Lambda} + R_2(\widehat{\boldsymbol{\Delta}}_{A^0})) \end{bmatrix}\right\|_F^2 \\
&\leq 2\left\|\boldsymbol{J}_{B,B}^{1/2}\left(\mathrm{vec}_{B\setminus A^0}(\boldsymbol{\Lambda} + R_2(\widehat{\boldsymbol{\Delta}}_{\tilde{A}})) - \boldsymbol{I}_{B,A^0}\boldsymbol{I}_{A^0,A^0}^{-1}\,\mathrm{vec}_{A^0}(\boldsymbol{\Lambda} + R_2(\widehat{\boldsymbol{\Delta}}_{A^0}))\right)\right\|_2^2 \\
&\quad +2\left\|\boldsymbol{J}_{A^0,A^0}^{1/2}\,\mathrm{vec}_{A^0}(R_2(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}) - R_2(\widehat{\boldsymbol{\Delta}}_{A^0}))\right\|_2^2.
\end{aligned} \tag{B.44}$$

Moreover, $\boldsymbol{J}_{B,B}^{-1}\boldsymbol{J}_{B,A^0} + \boldsymbol{I}_{B,A^0}\boldsymbol{I}_{A^0,A^0}^{-1} = \mathbf{0}$. Using this, we have that

$$\left\| \boldsymbol{J}_{B,B}^{1/2} \left(\text{vec}_B(\boldsymbol{\Lambda}) - \boldsymbol{I}_{B,A^0} \boldsymbol{I}_{A^0,A^0}^{-1} \text{vec}_{A^0}(\boldsymbol{\Lambda}) \right) \right\|_2^2$$

$$= \left\| \boldsymbol{J}_{B,B}^{-1/2} \left(\boldsymbol{J}_{B,B} \text{vec}_B(\boldsymbol{\Lambda}) + \boldsymbol{J}_{B,A^0} \text{vec}_{A^0}(\boldsymbol{\Lambda}) \right) \right\|_2^2 = \left\| \boldsymbol{J}_{B,B}^{-1/2} \boldsymbol{J}_{B,\tilde{A}} \text{vec}_{\tilde{A}}(\boldsymbol{\Lambda}) \right\|_2^2 . \quad \text{(B.45)}$$

This, together with (B.43) and (B.44), implies that

$$\left\| \sqrt{\boldsymbol{\Sigma}^0} (\widehat{\boldsymbol{\Delta}}_{\tilde{A}} - \widehat{\boldsymbol{\Delta}}_{A^0}) \sqrt{\boldsymbol{\Sigma}^0} \right\|_F^2$$

$$\leq \frac{1}{2} \left\| \boldsymbol{J}_{A^0,A^0}^{1/2} \text{vec}_{A^0}(R_2(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}) - R_2(\widehat{\boldsymbol{\Delta}}_{A^0})) \right\|_2^2 + \frac{1}{2} \left\| \boldsymbol{J}_{B,B}^{-1/2} \boldsymbol{J}_{B,B} \text{vec}_B(\boldsymbol{\Lambda}) \right\|_2^2 . \quad \text{(B.46)}$$

By (B.3), the covariance matrix of $\boldsymbol{J}_{B,B}^{-1/2} \boldsymbol{J}_{B,B} \text{vec}_{\tilde{A}}(\boldsymbol{\Lambda})$ is

$$\text{Var}\left(\boldsymbol{J}_{B,B}^{-1/2} \boldsymbol{J}_{B,B} \text{vec}_B(\boldsymbol{\Lambda}) \right) = n^{-1} \boldsymbol{J}_{B,B}^{-1/2} \boldsymbol{J}_{B,B} \text{Var}\left(\sqrt{n}\, \text{vec}_{\tilde{A}}(\boldsymbol{\Lambda}) \right) \boldsymbol{J}_{\tilde{A},B} \boldsymbol{J}_{B,B}^{-1/2}$$

$$= n^{-1} \boldsymbol{J}_{B,B}^{-1/2} \boldsymbol{J}_{B,\tilde{A}} \left(4\boldsymbol{J}^{-1} \right) \boldsymbol{J}_{\tilde{A},B} \boldsymbol{J}_{B,B}^{-1/2} = 4n^{-1} \boldsymbol{J}_{B,B}^{-1/2} \boldsymbol{J}_{B,B} \boldsymbol{J}_{B,B}^{-1/2} = 4n^{-1} \boldsymbol{I}_{|B| \times |B|} , \quad \text{(B.47)}$$

By Lemma 3, $\left\| \boldsymbol{J}_{B,B}^{-1/2} \boldsymbol{J}_{B,A} \text{vec}_{\tilde{A}}(\boldsymbol{\Lambda}) \right\|_2^2 \leq |B| \left\| \boldsymbol{J}_{B,B}^{-1/2} \boldsymbol{J}_{B,\tilde{A}} \text{vec}_A(\boldsymbol{\Lambda}) \right\|_\infty^2 = O_p\left(\frac{|B| \log |B|}{n} \right)$. Using this and (B.37), we bound (B.46) as follows:

$$\left\| \sqrt{\boldsymbol{\Sigma}^0} (\widehat{\boldsymbol{\Delta}}_{\tilde{A}} - \widehat{\boldsymbol{\Delta}}_{A^0}) \sqrt{\boldsymbol{\Sigma}^0} \right\|_F^2 \leq 2^{-1} \lambda_{\min}^{-2}(\boldsymbol{\Sigma}^0) \left\| R_2(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}) - R_2(\widehat{\boldsymbol{\Delta}}_{A^0}) \right\|_F^2 + O_p\left(\frac{|B| \log |B|}{n} \right)$$

$$\leq 2\lambda_{\min}^{-2}(\boldsymbol{\Sigma}^0) \max\left(\left\| R_2(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}) \right\|_F^2 , \left\| R_2(\widehat{\boldsymbol{\Delta}}_{A^0}) \right\|_F^2 \right) + O_p\left(\frac{|B| \log |B|}{n} \right)$$

$$\leq 8\kappa_0^2 \max\left(\|\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\|^4, \|\widehat{\boldsymbol{\Delta}}_{A^0}\|^4 \right) + O_p\left(\frac{|B| \log |B|}{n} \right).$$

Let $\Delta = \max\left(\|\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\|, \|\widehat{\boldsymbol{\Delta}}_{A^0}\| \right)$. Then combining the above bound with (B.42), we obtain

$$\frac{1}{3} \left| \text{tr}\left((\widehat{\boldsymbol{\Delta}}_{\tilde{A}} \boldsymbol{\Sigma}^0)^3 \right) - \text{tr}\left((\widehat{\boldsymbol{\Delta}}_{A^0} \boldsymbol{\Sigma}^0)^3 \right) \right|$$

$$\leq 2 \left\| \sqrt{\boldsymbol{\Sigma}^0} (\widehat{\boldsymbol{\Delta}}_{\tilde{A}} - \widehat{\boldsymbol{\Delta}}_{A^0}) \sqrt{\boldsymbol{\Sigma}^0} \right\|_F \max\left(\|\widehat{\boldsymbol{\Delta}}_{A^0}\|^2, \|\widehat{\boldsymbol{\Delta}}_{\tilde{A}}\|^2 \right)$$

$$\leq 4\Delta^2 \max\left(3\kappa_0 \Delta^2 , O_p\left(\sqrt{\frac{|B| \log |B|}{n}} \right) \right).$$

This together with (B.39) and (B.41) implies that the remainder term $R(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}, \widehat{\boldsymbol{\Delta}}_{A^0})$ defined in (B.32) is bounded by $n\Delta^2 \max\left(\kappa_0^2 \Delta^2 , O_p\left(\sqrt{\frac{|B| \log |B|}{n}} \right) \right)$ up to some positive constants. By Lemma 5, we have that $\Delta^2 = O_p\left(\frac{|\tilde{A}| \log p}{n} \right)$. This together with (B.39) and (B.41) yields that

$$R(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}, \widehat{\boldsymbol{\Delta}}_{A^0}) = O_p\left(\max\left(\frac{\kappa_0^2 |\tilde{A}|^2 \log^2(p+1)}{n}, |\tilde{A}| \log(p+1) \sqrt{\frac{|B| \log |B|}{n}} \right) \right)$$

Hence, if $|B|$ is fixed, $R(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}, \widehat{\boldsymbol{\Delta}}_{A^0}) = o_p(1)$, provided that $\frac{\kappa_0^2 |A|^2 \log^2 p}{n} \to 0$; and if $|\tilde{A} \backslash A^0| \to \infty$,

$R(\widehat{\boldsymbol{\Delta}}_{\tilde{A}}, \widehat{\boldsymbol{\Delta}}_{A^0})/\sqrt{|B|} = o_p(1)$, provided that $\frac{\kappa_0^2|\tilde{A}|^2 \log^2 p \log(|B|)}{n} \to 0$. This completes the proof.

C Proofs of Theorem 3 and 4

Proof of Theorem 3. Let $\Lambda_n(B)$ be the likelihood ratio test statistic defined in Theorem 1. A measure change from $\mathbb{P}_{\boldsymbol{\theta}^n}$ to $\mathbb{P}_{\boldsymbol{\theta}^0}$ yields that for any $u \geq 0$,

$$
\begin{aligned}
& \mathbb{P}_{\boldsymbol{\theta}^n}(\Lambda_n(B) \geq u) = \mathbb{E}_{\boldsymbol{\theta}^n}\mathbb{I}(\Lambda_n(B) \geq u) \\
= \; & \mathbb{E}_{\boldsymbol{\theta}^0}\left(\mathbb{I}(\Lambda_n(B) > u)\exp(\sqrt{n}\,\mathrm{vec}_B(\delta_n)^\top Z_n - \frac{n\,\mathrm{vec}_B(\delta_n)^\top \boldsymbol{I}_{B,B}\,\mathrm{vec}_B(\delta_n)}{2} + R_n(\boldsymbol{\theta}^0, \delta_n))\right),
\end{aligned}
$$

where $\mathbb{P}_{\boldsymbol{\theta}^n}$ is the probability measure under H_a, $Z_n = n^{-1/2}\frac{\partial L_n(\boldsymbol{\theta}^0)}{\partial \theta_B}$, $\boldsymbol{I}$ is the Fisher information matrix, and $R_n(\boldsymbol{\theta}^0, \delta_n)) = L_n(\boldsymbol{\theta}^n) - L_n(\boldsymbol{\theta}^0) - \sqrt{n}\,\mathrm{vec}_B(\delta_n)^\top Z_n + \frac{n\,\mathrm{vec}_B(\delta_n)^\top \boldsymbol{I}_{B,B}\,\mathrm{vec}_B(\delta_n)}{2}$. We will verify later that

$$
R_n(\boldsymbol{\theta}^0, \delta_n)) \xrightarrow{\mathbb{P}_{\theta^0}} 0 \tag{C.1}
$$

in the Gaussian graphical model and linear regression model.

For the Gaussian graphical model, we first verify (C.1). Now let $\boldsymbol{h}_n = \sqrt{n}\,\mathrm{vec}_B(\delta_n)$ with $\|\boldsymbol{h}_n\|_2 = h$. Then $Z_n = n^{-1/2}\frac{\partial L_n(\Omega)}{\partial \Omega_B} = \sqrt{n}\,\mathrm{vec}_B((\boldsymbol{\Omega}^0)^{-1} - \boldsymbol{S}) = \sqrt{n}\,\mathrm{vec}_B(\boldsymbol{\Lambda})$. It follows from the Taylor expansion of $\log\det(\cdot)$ that

$$
\begin{aligned}
& L_n(\boldsymbol{\theta}^n) - L_n(\boldsymbol{\theta}^0) = n\left(\log\det(\boldsymbol{\Omega}^n) - \mathrm{tr}(\boldsymbol{\Omega}^n\boldsymbol{S}) - \log\det(\boldsymbol{\Omega}^0) + \mathrm{tr}(\boldsymbol{\Omega}^0\boldsymbol{S})\right) \\
= \; & \boldsymbol{h}_n^\top\sqrt{n}\,\mathrm{vec}_B((\boldsymbol{\Omega}^0)^{-1} - \boldsymbol{S}) - \sqrt{n}\boldsymbol{h}_n^\top\,\mathrm{vec}_B((\boldsymbol{\Omega}^0)^{-1}) + n(\log\det(\boldsymbol{\Omega}^n) - \log\det(\boldsymbol{\Omega}^0)) \\
= \; & \boldsymbol{h}_n^\top Z_n - \frac{1}{2}\boldsymbol{h}_n^\top \boldsymbol{I}_{B,B}\boldsymbol{h}_n + r(\boldsymbol{\Omega}^n),
\end{aligned}
$$

where we have used (B.26) and

$$r(\mathbf{\Omega}^n) = n\sum_{i=3}^{\infty}(-1)^{i+1}\frac{\operatorname{tr}\left[(\sqrt{\Sigma^0}(\mathbf{\Omega}^n-\mathbf{\Omega}^0)\sqrt{\Sigma^0})^i\right]}{i} \tag{C.2}$$

By similar calculations as in (B.40), we have that

$$|r(\mathbf{\Omega}^n)| \leq \begin{cases} \frac{n}{3}\sum_{i=3}^{n}(\boldsymbol{h}_n^\top \boldsymbol{I}_{B,B}\boldsymbol{h}_n)^{i/2}\left(\frac{|B|^{1/4}}{\sqrt{n}}\right)^i & \text{if } |B|\to\infty \\ \frac{n}{3}\sum_{i=3}^{n}(\boldsymbol{h}_n^\top \boldsymbol{I}_{B,B}\boldsymbol{h}_n)^{i/2}\left(\frac{1}{\sqrt{n}}\right)^i & \text{if } |B| \text{ is fixed.} \end{cases} \tag{C.3}$$

Hence, when $|B|$ is fixed and n is large enough, we have that $|r(\mathbf{\Omega}^n)| \leq (\boldsymbol{h}_n^\top \boldsymbol{I}_{B,B}\boldsymbol{h}_n)^{3/2}n^{-1/2} \to 0$. When $|B| \to \infty$ but $|B|^{3/2}/n \to 0$, we have that $|r(\mathbf{\Omega}^n)| \leq (\boldsymbol{h}_n^\top \boldsymbol{I}_{B,B}\boldsymbol{h}_n)^{3/2}\frac{|B|^{3/4}}{n^{1/2}} \to 0$. Therefore,

$$R_n(\boldsymbol{\theta}^0,\delta_n) = L_n(\boldsymbol{\theta}^n) - L_n(\boldsymbol{\theta}^0) - \boldsymbol{h}_n^\top Z_n + \frac{1}{2}\boldsymbol{h}_n^\top \boldsymbol{I}_{B,B}\boldsymbol{h}_n = r(\mathbf{\Omega}^n) \to 0\,. \tag{C.4}$$

By (B.35), we have that, with probability tending to 1 under $P_{\boldsymbol{\theta}^0}$,

$$\Lambda_n(B) = \left\|\frac{1}{2}\boldsymbol{J}_{B,B}^{-1/2}\boldsymbol{J}_{B,\tilde{A}}\sqrt{n}\operatorname{vec}_{\tilde{A}}(\mathbf{\Lambda})\right\|_2^2 + R(\widehat{\mathbf{\Delta}}_{\tilde{A}},\widehat{\mathbf{\Delta}}_{A^0}). \tag{C.5}$$

Note that $\operatorname{Var}(\operatorname{vec}_{\tilde{A}}(\mathbf{\Lambda})) = 4\boldsymbol{I}$. Hence, by Lemmas 7 and 8,

$$\left(\frac{1}{2}\boldsymbol{J}_{B,B}^{-1/2}\boldsymbol{J}_{B,\tilde{A}}\sqrt{n}\operatorname{vec}_{\tilde{A}}(\mathbf{\Lambda}), \frac{1}{2}\sqrt{n}\operatorname{vec}_B(\mathbf{\Lambda})\right) \xrightarrow{d} (Z_1,Z_2) \sim N\left(\mathbf{0}, \begin{pmatrix} \boldsymbol{I}_{|B|\times|B|} & \boldsymbol{J}_{B,B}^{-1/2} \\ \boldsymbol{J}_{B,B}^{-1/2} & \boldsymbol{I}_{B,B}\end{pmatrix}\right), \tag{C.6}$$

where $\boldsymbol{J} = \boldsymbol{I}^{-1}$. Therefore,

$$Z_1 \sim N(0, I_{|B|\times|B|}) \text{ and } Z_2 \mid Z_1 = z_1 \sim N\left(\boldsymbol{J}_{B,B}^{-1/2}z_1\,, \boldsymbol{I}_{B,A^0}\boldsymbol{I}_{A^0,A^0}^{-1}\boldsymbol{I}_{A^0,B}\right) \tag{C.7}$$

where the fact that $\boldsymbol{J}_{B,B} = (\boldsymbol{I}_{B,B} - \boldsymbol{I}_{B,A^0}\boldsymbol{I}^{-1}_{A^0,A^0}\boldsymbol{I}_{A^0,B})^{-1}$ is used. Hence, for any θ_j; $j \in B^c$,

$$
\begin{aligned}
& P_{H_a}(\Lambda_n(B) \geq u) \to \mathbb{E}\left(\mathbb{I}(\|Z_1\|_2^2 \geq u)\exp(\boldsymbol{h}_n^\top Z_2 - \frac{1}{2}\boldsymbol{h}_n^\top \boldsymbol{I}_{B,B}\boldsymbol{h}_n)\right) \\
= \quad & \exp\left(-\frac{1}{2}\boldsymbol{h}_n^\top \boldsymbol{I}_{B,B}\boldsymbol{h}_n\right)\mathbb{E}_{Z_1}\left[\mathbb{I}(\|Z_1\|_2^2 \geq u)\mathbb{E}_{Z_2|Z_1}\left(\exp(\boldsymbol{h}_n^\top Z_2)\right]\right. \\
= \quad & \exp\left(-\frac{1}{2}\boldsymbol{h}_n^\top \boldsymbol{J}^{-1}_{B,B}\boldsymbol{h}_n\right)\mathbb{E}_{Z_1}\left[\mathbb{I}(\|Z_1\|_2^2 \geq u)\exp\left(Z_1^\top \boldsymbol{J}^{-1/2}_{B,B}\boldsymbol{h}_n\right)\right] \\
= \quad & \mathbb{E}_{Z_1}\mathbb{I}(\|Z_1 + \boldsymbol{J}^{-1/2}_{B,B}\boldsymbol{h}_n\|_2^2 \geq u) = \mathbb{P}\left(\|Z_1 + \boldsymbol{J}^{-1/2}_{B,B}\boldsymbol{h}_n\|_2^2 \geq u\right)
\end{aligned}
$$

where we have used the fact that $\boldsymbol{J}^{-1}_{B,B} = \boldsymbol{I}_{B,B} - \boldsymbol{I}_{B,A^0}\boldsymbol{I}^{-1}_{A^0,A^0}\boldsymbol{I}_{A^0,B}$. Hence, we must have $\Lambda_n(B) \xrightarrow{d} \|Z_1 + bmJ^{-1/2}_{B,B}\boldsymbol{h}_n\|_2^2$ with $Z_1 \sim N(\mathbf{0}, I_{|B|\times|B|})$ when $|B|$ is fixed. When $|B| \to \infty$, for any vector v with $\|v\|_2 = c|B|^{1/4}$ for some constant c, we have that

$$
\frac{\|Z+v\|_2^2 - |B|}{\sqrt{2|B|}} = \frac{\|Z\|_2^2 - |B|}{\sqrt{2|B|}} + \frac{\|v\|_2}{\sqrt{2}|B|^{1/4}}\left(\frac{2v^\top Z}{\|v\|_2|B|^{1/4}} + \frac{\|v\|_2}{|B|^{1/4}}\right) \xrightarrow{d} N\left(\frac{c^2}{\sqrt{2}}, 1\right), \quad \text{(C.8)}
$$

because the first term converges to $N(0,1)$ by CLT, and the second term converges $c^2/\sqrt{2}$ to since $\frac{2v^\top Z}{\|v\|_2|B|^{1/4}} \to 0$ in probability.

Consequently, the *local limiting power functions* for the proposed CMLR test is

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

where $\alpha > 0$ is the level of significance, $\boldsymbol{Z} \sim N(\mathbf{0}, \boldsymbol{I}_{|B|\times|B|})$ is a multivariate normal random variable, and $\boldsymbol{J}_{B,B}$ is the asymptotic variance of $\text{vec}_B(\widehat{\boldsymbol{\Omega}}^{(1)})$.

To make a comparison between the debiased lasso test proposed in [3], we consider the case when $|B| = 1$. Assume that $B = \{(i,j)\}$. In this case, the *local limiting power functions* for the proposed method is

$$\pi_{LR}(h, \theta_{B^c}) = \mathbb{P}\left(\left(Z + \frac{|h|}{\sigma_{\text{LR}}}\right)^2 > \chi^2_\alpha\right) = \mathbb{P}\left(\left|Z + \frac{|h|}{\sigma}\right| > z_{\alpha/2}\right) \tag{C.9}$$

where σ^2_{LR} is the asymptotic variance of $\hat{\omega}^{(1)}_{ij}$. In contrast, The *local limiting power functions* for the debiased lasso test proposed in [3] is

$$\pi_{debias}(h, \theta_{B^c}) = \mathbb{P}\left(\left|Z + \frac{|h|}{\sqrt{\omega^2_{ij} + \omega_{ii}\omega_{jj}}}\right| > z_{\alpha/2}\right) \tag{C.10}$$

where $Z \sim N(0, 1)$ is a standard normal random variable. By applying Corollary 1, we have that $\sigma^2_{\text{LR}} < \omega^2_{ij} + \omega_{ii}\omega_{jj}$, which implies that our $\pi_{LR}(h, \theta_{B^c}) \geq \pi_{debias}(h, \theta_{B^c})$. This completes the proof.

Proof of Theorem 4. The proof is similar to that of Theorem 3. Again, we first verify that (C.1) is satisfied for linear regression. Toward that end, let $\boldsymbol{h}_n = \sqrt{n}\,\text{vec}_B(\delta_n)$ with $\|\boldsymbol{h}_n\|_2 = h$. Notice that $L_n(\theta) = L_n(\beta, \sigma) = n\log(1/\sqrt{2\pi}\sigma) - (2\sigma^2)^{-1}\|y - X\beta\|^2_2$.

$$Z_n = n^{-1/2}\frac{\partial L_n(\beta^0)}{\partial \beta^0_B} = n^{-1/2}\sigma^{-2}\,\text{vec}_B\left(X^\top(y - X\beta^0)\right) = n^{-1/2}\sigma^{-2}\,\text{vec}_B(X^\top\boldsymbol{\epsilon}), \tag{C.11}$$

where $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I_{n\times n})$. Moreover, we have that

$$\begin{aligned} & L_n(\boldsymbol{\theta}^n) - L_n(\boldsymbol{\theta}^0) = (2\sigma^2)^{-1}\left(\|y - X\beta^0\|^2_2 - \|y - X(\beta^0 + \delta_n)\|^2_2\right) \\ = & \sqrt{n}\,\text{vec}_B(\delta_n)^\top n^{-1/2}\sigma^{-2}\,\text{vec}_B(X^\top(y - X\beta^0)) - (2\sigma^2)^{-1}\,\text{vec}_B(\delta_n)^\top(X^\top X)_{B,B}\,\text{vec}_B(\delta_n) \\ = & \boldsymbol{h}^\top_n Z_n - \frac{1}{2}\boldsymbol{h}^\top_n \boldsymbol{I}_{B,B}\boldsymbol{h}_n \end{aligned}$$

where $\boldsymbol{I} = (n\sigma^2)^{-1}X^\top X$. Hence (C.1) is satisfied with the remaining term to be exactly 0.

By similar arguments used in Theorem 2 and the fact that $\|\epsilon\|^2_2/n \xrightarrow{\mathbb{P}_{\beta^0}} 0$, we have that the likelihood ratio test statistic is

$$\Lambda_n(B) = \boldsymbol{\epsilon}^\top (\boldsymbol{P}_{A^0 \cup B} - \boldsymbol{P}_{A^0})\boldsymbol{\epsilon} + R(\boldsymbol{\epsilon}) \tag{C.12}$$

where $R(\boldsymbol{\epsilon}) \xrightarrow{\mathbb{P}_{\beta^0}} 0$. Moreover, since the matrix $\boldsymbol{P}_{A^0 \cup B} - \boldsymbol{P}_{A^0}$ is idempotent and has rank $|B|$, there must exist $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{|B|}$ such that $\boldsymbol{P}_{A^0 \cup B} - \boldsymbol{P}_{A^0} = \sum_{k=1}^{|B|} \boldsymbol{a}_k \boldsymbol{a}_k^\top$ and

$$\Lambda_n(B) = \sum_{k=1}^{|B|} (\boldsymbol{a}_k^\top \boldsymbol{\epsilon})^2 + R(\boldsymbol{\epsilon}) \tag{C.13}$$

Note that, under $\mathbb{P}_{\beta^0}$, we have that

$$\big((\boldsymbol{a}_1^\top \boldsymbol{\epsilon}, \ldots \boldsymbol{a}_{|B|}^\top \boldsymbol{\epsilon}), \mathrm{vec}_B(X^\top \boldsymbol{\epsilon})\big) = (Z_1, Z_2) \sim N\left(\boldsymbol{0}, \begin{pmatrix} I_{|B|\times|B|} & \boldsymbol{A}\boldsymbol{X}_B \\ \boldsymbol{X}_B^\top \boldsymbol{A}^\top & \boldsymbol{X}_B^\top \boldsymbol{X}_B \end{pmatrix}\right) \tag{C.14}$$

where $\boldsymbol{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{|B|})^\top \in \mathbb{R}^{|B|\times n}$.

Therefore,

$$Z_1 \sim N(0, I_{|B|\times|B|}) \text{ and } Z_2 \mid Z_1 = z_1 \sim N\left(\boldsymbol{X}_B^\top \boldsymbol{A}^\top z_1\,, \boldsymbol{X}_B^\top (I_{n\times n} - \boldsymbol{A}^\top \boldsymbol{A})\boldsymbol{X}_B\right) \tag{C.15}$$

Hence, for any β_j; $j \in B^c$ and any $u \geq 0$,

$$
\begin{aligned}
& P_{H_a}(\Lambda_n(B) \geq u) \\
\to \; & \mathbb{E}\left(\mathbb{I}(\|Z_1\|_2^2 \geq u) \exp(\boldsymbol{h}_n^\top Z_2 - \frac{1}{2}\boldsymbol{h}_n^\top \boldsymbol{X}_B^\top \boldsymbol{X}_B \boldsymbol{h}_n\right) \\
= \; & \exp\left(-\frac{1}{2}\boldsymbol{h}_n^\top \boldsymbol{X}_B^\top \boldsymbol{X}_B \boldsymbol{h}_n)\right) \mathbb{E}_{Z_1}\left[\mathbb{I}(\|Z_1\|_2^2 \geq u)\mathbb{E}_{Z_2|Z_1}\left(\exp(\boldsymbol{h}_n^\top Z_2)\right]\right. \\
= \; & \exp\left(-\frac{1}{2}\boldsymbol{h}_n^\top \boldsymbol{X}_B^\top \boldsymbol{X}_B \boldsymbol{h}_n\right) \mathbb{E}_{Z_1}\left[\mathbb{I}(\|Z_1\|_2^2 \geq u) \exp\left(Z_1^\top \boldsymbol{A}\boldsymbol{X}_B \boldsymbol{h}_n + \frac{1}{2}\boldsymbol{h}_n^\top \boldsymbol{X}_B^\top (I_{n\times n} - \boldsymbol{A}^\top \boldsymbol{A})\boldsymbol{X}_B \boldsymbol{h}_n\right)\right] \\
= \; & \exp\left(-\frac{1}{2}\boldsymbol{h}_n^\top \boldsymbol{X}_B^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{X}_B \boldsymbol{h}_n\right) \mathbb{E}_{Z_1}\left[\mathbb{I}(\|Z_1\|_2^2 \geq u) \exp\left(Z_1^\top \boldsymbol{A}\boldsymbol{X}_B \boldsymbol{h}_n\right)\right] \\
= \; & \mathbb{E}_{Z_1}\mathbb{I}(\|Z_1 + \boldsymbol{A}\boldsymbol{X}_B \boldsymbol{h}_n\|_2^2 \geq u) = \mathbb{P}\left(\|Z_1 + \boldsymbol{A}\boldsymbol{X}_B \boldsymbol{h}_n\|_2^2 \geq u\right)
\end{aligned}
$$

Hence, we must have $\Lambda_n(B) \xrightarrow{d} \|Z + \boldsymbol{A}\boldsymbol{X}_B \boldsymbol{h}_n\|_2^2$ with $Z \sim N(\mathbf{0}, I_{|B|\times|B|})$ when $|B|$ is fixed. When $|B| \to \infty$, a similar argument used in Theorem 3 can be applied.

Consequently, the *local limiting power functions* for the proposed CMLR test is

$$
\pi_{LR}(h, \beta_{B^c}) = \begin{cases} \mathbb{P}\left(\|Z + \boldsymbol{A}\boldsymbol{X}_B \boldsymbol{h}_n\|_2^2 \geq \chi^2_{\alpha,|B|}\right) & \text{if } |B| \text{ is fixed,} \\ \mathbb{P}\left(Z_1 + \frac{\|\boldsymbol{A}\boldsymbol{X}_B \boldsymbol{h}_n\|_2^2}{\sqrt{2|B|}} \geq z_\alpha\right) & \text{if } |B| \to \infty \quad . \end{cases} \tag{C.16}
$$

where $\alpha > 0$ is the level of significance, $Z \sim N(\mathbf{0}, \boldsymbol{I}_{|B|\times|B|})$ is a multivariate normal random variable, and $Z_1 \sim N(0,1)$ is a standard normal random variable.

Since $\boldsymbol{A}\boldsymbol{X}_B$ has full rank $|B|$, it is easy to see that when $\|\boldsymbol{h}_n\|_2 \to \infty$ and $|B|$ is finite, then $\pi_{LR}(h, \beta_{B^c}) \to 1$; and when $\|\boldsymbol{h}_n\|_2^2/\sqrt{|B|} \to \infty$ and $|B| \to \infty$, then $\pi_{LR}(h, \beta_{B^c}) \to 1$. This completes the proof.

References

[1] P. Billingsley. *Convergence of probability measures*, volume 493. John Wiley & Sons, 2009.

[2] L. D. Brown. Fundamentals of statistical exponential families with applications in statistical decision theory. *Lecture Notes-monograph series*, pages 1–279, 1986.

[3] J. Janková and S. Van de Geer. Honest confidence regions and optimality in high-dimensional precision matrix estimation. *TEST*, pages 1–20, 2016.