2020 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 21-24, 2020, ESPOO, FINLAND

SCAD-PENALIZED COMPLEX GAUSSIAN GRAPHICAL MODEL SELECTION

Jitendra K. Tugnait

Department of Electrical & Computer Engineering
Auburn University, Auburn, AL 36849, USA
tugnajk@auburn.edu

ABSTRACT

We consider the problem of estimating the conditional independence graph (CIG) of a sparse, high-dimensional proper complex-valued Gaussian graphical model (CGGM). For CGGMs, the problem reduces to estimation of the inverse covariance matrix with more unknowns than the sample size. We consider a smoothly clipped absolute deviation (SCAD) penalty instead of the ℓ_1-penalty to regularize the problem, and analyze a SCAD-penalized log-likelihood based objective function to establish consistency and sparsistency of a local estimator of inverse covariance in a neighborhood of the true value. A numerical example is presented to illustrate the advantage of SCAD-penalty over the usual ℓ_1-penalty.

Keywords: Complex Gaussian graphical models; undirected graph; SCAD penalty; consistency; sparsistency.

1. INTRODUCTION

We consider estimation of the conditional independence graph (CIG) of a proper complex-valued Gaussian graphical model (GGM). Given $\boldsymbol{x} \in \mathbb{C}^p$ with $\boldsymbol{x} \sim \mathcal{N}_c(0, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} \succ 0$, the conditional dependency structure among the p components $x_1, x_2, \cdots, x_p$ is represented using an undirected graph $\mathcal{G} = (V, \mathcal{E})$, where $V = \{1, 2, \cdots, p\} = [p]$ is the set of p nodes corresponding to x_is, and $\mathcal{E} \subset [p] \times [p]$ is the set of undirected edges that specify conditional dependencies among x_i's. In CIG $\mathcal{G}$, edge $\{i, j\} \notin \mathcal{E}$ iff x_i and x_j are conditionally independent given the remaining p-2 variables x_ℓ, $\ell \in [p]$, $\ell \neq i$, $\ell \neq j$. For complex GGM (CGGM), $\{i, j\} \notin \mathcal{E} \Leftrightarrow \Omega_{ij} = 0$ [1, Theorem 7.1] where Ω_{ij} denotes the (i, j)th component of $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$.

We consider the sparse, high-dimensional case. Given n i.i.d. realizations $\boldsymbol{x}(t)$, $t = 0, 1, \cdots, n-1$, of $\boldsymbol{x} \sim \mathcal{N}_c(0, \boldsymbol{\Sigma})$, we wish to estimate $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ when $n \leq p(p-1)/2$ (# of unknowns), or n is comparable to p, and only a "small" fraction of the elements of $\boldsymbol{\Omega}$ are nonzero (sparse). Sparse high-dimensional real-valued GGMs (RGGMs) have been extensively studied ([2–7], others) motivated by applications to gene networks, fMRI, social networks, etc. Study of (proper) complex GGM originated with [1] which is restricted to low-dimensional settings ($n \gg p$). In the context of frequency-domain formulation of graphical modeling of real-valued time series in high-dimensional settings (also a motivation for this paper, as in [9]), proper CGGMs have been considered implicitly in [8, 9] using ℓ_1 penalty. In [10, 11] a graphical lasso approach based on an ℓ_1-penalized log-likelihood objective function has been investigated.

In this paper we consider the smoothly clipped absolute deviation (SCAD) penalty instead of ℓ_1-penalty, following [4]. The SCAD penalty was exploited for real graphical model selection in [4]; work on complex GGMs is lacking. SCAD penalty can produce sparse set of solution like lasso, and approximately unbiased coefficients for large coefficients, unlike lasso. But this penalty is nonconvex, unlike lasso. Sufficient conditions for consistency (convergence in the Frobenius norm of the estimator $\hat{\boldsymbol{\Omega}}$ of $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$) and specification of its rate of convergence are provided in this paper, following the corresponding real-valued results of [4] (see also [2, 3]). Sufficient conditions for sparsistency (the property that all parameters that are zero are actually estimated as zero with probability tending to one), following real-valued results of [4], are also investigated.

In Sec. 2 we introduce notation and present the system model. Theoretical analysis of the proposed SCAD-penalized estimator is presented in Sec. 3 where we state Theorem 1 (consistency, proved in Sec. 4) and Theorem 2 (sparsistency, proved in Sec. 5). A numerical example is presented in Sec. 6.

2. PRELIMINARIES AND BACKGROUND

2.1. Notation

Given $\boldsymbol{A} \in \mathbb{C}^{p \times p}$, we use $\phi_{\min}(\boldsymbol{A})$, $\phi_{\max}(\boldsymbol{A})$, $|\mathbf{A}|$, $\text{tr}(\boldsymbol{A})$ and $\text{etr}(\boldsymbol{A})$ to denote the minimum eigenvalue, maximum eigenvalue, determinant, trace, and exponential of trace of $\boldsymbol{A}$, respectively. For $\boldsymbol{B} \in \mathbb{C}^{p \times q}$, we denote the operator norm, the Frobenius norm and the vectorized ℓ_1 norm, respectively, as $\|\boldsymbol{B}\| = \sqrt{\phi_{\max}(\boldsymbol{B}^H \boldsymbol{B})}$, $\|\boldsymbol{B}\|_F = \sqrt{\text{tr}(\boldsymbol{B}^H \boldsymbol{B})}$ and $\|\boldsymbol{B}\|_1 = \sum_{i,j} |B_{ij}|$ where B_{ij} is the (i, j)-th element of $\boldsymbol{B}$. We also denote B_{ij} by $[\boldsymbol{B}]_{ij}$. Given $\boldsymbol{A} \in \mathbb{C}^{p \times p}$, $\boldsymbol{A}^+ = \text{diag}(\boldsymbol{A})$ is a diagonal matrix with the same diagonal as $\boldsymbol{A}$, and $\boldsymbol{A}^- = \boldsymbol{A} - \boldsymbol{A}^+$ is $\boldsymbol{A}$ with all its diagonal elements set to zero. We use $\boldsymbol{A}^{-*}$ for $(\boldsymbol{A}^*)^{-1}$ where $\boldsymbol{A}^*$ is the complex conjugate of $\boldsymbol{A}$, and $\boldsymbol{A}^{-\top}$ for $(\boldsymbol{A}^\top)^{-1}$. For $\boldsymbol{y}_n, \boldsymbol{x}_n \in \mathbb{C}^p$, $\boldsymbol{y}_n \asymp \boldsymbol{x}_n$ means that $\boldsymbol{y}_n = \mathcal{O}(\boldsymbol{x}_n)$ and $\boldsymbol{x}_n = \mathcal{O}(\boldsymbol{y}_n)$, where the latter means there exists $0 < M < \infty$ such that $\|\boldsymbol{x}_n\| \leq M \|\boldsymbol{y}_n\|$ $\forall n \geq 1$. The notation $\boldsymbol{y}_n = \mathcal{O}_P(\boldsymbol{x}_n)$ for random vectors $\boldsymbol{y}_n, \boldsymbol{x}_n \in \mathbb{C}^p$ means that for any $\varepsilon > 0$, there exists $0 < M < \infty$ such that $P(\|\boldsymbol{y}_n\| \leq M \|\boldsymbol{x}_n\|) \geq 1 - \varepsilon$ $\forall n \geq 1$.

2.2. SCAD-Penalized Log-Likelihood

Let $\boldsymbol{x}(t) \in \mathbb{C}^{p_n}$, $t = 0, 1, \cdots, n-1$, be n i.i.d. observations of $\boldsymbol{x} \sim \mathcal{N}_c(\mathbf{0}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma} \succ \mathbf{0}$. The dimension p_n of $\boldsymbol{x}(t)$ is a non-decreasing function of the sample size n. Define $\boldsymbol{X} = [\boldsymbol{x}(0)\ \boldsymbol{x}(1)\ \cdots\ \boldsymbol{x}(n-1)]^H$. Define the sample covariance matrix

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \boldsymbol{X}^H \boldsymbol{X} = \frac{1}{n} \sum_{t=0}^{n-1} \boldsymbol{x}(t) \boldsymbol{x}^H(t). \tag{1}$$

This work was supported by NSF Grants CCF-1617610 and ECCS-2040536.

978-1-7281-6662-9/20/$31.00 ©2020 IEEE

The joint pdf of $\boldsymbol{X}$ can be expressed as [10]

$$f_{\boldsymbol{X}}(\boldsymbol{X}) = \frac{|\boldsymbol{\Omega}|^n}{\pi^{np_n}} \text{etr}(-n\hat{\boldsymbol{\Sigma}}\boldsymbol{\Omega}) \tag{2}$$

$$= \frac{|\boldsymbol{\Omega}|^{n/2}|\boldsymbol{\Omega}^*|^{n/2}}{\pi^{np_n}} \text{etr}\left(-\frac{n}{2}(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Omega} + \hat{\boldsymbol{\Sigma}}^*\boldsymbol{\Omega}^*)\right) . \tag{3}$$

We wish to estimate $\boldsymbol{\Omega}$ given $\boldsymbol{X}$.

We have the negative log-likelihood (up to some constants)

$$-\ln f_{\boldsymbol{X}}(\boldsymbol{X}) = \text{tr}(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Omega} + \hat{\boldsymbol{\Sigma}}^*\boldsymbol{\Omega}^*) - \ln|\boldsymbol{\Omega}| - \ln|\boldsymbol{\Omega}^*| . \tag{4}$$

In [10, 11], an ℓ_1-penalized cost function $\mathcal{L}_1(.)$ is minimized,

$$\mathcal{L}_1(\boldsymbol{X};\boldsymbol{\Omega},\boldsymbol{\Omega}^*) = -\ln f_{\boldsymbol{X}}(\boldsymbol{X}) + \lambda_n \|\boldsymbol{\Omega}^-\|_1 , \tag{5}$$

where $\lambda_n > 0$ is a tuning parameter and we estimate $\boldsymbol{\Omega} \succ \boldsymbol{0}$. This is the lasso penalty, leading to the term graphical lasso [5]. An alternative [4] (in the real-valued case) is to use the SCAD penalty function $P_{\lambda_n}(|\Omega_{ij}|)$ to modify (5) as

$$\mathcal{L}_s(\boldsymbol{X};\boldsymbol{\Omega},\boldsymbol{\Omega}^*) = -\ln f_{\boldsymbol{X}}(\boldsymbol{X}) + \sum_{i \neq j} P_{\lambda_n}(|\Omega_{ij}|) , \tag{6}$$

where, for some $a > 2$, the SCAD penalty is defined as

$$P_\lambda(\theta) = \begin{cases} \lambda|\theta| & \text{for } |\theta| \le \lambda \\ \frac{2a\lambda|\theta| - |\theta|^2 - \lambda^2}{2(a-1)} & \text{for } \lambda < |\theta| < a\lambda \\ \frac{\lambda^2(a+1)}{2} & \text{for } |\theta| \ge a\lambda \end{cases} . \tag{7}$$

The first-order derivative of $P_\lambda(\theta)$ w.r.t. $|\theta|$ is

$$P'_\lambda(\theta) = \begin{cases} \lambda & \text{for } |\theta| \le \lambda \\ \frac{a\lambda - |\theta|}{a-1} & \text{for } \lambda < |\theta| < a\lambda \\ 0 & \text{for } |\theta| \ge a\lambda \end{cases} , \tag{8}$$

and its second-order derivative is

$$P''_\lambda(\theta) = \begin{cases} 0 & \text{for } |\theta| \le \lambda \\ \frac{-1}{a-1} & \text{for } \lambda < |\theta| < a\lambda \\ 0 & \text{for } |\theta| \ge a\lambda \end{cases} . \tag{9}$$

The SCAD penalty was proposed by [12] and exploited for real graphical model selection in [4]. As a function of θ, the SCAD penalty is continuously differentiable on $(-\infty, 0) \cup (0, \infty)$ but singular at 0, with its derivatives zero outside the range $[-a\lambda, a\lambda]$. Compared to lasso (ℓ_1 penalty), this results in small coefficients being set to zero, a few other coefficients being shrunk towards zero while leaving large coefficients unchanged. Thus, SCAD can produce sparse set of solution like lasso, and approximately unbiased coefficients for large coefficients, unlike lasso. But this penalty is nonconvex, unlike lasso.

Remark 1: Note that unlike the formulations in [5–7] where $\boldsymbol{\Omega}$ is real-valued, we have complex-valued $\boldsymbol{\Omega}$ (as in [11]). So, as in [11], we use Wirtinger calculus [13, Appendix 2] coupled with corresponding definition of subdifferential/subgradients [14], to analyze and minimize $\mathcal{L}_s(\boldsymbol{X};\boldsymbol{\Omega},\boldsymbol{\Omega}^*)$ w.r.t. complex $\boldsymbol{\Omega}$ using the necessary and sufficient conditions for a local optimum. Consider a complex-valued $\boldsymbol{z} = \boldsymbol{x} + j\boldsymbol{y} \in \mathbb{C}^p$, $\boldsymbol{x}, \boldsymbol{y}$ reals, and a real-valued scalar function $g(\boldsymbol{z}) = g(\boldsymbol{z}, \boldsymbol{z}^*) = g(\boldsymbol{x}, \boldsymbol{y})$. In Wirtinger calculus, one views $g(\boldsymbol{z}, \boldsymbol{z}^*)$ as a function of two independent vectors $\boldsymbol{z}$ and $\boldsymbol{z}^*$, instead of a function a single $\boldsymbol{z}$. For $g(\boldsymbol{z})$ one defines its subdifferential $\partial g(\boldsymbol{z}_0)$ at a point $\boldsymbol{z}_0$ as [14]

$$\partial g(\boldsymbol{z}_0) = \Big\{ \boldsymbol{s} \in \mathbb{C}^p \, : \, g(\boldsymbol{z}) \ge g(\boldsymbol{z}_0) + 2\,\text{Re}\left(\boldsymbol{s}^H(\boldsymbol{z} - \boldsymbol{z}_0)\right) \text{ for all } \boldsymbol{z} \in \mathbb{C}^p \Big\} . \tag{10}$$

Similarly, with $h_k(x) := g(z_1, z_2, \cdots, z_{k-1}, x, z_{k+1}, \cdots, z_p)$, $x \in \mathbb{C}$, the partial subdifferential $\partial g_{z_{0k}}(\boldsymbol{z}) := \partial h_k(z_{0k})$ is the subdifferential $\partial h_k(z_{0k})$ of $h_k(x)$ at z_{0k}. Also [14]

$$\partial g(\boldsymbol{z}_0) = \frac{\partial g(\boldsymbol{z})}{\partial \boldsymbol{z}^*}\Big|_{\boldsymbol{z}=\boldsymbol{z}_0} \tag{11}$$

when this partial derivative exists and g is convex. □

3. THEORETICAL ANALYSIS

We make following two assumptions.

(A1) Define the true edge set $\mathcal{E}_0 = \{\{i,j\} \, : \, \Omega_{0ij} \neq 0, \; i \neq j\}$ where $\boldsymbol{\Omega}_0 \in \mathbb{C}^{p_n \times p_n}$ denotes the true inverse covariance matrix. Then card$(\mathcal{E}_0) \le s_{n0}$.

(A2) The minimum and maximum eigenvalues of the true covariance matrix $\boldsymbol{\Sigma}_0 = \boldsymbol{\Omega}_0^{-1} \succ \boldsymbol{0}$ satisfy

$$0 < \beta_{\min} \le \phi_{\min}(\boldsymbol{\Sigma_0}) \le \phi_{\max}(\boldsymbol{\Sigma_0}) \le \beta_{\max} < \infty .$$

Here $\beta_{\min}$ and $\beta_{\max}$ are not functions of n.

Define the estimator $\hat{\boldsymbol{\Omega}}_\lambda$ as minimizer of (6)

$$\hat{\boldsymbol{\Omega}}_\lambda = \arg\min_{\boldsymbol{\Omega} \succ \boldsymbol{0}} \mathcal{L}_s(\boldsymbol{X};\boldsymbol{\Omega},\boldsymbol{\Omega}^*) . \tag{12}$$

Theorem 1 establishes convergence in the Frobenius norm of the estimator $\hat{\boldsymbol{\Omega}}_\lambda$ to the true value, and also provides a rate of convergence. It is proved in Sec. 4. Its proof follows that of [4, Theorem 1] (see also [2]) pertaining to real GGMs.

Theorem 1 (Consistency): For $\tau > 2$, let

$$C_0 = 80 \max_i(\Sigma_{0ii})\sqrt{2\left(\tau + \ln(16)/\ln(p_n)\right)} . \tag{13}$$

Given real numbers $\alpha_1 \in (0, 1)$ and "small' $\alpha_2 > 0$, let

$$M = (2 + \alpha_2)(1 + \alpha_1)^2 C_0/\beta_{\min}^2 . \tag{14}$$

Suppose the regularization parameter λ_n is selected as

$$\lambda_n = \max(2C_0, M) r_n , \tag{15}$$

where, with s_{n0} as in assumption (A1),

$$r_n = \sqrt{\frac{(p_n + s_{n0})\ln(p_n)}{n}} = o(1), \;\; \text{(i.e. } \to 0 \text{ as } n \to \infty\text{)}, \tag{16}$$

assumptions (A1)-(A2) hold true, and $\min_{\{i,j\}\in\mathcal{E}_0} |\Omega_{0ij}| \ge a\lambda_n$, with $a > 2$ as in (7). Let

$$N_1 = 2\left(\ln(16) + \tau\,\ln(p_n)\right) \tag{17}$$

$$N_2 = \arg\min\left\{ n \, : \, r_n \le \frac{\alpha_1\beta_{\min}}{(1+\alpha_1)^2(2+\alpha_2)C_0} \right\} \tag{18}$$

$$N_3 = \arg\min\left\{ n \, : \, \lambda_n < \frac{\min_{\{i,j\}\in\mathcal{E}_0} |\Omega_{0ij}|}{a} \right\} . \tag{19}$$

If the sample size $n > \max\{N_1, N_2, N_3\}$, then there exists a local minimizer $\hat{\boldsymbol{\Omega}}_\lambda$ of $\mathcal{L}_s(\boldsymbol{X};\boldsymbol{\Omega},\boldsymbol{\Omega}^*)$ such that

$$\|\hat{\boldsymbol{\Omega}}_\lambda - \boldsymbol{\Omega}_0\|_F \le M r_n \tag{20}$$

with probability $> 1 - 1/p_n^{\tau-2}$. In terms of rate of convergence,

$$\|\hat{\boldsymbol{\Omega}}_\lambda - \boldsymbol{\Omega}_0\|_F = \mathcal{O}_P\left(\sqrt{(p_n + s_{n0})\ln(p_n)/n}\right) \quad \bullet \tag{21}$$

Theorem 2 regarding sparsistency of $\hat{\boldsymbol{\Omega}}_\lambda$ is stated below. Its proof closely follows that of [11, Theorem 2] pertaining to ℓ_1 penalty, which, in turn, follows that of [4, Theorem 2] pertaining to real GGMs.

Theorem 2 (Sparsistency): Suppose all assumptions and conditions of Theorem 1 hold true so that $\|\hat{\boldsymbol{\Omega}}_\lambda - \boldsymbol{\Omega}_0\|_F = \mathcal{O}_P(r_n)$. In addition, suppose that there exists a sequence $\eta_n \to 0$ such that $\|\hat{\boldsymbol{\Omega}}_\lambda - \boldsymbol{\Omega}_0\| = \mathcal{O}_P(\eta_n)$ and $\sqrt{\ln(p_n)/n} + \eta_n = \mathcal{O}(\lambda_n)$. Then with probability tending to one, $\hat{\Omega}_{\lambda ij} = 0$ for all $(i,j) \in \bar{\mathcal{E}}_0 = \{\{i,j\} : \Omega_{0ij} = 0,\ i \neq j\}$. $\bullet$

Remark 2: Since $\|\hat{\boldsymbol{\Omega}}_\lambda - \boldsymbol{\Omega}_0\| \leq \|\hat{\boldsymbol{\Omega}}_\lambda - \boldsymbol{\Omega}_0\|_F$, the choice $\eta_n = r_n$ satisfies $\|\hat{\boldsymbol{\Omega}}_\lambda - \boldsymbol{\Omega}_0\| = \mathcal{O}_P(\eta_n)$ as well as $\sqrt{\ln(p_n)/n} + \eta_n = \mathcal{O}(\lambda_n)$. This allows for $s_{n0} = \mathcal{O}(p_n^2)$ so long as $r_n = \sqrt{(p_n + s_{n0})\ln(p_n)/n} = o(1)$. This result is significantly better than that in [11] where the best one can do is $s_{n0} = \mathcal{O}(p_n)$ or $s_{n0} = \mathcal{O}(1)$ (see [11, Remark 2]). □

4. PROOF OF THEOREM 1

First we need to introduce some notation, recall some existing results, and develop several auxiliary results. Lemmas 1-3 stated below are from [11]. Lemma 1, regarding a tail bound on the sample covariance of (1), follows from the real-valued results of [3, Lemma 1].

Lemma 1: Under Assumption (A2), $\hat{\boldsymbol{\Sigma}}$ defined in (1) satisfies the tail bound

$$P\left(\max_{k,l}\left|[\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_0]_{kl}\right| > C_0\sqrt{\frac{\ln(p_n)}{n}}\right) \leq \frac{1}{p_n^{\tau-2}} \tag{22}$$

for $\tau > 2$, if the sample size $n > N_1$, where C_0 is defined in (13) and N_1 is defined in (17). $\bullet$

Lemma 2 deals with a Taylor series expansion using Wirtinger calculus; its proof is omitted for lack of space.

Lemma 2: For $\boldsymbol{\Omega} = \boldsymbol{\Omega}^H \succ \mathbf{0}$, define a real scalar function

$$c(\boldsymbol{\Omega}, \boldsymbol{\Omega}^*) = \ln|\boldsymbol{\Omega}| + \ln|\boldsymbol{\Omega}^*|\,. \tag{23}$$

Let $\boldsymbol{\Omega} = \boldsymbol{\Omega}_0 + \boldsymbol{\Delta}$ with $\boldsymbol{\Omega}_0 = \boldsymbol{\Omega}_0^H \succ \mathbf{0}$ and $\boldsymbol{\Delta} = \boldsymbol{\Delta}^H$. Then using Wirtinger calculus, the Taylor series expansion of $c(\boldsymbol{\Omega}, \boldsymbol{\Omega}^*)$ is given by

$$\begin{aligned} c(\boldsymbol{\Omega}, \boldsymbol{\Omega}^*) &= c(\boldsymbol{\Omega}_0, \boldsymbol{\Omega}_0^*) + \mathrm{tr}(\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Delta} + \boldsymbol{\Omega}_0^{-*}\boldsymbol{\Delta}^*) \\ &- \frac{1}{2}(\mathrm{vec}(\boldsymbol{\Delta}))^H(\boldsymbol{\Omega}_0^{-*} \otimes \boldsymbol{\Omega}_0^{-1})\mathrm{vec}(\boldsymbol{\Delta}) \\ &- \frac{1}{2}(\mathrm{vec}(\boldsymbol{\Delta}^*))^H(\boldsymbol{\Omega}_0^{-1} \otimes \boldsymbol{\Omega}_0^{-*})\mathrm{vec}(\boldsymbol{\Delta}^*) + \mathrm{h.o.t.} \end{aligned} \tag{24}$$

where h.o.t. stands for higher-order terms in $\boldsymbol{\Delta}$ and $\boldsymbol{\Delta}^*$. $\bullet$

Lemma 2 regarding Taylor series expansion immediately leads to Lemma 3 regarding Taylor series with integral remainder, needed to follow the proof of [2,4] pertaining to the real-valued case.

Lemma 3: With $c(\boldsymbol{\Omega}, \boldsymbol{\Omega}^*)$ and $\boldsymbol{\Omega} = \boldsymbol{\Omega}_0 + \boldsymbol{\Delta}$ as in Lemma 2, the Taylor series expansion of $c(\boldsymbol{\Omega}, \boldsymbol{\Omega}^*)$ in integral remainder form is given by (v is real)

$$\begin{aligned} c(\boldsymbol{\Omega}, \boldsymbol{\Omega}^*) &= c(\boldsymbol{\Omega}_0, \boldsymbol{\Omega}_0^*) + \mathrm{tr}(\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Delta} + \boldsymbol{\Omega}_0^{-*}\boldsymbol{\Delta}^*) \\ &- \boldsymbol{g}^H(\boldsymbol{\Delta})\left(\int_0^1 (1-v)\boldsymbol{H}(\boldsymbol{\Omega}_0, \boldsymbol{\Delta}, v)\,dv\right)\boldsymbol{g}(\boldsymbol{\Delta}) \end{aligned} \tag{25}$$

where

$$\boldsymbol{g}(\boldsymbol{\Delta}) = \begin{bmatrix} \mathrm{vec}(\boldsymbol{\Delta}) \\ \mathrm{vec}(\boldsymbol{\Delta}^*) \end{bmatrix}, \quad \boldsymbol{H}(\boldsymbol{\Omega}_0, \boldsymbol{\Delta}, v) = \begin{bmatrix} \boldsymbol{H}_{11} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{H}_{22} \end{bmatrix} \tag{26}$$

$$\boldsymbol{H}_{11} = (\boldsymbol{\Omega}_0 + v\boldsymbol{\Delta})^{-*} \otimes (\boldsymbol{\Omega}_0 + v\boldsymbol{\Delta})^{-1} \tag{27}$$

and

$$\boldsymbol{H}_{22} = (\boldsymbol{\Omega}_0 + v\boldsymbol{\Delta})^{-1} \otimes (\boldsymbol{\Omega}_0 + v\boldsymbol{\Delta})^{-*} \quad \bullet \tag{28}$$

We now turn to the proof of Theorem 1.

Proof of Theorem 1: Let

$$Q(\boldsymbol{\Omega}) := \mathcal{L}_s(\boldsymbol{X}; \boldsymbol{\Omega}, \boldsymbol{\Omega}^*) - \mathcal{L}_s(\boldsymbol{X}; \boldsymbol{\Omega}_0, \boldsymbol{\Omega}_0^*)\,. \tag{29}$$

By (4), (6) and (23), we have

$$\begin{aligned} Q(\boldsymbol{\Omega}) &= \mathrm{tr}\left(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Omega} + \hat{\boldsymbol{\Sigma}}^*\boldsymbol{\Omega}^*\right) - c(\boldsymbol{\Omega}, \boldsymbol{\Omega}^*) + \sum_{i \neq j} P_{\lambda_n}(|\Omega_{ij}|) \\ &-\mathrm{tr}\left(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Omega}_0 + \hat{\boldsymbol{\Sigma}}^*\boldsymbol{\Omega}_0^*\right) + c(\boldsymbol{\Omega}_0, \boldsymbol{\Omega}_0^*) - \sum_{i \neq j} P_{\lambda_n}(|\Omega_{0ij}|)\,. \end{aligned} \tag{30}$$

The estimate $\hat{\boldsymbol{\Omega}}_\lambda$, denoted by $\hat{\boldsymbol{\Omega}}$ hereafter suppressing dependence upon λ, minimizes $Q(\boldsymbol{\Omega})$, or equivalently, $\hat{\boldsymbol{\Delta}} = \hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0$ minimizes $G(\boldsymbol{\Delta}) := Q(\boldsymbol{\Omega}_0 + \boldsymbol{\Delta})$.

We will follow, for the most part, the proof of [4, Theorem 1] pertaining to real GGMs. Consider the set

$$\Theta_n(M) := \left\{\boldsymbol{\Delta} \,:\, \boldsymbol{\Delta} = \boldsymbol{\Delta}^H,\ \|\boldsymbol{\Delta}\|_F = Mr_n\right\} \tag{31}$$

where M is a constant defined in (14), and

$$r_n = \sqrt{(p_n + s_{n0})\ln(p_n)/n}, \quad \lim_{n\to\infty} r_n = 0\,. \tag{32}$$

Observe that $G(\hat{\boldsymbol{\Delta}}) \leq G(\mathbf{0}) = 0$. Therefore, if we can show that

$$\inf_{\boldsymbol{\Delta}}\{G(\boldsymbol{\Delta}) \,:\, \boldsymbol{\Delta} \in \Theta_n(M)\} > 0\,, \tag{33}$$

the minimizer $\hat{\boldsymbol{\Delta}}$ must be inside $\Theta_n(M)$, and hence

$$\|\hat{\boldsymbol{\Delta}}\|_F \leq Mr_n\,. \tag{34}$$

Using Lemma 3 and noting that $\boldsymbol{\Omega}^{-1} = \boldsymbol{\Sigma}$, we rewrite $G(\boldsymbol{\Delta})$ as

$$G(\boldsymbol{\Delta}) = A_1 + A_2 + A_3\,, \tag{35}$$

$$A_1 = \mathrm{tr}\left((\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_0)\boldsymbol{\Delta} + (\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_0)^*\boldsymbol{\Delta}^*\right)\,, \tag{36}$$

$$A_2 = \boldsymbol{g}^H(\boldsymbol{\Delta})\left(\int_0^1 (1-v)\boldsymbol{H}(\boldsymbol{\Omega}_0, \boldsymbol{\Delta}, v)\,dv\right)\boldsymbol{g}(\boldsymbol{\Delta})\,, \tag{37}$$

$$A_3 = \sum_{i \neq j}\left(P_{\lambda_n}(|\Omega_{0ij} + \Delta_{ij}|) - P_{\lambda_n}(|\Omega_{0ij}|)\right)\,. \tag{38}$$

By the structure of Hermitian $\boldsymbol{H}(\boldsymbol{\Omega}_0, \boldsymbol{\Delta}, v)$, we have

$$\begin{aligned} &\phi_{\min}(\boldsymbol{H}(\boldsymbol{\Omega}_0, \boldsymbol{\Delta}, v)) = \phi_{\min}(\boldsymbol{H}_{11}) = \phi_{\min}(\boldsymbol{H}_{22}) \\ &\quad = \phi_{\min}^2((\boldsymbol{\Omega}_0 + v\boldsymbol{\Delta})^{-1}) = \phi_{\max}^{-2}(\boldsymbol{\Omega}_0 + v\boldsymbol{\Delta})\,. \end{aligned} \tag{39}$$

Since $\boldsymbol{x}^H\boldsymbol{A}\boldsymbol{x} \geq \phi_{\min}(\boldsymbol{A})\|\boldsymbol{x}\|^2$, we have

$$\begin{aligned} A_2 &\geq \|\boldsymbol{g}(\boldsymbol{\Delta})\|^2\phi_{\min}\left(\int_0^1 (1-v)\boldsymbol{H}(\boldsymbol{\Omega}_0, \boldsymbol{\Delta}, v)\,dv\right) \\ &\geq 2\|\mathrm{vec}(\boldsymbol{\Delta})\|^2\int_0^1 (1-v)\,dv \min_{0\leq v\leq 1}\phi_{\min}(\boldsymbol{H}(\boldsymbol{\Omega}_0, \boldsymbol{\Delta}, v)) \\ &= \|\boldsymbol{\Delta}\|_F^2 \min_{0\leq v\leq 1}\phi_{\max}^{-2}(\boldsymbol{\Omega}_0 + v\boldsymbol{\Delta})\,, \end{aligned} \tag{40}$$

where we have used the facts that $\int_0^1 (1-v)\,dv = 1/2$. Since

$$\phi_{\max}(\mathbf{\Omega}_0 + v\mathbf{\Delta}) \le \|\mathbf{\Omega}_0 + v\mathbf{\Delta}\| \le \|\mathbf{\Omega}_0\| + v\|\mathbf{\Delta}\|\,, \tag{41}$$

we have, for $0 \le v \le 1$,

$$\phi_{\max}^{-2}(\mathbf{\Omega}_0 + v\mathbf{\Delta}) \ge (\|\mathbf{\Omega}_0\| + v\|\mathbf{\Delta}\|)^{-2} \ge (\|\mathbf{\Omega}_0\| + \|\mathbf{\Delta}\|)^{-2}\,. \tag{42}$$

Thus,

$$A_2 \ge \frac{\|\mathbf{\Delta}\|_F^2}{(\|\mathbf{\Omega}_0\| + \|\mathbf{\Delta}\|)^2} \ge \|\mathbf{\Delta}\|_F^2 \left(\beta_{\min}^{-1} + M r_n\right)^{-2} \tag{43}$$

where we have used the fact that $\|\mathbf{\Omega}_0\| = \|\mathbf{\Sigma}_0^{-1}\| = \phi_{\max}(\mathbf{\Sigma}_0^{-1}) = (\phi_{\min}(\mathbf{\Sigma}_0))^{-1} \le \beta_{\min}^{-1}$ and $\|\mathbf{\Delta}\| \le \|\mathbf{\Delta}\|_F = M r_n = \mathcal{O}(r_n)$.

We now consider A_1 in (36). Define the set $\tilde{\mathcal{E}}_0 = \mathcal{E}_0 \cup \{\{i,j\} : i = j\}$. Let $\bar{\mathcal{E}}_0$ denote the complement of $\tilde{\mathcal{E}}_0$, given by $\bar{\mathcal{E}}_0 = \{\{i,j\} : \Omega_{0ij} = 0,\ i \neq j\}$. Also, for an index set $\boldsymbol{B}$ and a matrix $\boldsymbol{C} \in \mathbb{C}^{p\times p}$, we write $\boldsymbol{C}_{\boldsymbol{B}}$ to denote a matrix in $\mathbb{C}^{p\times p}$ such that $[\boldsymbol{C}_{\boldsymbol{B}}]_{ij} = C_{ij}$ if $(i,j) \in \boldsymbol{B}$, and $[\boldsymbol{C}_{\boldsymbol{B}}]_{ij} = 0$ if $(i,j) \notin \boldsymbol{B}$. Then, by definition, $\mathbf{\Delta}^- = \mathbf{\Delta}^-_{\mathcal{E}_0} + \mathbf{\Delta}^-_{\bar{\mathcal{E}}_0}$, and $\|\mathbf{\Delta}^-\|_1 = \|\mathbf{\Delta}^-_{\mathcal{E}_0}\|_1 + \|\mathbf{\Delta}^-_{\bar{\mathcal{E}}_0}\|_1$. We have

$$A_1 = L_1 + L_2 \tag{44}$$

where (note that tr$((\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_0)\mathbf{\Omega})$ = tr$((\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_0)^*\mathbf{\Omega}^*)$)

$$L_1 = 2\sum_{\{i,j\}\in\mathcal{E}_0} [\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_0]_{ij}\Delta_{ji} + 2\sum_i [\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_0]_{ii}\Delta_{ii}\,, \tag{45}$$

$$L_2 = 2\sum_{\{i,j\}\in\bar{\mathcal{E}}_0} [\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_0]_{ij}\Delta_{ji}\,. \tag{46}$$

To bound L_1, using Lemma 1, with probability $> 1 - 1/p_n^{\tau-2}$,

$$\begin{aligned} |L_1| &\le 2\|\mathbf{\Delta}^-_{\mathcal{E}_0} + \mathbf{\Delta}^+\|_1 \max_{i,j} \left|[\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_0]_{ij}\right| \\ &\le 2\|\mathbf{\Delta}^-_{\mathcal{E}_0} + \mathbf{\Delta}^+\|_1\, C_0\sqrt{\ln(p_n)/n}\,. \end{aligned} \tag{47}$$

By Cauchy-Schwartz inequality, with $g_n := \sqrt{s_{n0} + p_n}$,

$$\|\mathbf{\Delta}^-_{\mathcal{E}_0} + \mathbf{\Delta}^+\|_1 \le g_n\, \|\mathbf{\Delta}^-_{\mathcal{E}_0} + \mathbf{\Delta}^+\|_F \le g_n\, \|\mathbf{\Delta}\|_F\,. \tag{48}$$

Therefore,

$$|L_1| \le 2C_0\, \|\mathbf{\Delta}\|_F \sqrt{(p_n + s_{n0})\ln(p_n)/n}\,. \tag{49}$$

We will consider L_2 with part of A_3, where

$$A_3 = L_3 + L_4\,, \tag{50}$$

$$L_3 = \sum_{\{i,j\}\in\mathcal{E}_0} \left(P_{\lambda_n}(|\Omega_{0ij} + \Delta_{ij}|) - P_{\lambda_n}(|\Omega_{0ij}|)\right)\,, \tag{51}$$

$$L_4 = \sum_{\{i,j\}\in\bar{\mathcal{E}}_0} P_{\lambda_n}(|\Delta_{ij}|)\,, \tag{52}$$

and we have used the fact that $\Omega_{0ij} = 0$ for $\{i,j\} \in \bar{\mathcal{E}}_0$ and $P_{\lambda_n}(0) = 0$. Since $\|\mathbf{\Delta}\|_F = M r_n$, we must have $|\Delta_{ij}| \le M r_n$. For $\lambda_n \ge M r_n$, $P_{\lambda_n}(|\Delta_{ij}|) = \lambda_n|\Delta_{ij}|$ for $|\Delta_{ij}| \le M r_n$. Consider L_4 with L_2. By Lemma 1, with probability $> 1 - 1/p_n^{\tau-2}$,

$$\begin{aligned} L_4 - |L_2| &\ge \sum_{\{i,j\}\in\bar{\mathcal{E}}_0} \left(\lambda_n|\Delta_{ij}| - 2|[\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_0]_{ij}| \cdot |\Delta^*_{ij}|\right) \\ &\ge \sum_{\{i,j\}\in\bar{\mathcal{E}}_0} \left(\lambda_n - 2C_0\sqrt{\ln(p_n)/n}\right)|\Delta_{ij}| \\ &= \lambda_n \left(1 - 2\frac{C_0}{\lambda_n}\sqrt{\ln(p_n)/n}\right) \sum_{\{i,j\}\in\bar{\mathcal{E}}_0} |\Delta_{ij}| > 0 \end{aligned} \tag{53}$$

since $\frac{2C_0}{\lambda_n}\sqrt{\ln(p_n)/n} < 1$ for λ_n selected as in (15).

It remains to bound $|L_3|$. A Taylor series expansion of $P_{\lambda_n}(\theta)$ for $\theta > 0$, around $\theta_0 > 0$, is given by

$$P_{\lambda_n}(\theta) = P_{\lambda_n}(\theta_0) + P'_{\lambda_n}(\theta_0)(\theta - \theta_0) + P''_{\lambda_n}(\tilde{\theta})\frac{(\theta-\theta_0)^2}{2} \tag{54}$$

where $\tilde{\theta} = \theta_0 + \gamma(\theta - \theta_0)$ for some $\gamma \in [0,1]$. Setting $\theta_0 = |\Omega_{0ij}|$ and $\theta = |\Omega_{0ij} + \Delta_{ij}|$, and noting that $P''_{\lambda_n}(\tilde{\theta}) \le 0$ for any $\tilde{\theta} > 0$, and $|\Omega_{0ij}| > 0$ for $\{i,j\} \in \mathcal{E}_0$, we have

$$\begin{aligned} P_{\lambda_n}(|\Omega_{0ij} + \Delta_{ij}|) &\le P_{\lambda_n}(|\Omega_{0ij}|) \\ &\quad + P'_{\lambda_n}(|\Omega_{0ij}|)(|\Omega_{0ij} + \Delta_{ij}| - |\Omega_{0ij}|)\,. \end{aligned} \tag{55}$$

Therefore

$$\begin{aligned} |L_3| &\le \sum_{\{i,j\}\in\mathcal{E}_0} \left|P'_{\lambda_n}(|\Omega_{0ij}|)\right| \cdot \left||\Omega_{0ij} + \Delta_{ij}| - |\Omega_{0ij}|\right| \\ &\le \sum_{\{i,j\}\in\mathcal{E}_0} \left|P'_{\lambda_n}(|\Omega_{0ij}|)\right| \cdot |\Delta_{ij}| = 0 \text{ for } n \ge N_3\,, \end{aligned} \tag{56}$$

where we have used the properties of $P'_\lambda(\theta)$ for $\theta > 0$.

Using the bounds on $A_1 = L_1 + L_2$, A_2, and $A_3 = L_3 + L_4$, with probability $> 1 - 1/p_n^{\tau-2}$, we have

$$\begin{aligned} G(\mathbf{\Delta}) &\ge -|L_1| + L_4 - |L_2| - |L_3| \ + \|\mathbf{\Delta}\|_F^2 \left(\beta_{\min}^{-1} + M r_n\right)^{-2} \\ &\ge -2C_0\, \|\mathbf{\Delta}\|_F \sqrt{\frac{(p_n + s_{n0})\ln(p_n)}{n}} + \|\mathbf{\Delta}\|_F^2 \left(\beta_{\min}^{-1} + M r_n\right)^{-2} \end{aligned} \tag{57}$$

where we have used the fact that $L_4 - |L_2| > 0$, and $|L_3| = 0$ for large n. Using $\|\mathbf{\Delta}\|_F = M r_n = M\sqrt{(p_n + s_{n0})\ln(p_n)/n}$,

$$G(\mathbf{\Delta}) \ge \|\mathbf{\Delta}\|_F^2 \left[\left(\beta_{\min}^{-1} + M r_n\right)^{-2} - 2\frac{C_0}{M}\right]\,. \tag{58}$$

For $n \ge N_2$, if we pick M as specified in (14), we obtain $M r_n \le M r_{N_2} \le \alpha_1/\beta_{\min}$. It then follows that

$$\left(\beta_{\min}^{-1} + M r_n\right)^{-2} \ge \frac{\beta_{\min}^2}{(1+\alpha_1)^2} = \frac{(2+\alpha_2)C_0}{M} > 2\frac{C_0}{M}\,,$$

implying $G(\mathbf{\Delta}) > 0$. This proves the desired result. ■

5. PROOF OF THEOREM 2

As noted in Remark 1, the notation $\partial_{\hat{\Omega}_{\lambda ik}}\mathcal{L}(\mathbf{X};\mathbf{\Omega},\mathbf{\Omega}^*)$ denotes the partial subdifferential of $\mathcal{L}_s(\mathbf{X};\mathbf{\Omega},\mathbf{\Omega}^*)$ at the (i,k)th component $\Omega_{\lambda ik} = \hat{\Omega}_{\lambda ik}$. When the corresponding partial derivative exists, we have

$$\partial_{\hat{\Omega}_{\lambda ik}}\mathcal{L}_s(\mathbf{X};\mathbf{\Omega},\mathbf{\Omega}^*) = \frac{\partial\mathcal{L}_s(\mathbf{X};\mathbf{\Omega},\mathbf{\Omega}^*)}{\partial\Omega^*_{ik}}\Big|_{\mathbf{\Omega}=\hat{\mathbf{\Omega}}_\lambda}\,.$$

Using (4), (6) and (7), we have

$$\begin{aligned} \partial_{\hat{\Omega}_{\lambda ik}}\mathcal{L}_s(\mathbf{X};\mathbf{\Omega},\mathbf{\Omega}^*) &= \hat{\Sigma}^*_{ki} - [\hat{\mathbf{\Omega}}_\lambda^{-*}]_{ki} + \frac{t}{2}P'_{\lambda_n}(|\hat{\Omega}_{\lambda ik}|) \\ &= \hat{\Sigma}_{ik} - \check{\Sigma}_{\lambda ik} + \frac{t}{2}P'_{\lambda_n}(|\hat{\Omega}_{\lambda ik}|) \text{ where } \check{\mathbf{\Sigma}}_\lambda := \hat{\mathbf{\Omega}}_\lambda^{-1} \end{aligned} \tag{59}$$

and t is given by

$$t = \begin{cases} \hat{\Omega}_{\lambda ik}/|\hat{\Omega}_{\lambda ik}| =: \text{sign}(\hat{\Omega}_{\lambda ik}) & \text{if } \hat{\Omega}_{\lambda ik} \neq 0 \\ \in \{u : |u| \le 1,\ u \in \mathbb{C}\} & \text{if } \hat{\Omega}_{\lambda ik} = 0\,. \end{cases} \tag{60}$$

To prove the desired result, the term $\frac{P'_{\lambda_n}(|\hat{\Omega}_{\lambda ik}|)}{2}\hat{\Omega}_{\lambda ik}/|\hat{\Omega}_{\lambda ik}|$ on the right-side of (59) must dominate the term $\hat{\Sigma}_{ik} - \check{\Sigma}_{\lambda ik}$ whenever true value $\Omega_{0ij} = 0$. Note that for Ω_{ik} in a small neighborhood of 0 (but excluding 0), $P_{\lambda_n}(|\Omega_{ik}|) = \lambda_n|\Omega_{ik}| > 0$ and $P'_{\lambda_n}(|\Omega_{ik}|) = \lambda_n > 0$. Then sign of the left-side of (59) is the same as $\text{sign}(\hat{\Omega}_{\lambda ik})$ with probability tending to one, which yields the desired result, as is shown in what follows. At the optimal solution, by the KKT conditions, one must have (59) equal to zero. Suppose that for $(i,k) \in \bar{\mathcal{E}}_0$, one has $\hat{\Omega}_{ik} \neq 0$ when $\hat{\Sigma}_{ik} - \check{\Sigma}_{\lambda ik} + \frac{P'_{\lambda_n}(|\hat{\Omega}_{\lambda ik}|)}{2}\hat{\Omega}_{\lambda ik}/|\hat{\Omega}_{\lambda ik}| = 0$. Therefore, with $\hat{\Omega}_{\lambda ikR} = \text{Re}(\hat{\Omega}_{\lambda ik})$ and $\hat{\Omega}_{\lambda ikI} = \text{Im}(\hat{\Omega}_{\lambda ik})$, $\hat{\Omega}_{ik} \neq 0$ implies that $\hat{\Omega}_{\lambda ikR} \neq 0$ and/or $\hat{\Omega}_{\lambda ikI} \neq 0$. Suppose that $\hat{\Omega}_{\lambda ikR} < 0$, implying that for some $\delta > 0$, $\hat{\Omega}_{\lambda ikR} + \delta < 0$, since, by Theorem 1, $\hat{\Omega}_{\lambda ikR}$ converges to $\Omega_{0ikR} = 0$ for $(i,j) \in \bar{\mathcal{E}}_0$. Since $\hat{\Omega}_{\lambda ijR}$ minimizes $\mathcal{L}_s(\mathbf{X}; \boldsymbol{\Omega}, \boldsymbol{\Omega}^*)$, and $\partial_{\hat{\Omega}_{\lambda ik}}\mathcal{L}_s(\mathbf{X}; \boldsymbol{\Omega}, \boldsymbol{\Omega}^*) = \frac{1}{2}\left(\frac{\partial\mathcal{L}_s(\mathbf{X};\boldsymbol{\Omega},\boldsymbol{\Omega}^*)}{\partial\hat{\Omega}_{\lambda ikR}} + j\frac{\partial\mathcal{L}_s(\mathbf{X};\boldsymbol{\Omega},\boldsymbol{\Omega}^*)}{\partial\hat{\Omega}_{\lambda ikI}}\right) = 0$, we have $I_1 := \frac{\partial\mathcal{L}_s(\mathbf{X};\boldsymbol{\Omega},\boldsymbol{\Omega}^*)}{\partial(\hat{\Omega}_{\lambda ikR}+\delta)} > 0$ for $\delta > 0$. If λ_n (=$P'_{\lambda_n}(|\hat{\Omega}_{\lambda ik}|)$) dominates $\hat{\Sigma}_{ik} - \check{\Sigma}_{\lambda ik}$ in (59), $I_1 > 0$ implies that $\hat{\Omega}_{\lambda ikR} + \delta > 0$, contradicting the assumption that $\hat{\Omega}_{\lambda ikR} + \delta < 0$. Therefore, $\hat{\Omega}_{\lambda ikR} \not< 0$. Now suppose that $\hat{\Omega}_{\lambda ikR} > 0$, implying that for some $\delta > 0$, $\hat{\Omega}_{\lambda ikR} - \delta > 0$, since, by Theorem 1, $\hat{\Omega}_{\lambda ikR}$ converges to $\Omega_{0ikR} = 0$ for $(i,j) \in \bar{\mathcal{E}}_0$. Since $\hat{\Omega}_{\lambda ijR}$ minimizes $\mathcal{L}_s(\mathbf{X}; \boldsymbol{\Omega}, \boldsymbol{\Omega}^*)$, hence $\frac{\partial\mathcal{L}_s(\mathbf{X};\boldsymbol{\Omega},\boldsymbol{\Omega}^*)}{\partial\hat{\Omega}_{\lambda ikR}} = 0$, we have $I_2 := \frac{\partial\mathcal{L}_s(\mathbf{X};\boldsymbol{\Omega},\boldsymbol{\Omega}^*)}{\partial(\hat{\Omega}_{\lambda ikR}-\delta)} < 0$ for $\delta > 0$. If λ_n (=$P'_{\lambda_n}(|\hat{\Omega}_{\lambda ik}|)$) dominates $\hat{\Sigma}_{ik} - \check{\Sigma}_{\lambda ik}$ in (59), $I_2 < 0$ implies that $\hat{\Omega}_{\lambda ikR} - \delta < 0$, contradicting the assumption that $\hat{\Omega}_{\lambda ikR} - \delta > 0$. Therefore, $\hat{\Omega}_{\lambda ikR} = 0$ for $(i,k) \in \bar{\mathcal{E}}_0$, with probability tending to one. Similar arguments prove that $\hat{\Omega}_{\lambda ikI} = 0$ for $(i,k) \in \bar{\mathcal{E}}_0$, thus $\hat{\Omega}_{\lambda ik} = 0$, with probability tending to one, if λ_n dominates $\hat{\Sigma}_{ik} - \check{\Sigma}_{\lambda ik}$.

It remains to investigate the conditions under which λ_n (=$P'_{\lambda_n}(|\hat{\Omega}_{\lambda ik}|)$ for small nonzero $\hat{\Omega}_{\lambda ik}$) dominates $\hat{\Sigma}_{ik} - \check{\Sigma}_{\lambda ik}$. Rewrite

$$\hat{\Sigma}_{ik} - \check{\Sigma}_{\lambda ik} = \underbrace{\hat{\Sigma}_{ik} - \Sigma_{0ik}}_{=:I_3} + \underbrace{\Sigma_{0ik} - \check{\Sigma}_{\lambda ik}}_{=:I_4} . \tag{61}$$

By Lemma 2, $\max_{i,k}|I_3| = \mathcal{O}_P\left(\sqrt{\frac{\ln(p_n)}{n}}\right)$. By [4, Lemma 1],

$$\begin{aligned} |I_4| \leq& \|\boldsymbol{\Sigma}_0 - \check{\boldsymbol{\Sigma}}_\lambda\| = \|\check{\boldsymbol{\Sigma}}_\lambda(\hat{\boldsymbol{\Omega}}_\lambda - \boldsymbol{\Omega}_0)\boldsymbol{\Sigma}_0\| \\ \leq& \|\check{\boldsymbol{\Sigma}}_\lambda\| \cdot \|(\hat{\boldsymbol{\Omega}}_\lambda - \boldsymbol{\Omega}_0)\| \cdot \|\boldsymbol{\Sigma}_0\| . \end{aligned} \tag{62}$$

By Assumption (A2), $\|\boldsymbol{\Sigma}_0\| = \mathcal{O}(1)$. Furthermore,

$$\begin{aligned} \|\check{\boldsymbol{\Sigma}}_\lambda\| =& \|\hat{\boldsymbol{\Omega}}_\lambda^{-1}\| = \phi_{\min}^{-1}(\hat{\boldsymbol{\Omega}}_\lambda) \\ \leq& \left(\phi_{\min}(\boldsymbol{\Omega}_0) + \phi_{\min}(\hat{\boldsymbol{\Omega}}_\lambda - \boldsymbol{\Omega}_0)\right)^{-1} \\ =& (\mathcal{O}_P(1) + \mathcal{O}_P(\eta_n))^{-1} = \mathcal{O}_P(1) , \end{aligned} \tag{63}$$

where we have used the fact that since $\|\hat{\boldsymbol{\Omega}}_\lambda - \boldsymbol{\Omega}_0\| = \mathcal{O}_P(\eta_n)$, $\phi_{\min}(\hat{\boldsymbol{\Omega}}_\lambda - \boldsymbol{\Omega}_0) \leq \|\hat{\boldsymbol{\Omega}}_\lambda - \boldsymbol{\Omega}_0\| = \mathcal{O}_P(\eta_n)$, and by Weyl's inequality, $\phi_{\min}(\boldsymbol{A} + \boldsymbol{B}) \geq \phi_{\min}(\boldsymbol{A}) + \phi_{\min}(\boldsymbol{B})$. Hence,

$$\max_{i,k}|I_4| = \mathcal{O}_P\left(\|\hat{\boldsymbol{\Omega}}_\lambda - \boldsymbol{\Omega}_0\|\right) = \mathcal{O}_P(\eta_n) . \tag{64}$$

It then follows that

$$|\hat{\Sigma}_{ik} - \check{\Sigma}_{\lambda ik}| \leq |I_3| + |I_4| = \mathcal{O}_P\left(\sqrt{\frac{\ln(p_n)}{n}} + \eta_n\right) . \tag{65}$$

Suppose $\mathcal{O}(\lambda_n) = \sqrt{\ln(p_n)/n} + \eta_n$. Then $\frac{\lambda_n}{2}\hat{\Omega}_{\lambda ik}/|\hat{\Omega}_{\lambda ik}|$ (=$\frac{P'_{\lambda_n}(|\hat{\Omega}_{\lambda ik}|)}{2}\hat{\Omega}_{\lambda ik}/|\hat{\Omega}_{\lambda ik}|$) dominates $|\hat{\Sigma}_{ik} - \check{\Sigma}_{\lambda ik}|$ with probability tending to one. This completes the proof. ■

6. SIMULATION EXAMPLE

We use the example from [11] (similar to a real-valued example in [7]), where Hermitian $\breve{\boldsymbol{\Omega}}$ has ones on its diagonal. With probability q, off-diagonal elements of $\breve{\boldsymbol{\Omega}}$ are complex random variables with independent real and imaginary parts, uniformly distributed over $\{[-0.4\ \ -0.1] \cup [0.1\ 0.4]\}$, otherwise zero (probability $1-q$). Set $\boldsymbol{\Omega} = \breve{\boldsymbol{\Omega}} + \beta\mathbf{I}$ with β picked to make $\boldsymbol{\Omega} \succ \mathbf{0}$. With $\boldsymbol{\Phi}\boldsymbol{\Phi}^H = \boldsymbol{\Omega}^{-1}$, we generate $\mathbf{x} = \boldsymbol{\Phi}\mathbf{w}$ with $\mathbf{w} \sim \mathcal{N}_c(0, 2\mathbf{I})$. As in [11], we pick $q = 0.05$, leading to a sparse Erdös-Rényi graph with only about 5% of connected edges.

We used an AM (alternating minimization) method [15, 16] based on variable splitting and penalty technique as detailed in [10, 11], using two "passes". As recommended in [4], since SCAD penalty is nonconvex, in first pass use ℓ_1 penalty ($\rho > 0$ is "large") to estimate $\boldsymbol{\Omega}$ and $\boldsymbol{W}$ via

$$\begin{aligned} (\hat{\boldsymbol{\Omega}}^{(1)}, \hat{\boldsymbol{W}}^{(1)}) = \arg\min_{\boldsymbol{\Omega}\succ 0, \boldsymbol{W}} \Big\{&\text{tr}(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Omega} + \hat{\boldsymbol{\Sigma}}^*\boldsymbol{\Omega}^*) - \ln(|\boldsymbol{\Omega}|) \\ &- \ln(|\boldsymbol{\Omega}^*|) + \lambda_n \|\boldsymbol{W}^-\|_1 + \rho\|\boldsymbol{\Omega} - \boldsymbol{W}\|_F^2\Big\} . \end{aligned} \tag{66}$$

This estimator is globally convergent [10, 11]. Then linearize the SCAD penalty around $\boldsymbol{\Omega} = \hat{\boldsymbol{W}}^{(1)}$ and solve the convex problem

$$\begin{aligned} (\hat{\boldsymbol{\Omega}}^{(2)}, \hat{\boldsymbol{W}}^{(2)}) = \arg\min_{\boldsymbol{\Omega}\succ 0, \boldsymbol{W}} \Big\{&\text{tr}(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Omega} + \hat{\boldsymbol{\Sigma}}^*\boldsymbol{\Omega}^*) - \ln(|\boldsymbol{\Omega}|) \\ &- \ln(|\boldsymbol{\Omega}^*|) + \sum_{i\neq j} P'_{\lambda_n}(|\hat{W}_{ij}^{(1)}|)|W_{ij}| + \rho\|\boldsymbol{\Omega} - \boldsymbol{W}\|_F^2\Big\} . \end{aligned} \tag{67}$$

Convergence criterion was picked to be fractional improvement in cost (66) (or (67)) ≤ 0.001. Since the sparsity penalty is on $\boldsymbol{W}$, it was used to compute the error norm and to determine the edges: $\{i,j\} \in \mathcal{E}$ if $|W_{ij}| > 0$, else $\{i,j\} \notin \mathcal{E}$.

Simulation results based on 100 runs are shown in Fig. 1 for $p = 400$, with varying n =50, 100, 200, 400, 800, 1600, 3200 and 6400. We compare the SCAD solution with the solution to ℓ_1-penalized cost (5). The performance metrics used are:

- The F_1-score defined as

  $$F_1 = 2 \times \text{precision} \times \text{recall}/(\text{precision} + \text{recall}) ,$$

  precision $= |\hat{\mathcal{E}} \cap \mathcal{E}_0|/|\hat{\mathcal{E}}|$, recall $= |\hat{\mathcal{E}} \cap \mathcal{E}_0|/|\mathcal{E}_0|$, $\mathcal{E}_0$ and $\hat{\mathcal{E}}$ denote the true and estimated edge sets, respectively. If we denote TP= true positive (fraction of connected edges picked as connected), FP= false positive (fraction of missing edges picked as connected) and FN= false negative, then alternatively precision=TP/(TP+FP) and recall=TP/(TP+FN).
- Frobenius error norm = $\|\hat{\boldsymbol{W}} - \boldsymbol{\Omega}_0\|_F$ where $\hat{\boldsymbol{W}} = \hat{\boldsymbol{W}}^{(2)}$ for SCAD penalty and $\hat{\boldsymbol{W}} = \hat{\boldsymbol{W}}^{(1)}$ for ℓ_1 penalty. The error norm is labeled as "SCAD error" or "lasso error" in Fig. 1.

For (5), we set $\lambda_n = C_\ell\sqrt{\log(p)/n}$ (as suggested by [11, Theorem 1]), $\rho = 2$, and we set $\lambda_n = C_s\sqrt{(p+s_0)\log(p)/n}$ for (66)-(67), with $s_0 = 0.05\,p(p-1)$, number of off-diagonal nonzero elements in $\boldsymbol{\Omega}_0$, (as suggested by Theorem 1), $\rho = 2$ and $a = 3.7$ in (7). Several values of C_ℓ and C_s were tried. The results for three choices for each of these two approaches are displayed in Fig. 1.

In particular, the tuning parameter λ_n for ℓ_1-penalized cost (5) was picked as $\lambda_n = \lambda_{\ell 0}\sqrt{25\ln(p)/(n\ln(100))}$; i.e., when $n = 25$ and $p = 100$, $\lambda_n = \lambda_{\ell 0}$, and for other values of n, p, it is scaled according to [11, Theorem 1]. The results for three values $\lambda_{\ell 0} = 0.3$, 0.45, 0.6, (labeled "lasso, ... , $\lambda = 0.3$" or 0.45 or 0.6) are shown in Fig. 1. The tuning parameter λ_n for the proposed SCAD cost (66)-(67), was picked as $\lambda_n = \lambda_{s0}\sqrt{25(p+s_0)\ln(p)/(9900n\ \ln(100))}$; i.e., when $n = 25$ and $p = 100$, $\lambda_n = \lambda_{s0}$, and for other values of n, p, it is scaled according to (15) in Theorem 1. The results for three values $\lambda_{s0} = 0.08$, 0.10, 0.12, (labeled "SCAD, ... , $\lambda = 0.08$" or 0.10 or 0.12) are shown in Fig. 1.

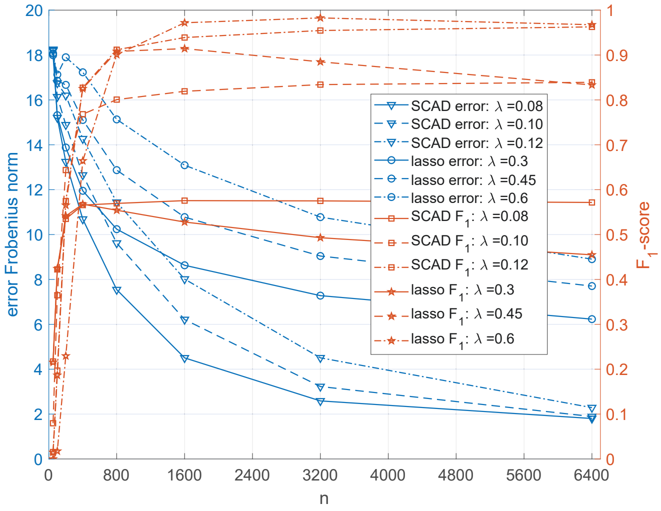

Fig. 1: *Error norm $\|\hat{\boldsymbol{W}} - \boldsymbol{\Omega}_0\|_F$ and corresponding F_1 values; p =400. The λ values shown refer to λ_n when $(p, n) = (100, 25)$; for other values, it is scaled according to (15) for SCAD, and [11, Theorem 1] for lasso; see the text.*

It is seen that for comparable F_1 values (a widely used measure of classification performance), the SCAD-penalized approach yields significantly smaller errors in estimating the inverse covariance matrix compared to the ℓ_1-penalized approach. For a given n, as penalty parameter is increased, both F_1-score and error norm increase, the former is desirable while the latter is not. It is desirable to maximize TP subject to FP $\leq$ some threshold value (say, 0.05). It is seen that for comparable values of F_1-score, the error norm in estimating $\boldsymbol{\Omega}_0$ is significantly smaller for SCAD-penalized approach compared to the lasso approach. We also see that as number of samples n increases, F_1-score values (generally) increase while the error norm values decrease, as expected.

7. CONCLUSIONS

We considered the problem of inferring the conditional independence graph of complex-valued, proper, multivariate Gaussian vectors in high dimensions, which is tantamount to estimating the inverse covariance with more unknowns than the sample size. We analyzed a SCAD-penalized log-likelihood based objective function to establish consistency and sparsistency of a local estimator of the inverse covariance in a neighborhood of the true value. An AM algorithm based on variable splitting and penalty method was used to optimize the objective function. A numerical examples was presented to illustrate the advantage of SCAD-penalty over the usual ℓ_1-penalty. However, SCAD penalty is nonconvex and therefore, can yield only a local optimum, unlike lasso.

8. REFERENCES

[1] H.H. Andersen, M. Hojbjerre, D. Sorensen, and P.S. Eriksen, *Linear and Graphical Models for the Multivariate Complex Normal Distribution*, Lecture Notes in Statistics, vol. 101. New York: Springer-Verlag, 1995.

[2] A.J. Rothman, P.J. Bickel, E. Levina and J. Zhu, "Sparse permutation invariant covariance estimation," *Electronic J. Statistics*, vol. 2, pp. 494-515, 2008.

[3] P. Ravikumar, M.J. Wainwright, G. Raskutti and B. Yu, "High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence," *Electronic J. Statistics*, vol. 5, pp. 935-980, 2011.

[4] C. Lam and J. Fan, "Sparsistency and rates of convergence in large covariance matrix estimation," *Ann. Statist.*, vol. 37, no. 6B, pp. 4254-4278, 2009.

[5] J. Friedman, T. Hastie and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432-441, July 2008.

[6] O. Banerjee, L.E. Ghaoui and A. d'Aspremont, "Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data," *J. Machine Learning Research*, vol. 9, pp. 485-516, 2008.

[7] P. Danaher, P. Wang and D.M. Witten, "The joint graphical lasso for inverse covariance estimation across multiple classes," *J. Royal Statistical Society, Series B (Methodological)*, vol. 76, pp. 373-397, 2014.

[8] A. Jung, G. Hannak and N. Goertz, "Graphical LASSO based model selection for time series," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1781-1785, Oct. 2015.

[9] J.K. Tugnait, "Graphical modeling of high-dimensional time series," in *Proc. 52nd Asilomar Conference on Signals, Systems & Computers*, Pacific Grove, CA, Oct. 28-31, 2018, pp. 840-844.

[10] J.K. Tugnait, "Graphical lasso for high-dimensional complex Gaussian graphical model selection," in *Proc. 2019 IEEE ICASSP*, Brighton, UK, May 2019, pp. 2952-2956.

[11] J.K. Tugnait, "On sparse complex Gaussian graphical model selection," in *Proc. 2019 IEEE Intern. Workshop on Machine Learning for Signal Processing (MLSP 2019)*, Pittsburgh, PA, Oct. 13-16, 2019.

[12] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *J. American Statistical Assoc.*, vol. 96, pp. 1348-1360, Dec. 2001.

[13] P.J. Schreier and L.L. Scharf, *Statistical Signal Processing of Complex-Valued Data*, Cambridge, UK: Cambridge Univ. Press, 2010.

[14] E. Ollila, "Direction of arrival estimation using robust complex Lasso," in *Proc. 2016 10th European Conf. Antennas Propag. (EuCAP)*, Davos, April 2016, pp. 1-5.

[15] A. Beck, "On the convergence of alternating minimization for convex programming with applications to iteratively reweighted least squares and decomposition schemes," *SIAM J. Optimization*, vol. 25, No. 1, pp. 185-209, 2015.

[16] P. Zheng, T. Askham, S.L. Brunton, J.N. Kutz and A.V. Aravkin, "A unified framework for sparse relaxed regularized regression: SR3," *IEEE Access*, vol. 7, pp. 1404-1423, 2019.