Calibrated Precision Matrix Estimation for High-Dimensional Elliptical Distributions

Tuo Zhao and Han Liu

***Abstract*— We propose a semiparametric method for estimating a precision matrix of high-dimensional elliptical distributions. Unlike most existing methods, our method naturally handles heavy tailness and conducts parameter estimation under a calibration framework, thus achieves improved theoretical rates of convergence and finite sample performance on heavy-tail applications. We further demonstrate the performance of the proposed method using thorough numerical experiments.**

***Index Terms*— Precision matrix, calibrated estimation, elliptical distribution, heavy-tailness, semiparametric model.**

I. Introduction

We consider the problem of precision matrix estimation. Let $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ be a d-dimensional random vector with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$, where $\boldsymbol{\Sigma}_{kj} = \mathbb{E}X_kX_j - \mathbb{E}X_k\mathbb{E}X_j$. We want to estimate the precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ based on n independent observations. In this paper we focus on high dimensional settings where $d/n \to \infty$. To handle the curse of dimensionality, we assume that $\boldsymbol{\Omega}$ is sparse (i.e., many off-diagonal entries of $\boldsymbol{\Omega}$ are zero).

A popular statistical model for precision matrix estimation is multivariate Gaussian, i.e., $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Under Gaussian models, sparse precision matrix encodes the conditional independence relationship of the random variables [8], [21], which has motivates numerous applications in different research areas [3], [15], [36]. In the past decade, many precision matrix estimation methods have been proposed for Gaussian distributions. For more details, let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ be n independent observations of $\boldsymbol{X}$, we define the sample covariance matrix as

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T, \tag{1}$$

Manuscript received July 29, 2013; revised May 9, 2014; accepted September 11, 2014. Date of publication September 30, 2014; date of current version November 18, 2014. This work was supported in part by the National Science Foundation under Grant IIS1408910 and Grant IIS1332109 and in part by the National Institutes of Health under Grant R01MH102339, Grant R01GM083084, and Grant R01HG06841. This paper was presented at the 27th Annual Conference on Neural Information Processing Systems in 2013.

T. Zhao is with the Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544 USA, and also with the Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: tour@cs.jhu.edu).

H. Liu is with the Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: hanliu@princeton.edu).

Communicated by N. Cesa-Bianchi, Associate Editor for Pattern Recognition, Statistical Learning and Inference.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2014.2360980

where $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i$. [1], [11], [38] propose the penalized Gaussian log-likelihood method named graphical lasso (GLASSO), which solves

$$\widehat{\boldsymbol{\Omega}} = \underset{\boldsymbol{\Omega}}{\operatorname{argmin}} -\log|\boldsymbol{\Omega}| + \operatorname{tr}(\mathbf{S}\boldsymbol{\Omega}) + \lambda\sum_{k,j}|\boldsymbol{\Omega}_{kj}|, \tag{2}$$

where $\lambda > 0$ is a regularization parameter for controlling the bias-variance tradeoff. In another line of research, [5], [37] propose pseudo-likelihood methods to estimate the precision matrix. Their methods adopt a column-by column estimation scheme and are more amenable to theoretical analysis. More specifically, given a matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, let $\mathbf{A}_{*j} = (\mathbf{A}_{1j}, \ldots, \mathbf{A}_{dj})^T$ denote the j^{th} column of $\mathbf{A}$, we define $||\mathbf{A}_{*j}||_1 = \sum_k |\mathbf{A}_{kj}|$ and $||\mathbf{A}_{*j}||_\infty = \max_k |\mathbf{A}_{kj}|$. [5] propose the CLIME estimator, which solves

$$\begin{aligned}\widehat{\boldsymbol{\Omega}}_{*j} = &\underset{\boldsymbol{\Omega}_{*j}}{\operatorname{argmin}}\ ||\boldsymbol{\Omega}_{*j}||_1\\ &\text{s.t.}\ \ ||\mathbf{S}\boldsymbol{\Omega}_{*j} - \mathbf{I}_{*j}||_\infty \le \lambda, \quad \forall j = 1, \ldots, d,\end{aligned} \tag{3}$$

to estimate the j^{th} column of the precision matrix. Moreover, let $||\mathbf{A}||_1 = \max_j ||\mathbf{A}_{*j}||_1$ be the matrix ℓ_1 norm of $\mathbf{A}$, and $||\mathbf{A}||_2$ be the largest singular value of $\mathbf{A}$, (i.e., the spectral norm of $\mathbf{A}$), [5] show that if we choose

$$\lambda \asymp ||\boldsymbol{\Omega}||_1 \cdot \sqrt{\frac{\log d}{n}}, \tag{4}$$

the CLIME estimator in (3) attains the rates of convergence

$$||\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}||_p = O_P\left(||\boldsymbol{\Omega}||_1^2 \cdot s\sqrt{\frac{\log d}{n}}\right), \tag{5}$$

where $s = \max_j \sum_k I(\boldsymbol{\Omega}_{kj} \neq 0)$, and $p = 1, 2$. Scalable software packages for GLASSO and CLIME have been developed which scale to thousands of dimensions [16], [22], [40].

Though significant progress has been for estimating Gaussian graphical models, most existing methods have two drawbacks: (i) They generally require the underlying distribution to be light-tailed [5], [7]. When this assumption is violated, these sample covariance matrix-based methods may have poor performance. (ii) They generally use the same tuning parameter to regularize the estimation, which is not adaptive to the individual sparseness of each column (More details will be provided in §III.B) and may lead to inferior finite sample performance. In another word, the regularization for estimating different columns of the precision matrix is not calibrated.

To overcome the above drawbacks, we propose a new sparse precision matrix estimation method, named EPIC (Estimating Precision matrIx with Calibration), which simultaneously

handles data heavy-tailness and conducts calibrated estimation. To relax the tail conditions, we adopt a combination of the rank-based transformed Kendall's tau estimator and Catoni's M-estimator [7], [18]. Such a semiparametric combination has shown better statistical properties than those of the sample covariance matrix for the heavy-tailed elliptical distributions [6], [7], [10], [17]. We will explain more details in § II and § IV. To calibrate the parameter estimation, we exploit a new framework proposed by [12]. Under this framework, the optimal tuning parameter does not depend on any unknown quantity of the data distribution, thus the EPIC estimator is tuning insensitive [25]. Computationally, the EPIC estimator is formulated as a convex program, which can be efficiently solved by the parametric simplex method [34]. Theoretically, we show that the EPIC estimator attains improved rates of convergence than the one in (5) under mild conditions. Numerical experiments on both simulated and real datasets show that the EPIC method outperforms existing precision matrix estimation methods.

The rest of this paper is organized as follows: In §II, we briefly review the elliptical family; In §III, we describe the proposed method and derive the computational algorithm; In §IV, we analyze the statistical properties of the EPIC estimator; In §V and §VI, we conduct numerical experiments on both simulated and real datasets to illustrate the effectiveness of the proposed method; In §VII, we discuss other related precision matrix estimation methods and compare them with our method [23]–[25].

II. Background

We start with some notations. Let $\boldsymbol{v} = (v_1, \ldots, v_d)^T \in \mathbb{R}^d$ be a vector, we define vector norms: $||\boldsymbol{v}||_1 = \sum_{j=1}^d |v_j|$, $||\boldsymbol{v}||_2^2 = \sum_{j=1}^d v_j^2$, $||\boldsymbol{v}||_\infty = \max_{1\le j\le d} |v_j|$. Let $\mathcal{S}$ be a subspace of $\mathbb{R}^d$, we use $\boldsymbol{v}_\mathcal{S}$ to denote the projection of $\boldsymbol{v}$ onto $\mathcal{S}$: $\boldsymbol{v}_\mathcal{S} = \operatorname{argmin}_{\boldsymbol{u}\in\mathcal{S}} ||\boldsymbol{u} - \boldsymbol{v}||_2^2$. We also define the orthogonal complement of $\mathcal{S}$ as $\mathcal{S}^\perp = \{\boldsymbol{u} \in \mathbb{R}^d | \boldsymbol{u}^T\boldsymbol{v} = 0, \text{ for any } \boldsymbol{v} \in \mathcal{S}\}$. Given a matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, let $\mathbf{A}_{*j} = (\mathbf{A}_{1j}, \ldots, \mathbf{A}_{dj})^T$ and $\mathbf{A}_{k*} = (\mathbf{A}_{k1}, \ldots, \mathbf{A}_{kd})^T$ denote the j^{th} column and k^{th} row of $\mathbf{A}$ in vector forms, we define matrix norms: $||\mathbf{A}||_1 = \max_j ||\mathbf{A}_{*j}||_1$, $||\mathbf{A}||_2 = \psi_{\max}(\mathbf{A})$, $||\mathbf{A}||_\infty = \max_k ||\mathbf{A}_{k*}||_1$, $||\mathbf{A}||_{\text{F}}^2 = \sum_j ||\mathbf{A}_{*j}||_2^2$, $||\mathbf{A}||_{\max} = \max_j ||\mathbf{A}_{*j}||_\infty$, where $\psi_{\max}(\mathbf{A})$ is the largest singular value of $\mathbf{A}$. We use $\Lambda_{\max}(\mathbf{A})$ and $\Lambda_{\min}(\mathbf{A})$ to denote the largest and smallest eigenvalues of $\mathbf{A}$. Moreover, we define the projection of $\mathbf{A}_{*j}$ onto $\mathcal{S}$ as $\mathbf{A}_{\mathcal{S}j} = \operatorname{argmin}_{\boldsymbol{u}\in\mathcal{S}} ||\boldsymbol{u} - \mathbf{A}_{*j}||_2^2$.

We then briefly review the elliptical family, which has the following definition.

Definition 2.1 ([10]): Given $\boldsymbol{\mu} \in \mathbb{R}^d$ and a symmetric positive semidefinite matrix $\boldsymbol{\Sigma}$ with $\operatorname{rank}(\boldsymbol{\Sigma}) = r \le d$, we say that a d-dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ follows an elliptical distribution with parameter $\boldsymbol{\mu}$, ξ, and $\boldsymbol{\Sigma}$ denoted by

$$\boldsymbol{X} \sim EC(\boldsymbol{\mu}, \xi, \boldsymbol{\Sigma}), \tag{6}$$

if $\boldsymbol{X}$ has a stochastic representation

$$\boldsymbol{X} \overset{d}{=} \boldsymbol{\mu} + \xi\mathbf{A}\boldsymbol{U} \tag{7}$$

where $\xi \ge 0$ is a continuous random variable independent of $\boldsymbol{U}$. Here $\boldsymbol{U} \in \mathbb{S}^{r-1}$ is uniformly distributed on the unit sphere in $\mathbb{R}^r$, and $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^T$.

Note that $\mathbf{A}$ and ξ in (7) can be properly rescaled without changing the distribution. Thus existing literature usually imposes an additional constraint $||\boldsymbol{\Sigma}||_{\max} = 1$ to make the distribution identifiable [10]. However, such a constraint does not necessarily make $\boldsymbol{\Sigma}$ the covariance matrix of $\boldsymbol{X}$. Since we are interested in estimating the precision matrix in this paper, we require $\mathbb{E}(\xi^2) < \infty$ and $\operatorname{rank}(\boldsymbol{\Sigma}) = d$ such that the precision matrix of the elliptical distribution exists. Under this assumption, we use an alternative constraint $\mathbb{E}(\xi^2) = d$, which not only makes the distribution identifiable but also has $\boldsymbol{\Sigma}$ defined as the conventional covariance matrix (e.g., as in the Gaussian distribution).

Remark 1: $\boldsymbol{\Sigma}$ can be factorized as $\boldsymbol{\Sigma} = \boldsymbol{\Theta}\mathbf{Z}\boldsymbol{\Theta}$, where $\mathbf{Z}$ is the Pearson correlation matrix, and $\boldsymbol{\Theta} = \operatorname{diag}(\theta_1, \ldots, \theta_d)$ with θ_j as the standard deviation of X_j. Since $\boldsymbol{\Theta}$ is a diagonal matrix, we can rewrite the precision matrix $\boldsymbol{\Omega}$ as $\boldsymbol{\Omega} = \boldsymbol{\Theta}^{-1}\boldsymbol{\Gamma}\boldsymbol{\Theta}^{-1}$, where $\boldsymbol{\Gamma} = \mathbf{Z}^{-1}$ is the inverse correlation matrix.

Remark 2: As a generalization of the Gaussian family, the elliptical family has been widely applied to many research areas such as dimensionality reduction [19], portfolio theory [14], and data visualization [33]. Many of these applications rely on an effective estimator of the precision matrix for elliptical distributions.

III. Method

Motivated by the above discussion, the EPIC method has three steps: We first use the transformed Kendall's tau estimator and Catoni's M-estimator to obtain $\widehat{\mathbf{Z}}$ and $\widehat{\boldsymbol{\Theta}}$ respectively; We then plug $\widehat{\mathbf{Z}}$ into a calibrated inverse correlation matrix estimation procedure to obtain $\widehat{\boldsymbol{\Gamma}}$; At last we assemble $\widehat{\boldsymbol{\Gamma}}$ and $\widehat{\boldsymbol{\Theta}}$ to obtain $\widehat{\boldsymbol{\Omega}}$. We explain more details about these three steps in the following subsections.

A. Correlation Matrix and Standard Deviation Estimation

To estimate $\mathbf{Z}$, we adopt the transformed Kendall's tau estimator proposed in [10] and [23]. More specifically, we define a population version of the Kendall's tau statistic between X_j and X_k as follows,

$$\begin{aligned}\tau_{kj} = &\mathbb{P}\left((X_j - \widetilde{X}_j)(X_k - \widetilde{X}_k) > 0\right)\\ &-\mathbb{P}\left((X_j - \widetilde{X}_j)(X_k - \widetilde{X}_k) < 0\right),\end{aligned}$$

where $\widetilde{X}_j$ and $\widetilde{X}_k$ are independent copies of X_j and X_k respectively. For elliptical distributions, [10], [23] show that $\mathbf{Z}_{kj}$'s and τ_{kj}'s have the following relationship

$$\mathbf{Z} = [\mathbf{Z}_{kj}] = \left[\sin\left(\frac{\pi}{2}\tau_{kj}\right)\right]. \tag{8}$$

Therefore given $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be n independent observations of $\boldsymbol{X}$, where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})^T$, we first calculate a sample version of the Kendall's tau statistic between X_j and X_k by

$$\widehat{\tau}_{kj} = \frac{2\sum_{i<i'} \operatorname{sign}((x_{ik} - x_{i'k})(x_{ij} - x_{i'j}))}{n(n-1)}$$

for all $k \neq j$, and 1 otherwise. We then obtain a correlation matrix estimator by the same entrywise transformation as (8),

$$\widehat{\mathbf{Z}} = [\widehat{\mathbf{Z}}_{kj}] = \left[\sin\left(\frac{\pi}{2}\widehat{\tau}_{kj}\right)\right]. \tag{9}$$

To estimate $\mathbf{\Theta}$, we exploit the Catoni's M-estimator proposed in [7]. For heavy-tailed distributions, [7] show that the Catoni's M-estimator has better theoretical and empirical performance than the sample moment-based estimator. In particular, let $\psi(t) = \text{sign}(t) \cdot \log(1 + |t| + t^2/2)$ be a univariate function where $\text{sign}(0) = 0$. Let $\widehat{\mu}_j$ and $\widehat{m}_j$ be the estimator of $\mathbb{E}X_j$ and $\mathbb{E}X_j^2$ respectively which solve the following two equations:

$$\sum_{i=1}^{n} \psi\left((x_{ij} - \mu_j)\sqrt{\frac{2}{nK_{\max}}}\right) = 0, \tag{10}$$

$$\sum_{i=1}^{n} \psi\left((x_{ij}^2 - m_j)\sqrt{\frac{2}{nK_{\max}}}\right) = 0. \tag{11}$$

Here $K_{\max}$ is a preset upper bound of $\max_j \text{Var}(X_j)$ and $\max_j \text{Var}(X_j^2)$. [7] shows that the solutions to (10) and (11) must exist and can be efficiently solved by the Newton-Raphson algorithm [31]. Once we obtain $\widehat{m}_j$ and $\widehat{\mu}_j$, we estimate the marginal standard deviation θ_j by

$$\widehat{\theta}_j = \sqrt{\max\{\widehat{m}_j - \widehat{\mu}_j^2, K_{\min}\}}, \tag{12}$$

where $K_{\min}$ is a preset lower bound of $\min_j \theta_j^2$.

Remark 3: We choose the combination of the transformed Kendall's tau estimator and Catoni's M-estimator instead of sample covariance matrix, because we are handling heavy-tailed elliptical distributions. For light-tailed distributions (e.g. Gaussian distribution), we can still use the sample correlation matrix and sample standard deviation to estimate the $\mathbf{Z}$ and $\mathbf{\Theta}$. The extension of our proposed methodology and theory is straightforward. See more details in §IV.

B. Calibrated Inverse Correlation Matrix Estimation

We then plug the transformed Kendall's tau estimator $\widehat{\mathbf{Z}}$ into the following convex program,

$$\begin{aligned}(\widehat{\mathbf{\Gamma}}_{*j}, \widehat{\tau}_j) = &\underset{\mathbf{\Gamma}_{*j}, \tau_j}{\text{argmin}} \ ||\mathbf{\Gamma}_{*j}||_1 + c\tau_1 \\ \text{s.t.} \ \ &||\widehat{\mathbf{Z}}\mathbf{\Gamma}_{*j} - \mathbf{I}_{*j}||_\infty \leq \lambda\tau_j, \ ||\mathbf{\Gamma}_{*j}||_1 \leq \tau_j,\end{aligned} \tag{13}$$

for all $j = 1, \ldots, d$, where c can be any constant between 0 and 1 (e.g., $c = 0.5$). Here τ_j serves as an auxiliary variable to calibrate the regularization [12], [32]. Both the objective function and constraints in (13) contain τ_j to prevent from choosing τ_j either too large or too small.

To gain more intuition of the formulation of (13), we first consider estimating the j^{th} column of the inverse correlation matrix using the CLIME method in a regularization form as follows,

$$\widehat{\mathbf{\Gamma}}_{*j} = \underset{\mathbf{\Gamma}_{*j}}{\text{argmin}} \, ||\mathbf{\Gamma}_{*j}||_1 + \nu||\widehat{\mathbf{Z}}\mathbf{\Gamma}_{*j} - \mathbf{I}_{*j}||_\infty, \tag{14}$$

where $\nu > 0$ is the regularization parameter. The next proposition presents an alternative formulation of (14).

Proposition III.1: The following optimization problem

$$\begin{aligned}(\widehat{\mathbf{\Gamma}}_{*j}, \widehat{\tau}_j) = &\underset{\mathbf{\Gamma}_{*j}, \tau_j}{\text{argmin}} \ ||\mathbf{\Gamma}_{*j}||_1 + c\tau_j \\ \text{s.t.} \ \ &||\widehat{\mathbf{Z}}\mathbf{\Gamma}_{*j} - \mathbf{I}_{*j}||_\infty \leq \frac{c}{\nu}\tau_j.\end{aligned} \tag{15}$$

has the same solution as (14).

The proof of Proposition III.1 is provided in Appendix A. If we set $\nu/c = \lambda$, then the only difference between (13) and (15) is that (13) contains a constraint $||\mathbf{\Gamma}_{*j}||_1 \leq \tau_j$. Due to the complementary slackness, this additional constraint encourages the regularization $\lambda\tau_j$ to be proportional to the ℓ_1 norm of the j^{th} column (weak sparseness). From the theoretical analysis in §IV, we see that the regularization is calibrated in this way.

In the rest of this subsection, we omit the index j in (13) for notational simplicity. We denote $\mathbf{\Gamma}_{*j}$, $\mathbf{I}_{*j}$, and τ_j by $\boldsymbol{\gamma}$, $\boldsymbol{e}$, and τ respectively. By reparametrizing $\boldsymbol{\gamma} = \boldsymbol{\gamma}^+ - \boldsymbol{\gamma}^-$, we can rewrite (13) as the following linear program,

$$\begin{aligned}(\widehat{\boldsymbol{\gamma}}^+, \widehat{\boldsymbol{\gamma}}^-, \widehat{\tau}) = &\underset{\boldsymbol{\gamma}^+, \boldsymbol{\gamma}^-, \tau}{\text{argmin}} \ \ \mathbf{1}^T\boldsymbol{\gamma}^+ + \mathbf{1}^T\boldsymbol{\gamma}^- + c\tau \\ \text{s.t.} \ \ &\begin{bmatrix} \widehat{\mathbf{Z}} & -\widehat{\mathbf{Z}} & -\boldsymbol{\lambda} \\ -\widehat{\mathbf{Z}} & \widehat{\mathbf{Z}} & -\boldsymbol{\lambda} \\ \mathbf{1}^T & \mathbf{1}^T & -1 \end{bmatrix} \begin{bmatrix} \boldsymbol{\gamma}^+ \\ \boldsymbol{\gamma}^- \\ \tau \end{bmatrix} \leq \begin{bmatrix} \boldsymbol{e} \\ -\boldsymbol{e} \\ 0 \end{bmatrix}, \\ &\boldsymbol{\gamma}^+ \geq \mathbf{0}, \quad \boldsymbol{\gamma}^- \geq \mathbf{0}, \quad \tau \geq 0,\end{aligned} \tag{16}$$

where $\boldsymbol{\lambda} = \lambda\mathbf{1}$. Though (16) can be solved by general linear program solvers (e.g. the simplex method as suggested in [5]), these general solvers cannot scale to large problems. In Appendix B, we provide a more efficient parametric simplex method [34], which naturally exploits the underlying sparsity structure, and attains better empirical performance than the simplex method.

C. Symmetric Precision Matrix Estimation

Once we get the inverse correlation matrix estimate $\widehat{\mathbf{\Gamma}}$, we estimate the precision matrix by

$$\widetilde{\mathbf{\Omega}} = \widehat{\mathbf{\Theta}}^{-1}\widehat{\mathbf{\Gamma}}\widehat{\mathbf{\Theta}}^{-1}.$$

Remark 4: A possible alternative is that we first assemble a covariance matrix estimator

$$\widehat{\mathbf{S}} = \widehat{\mathbf{\Theta}}\widehat{\mathbf{Z}}\widehat{\mathbf{\Theta}}, \tag{17}$$

then directly estimate $\mathbf{\Omega}$ by solving

$$\begin{aligned}(\widehat{\mathbf{\Omega}}_{*j}, \widehat{\tau}_j) = &\underset{\mathbf{\Omega}_{*j}, \tau_j}{\text{argmin}} \ ||\mathbf{\Omega}_{*j}||_1 + c\tau_j \\ \text{s.t.} \ \ &||\widehat{\mathbf{S}}\mathbf{\Omega}_{*j} - \mathbf{I}_{*j}||_\infty \leq \lambda\tau_j, \ ||\mathbf{\Omega}_{*j}||_1 \leq \tau_j\end{aligned}$$

for all $j = 1, \ldots, d$. However, such a direct estimation procedure makes the regularization parameter selection sensitive to marginal variability. See [20], [26], [29] for more discussions of the ensemble rule.

The EPIC method does not guarantee the symmetry of $\widetilde{\mathbf{\Omega}}$. To get a symmetric estimate, we take an additional projection procedure to obtain a symmetric estimator

$$\widehat{\mathbf{\Omega}} = \underset{\mathbf{\Omega}}{\text{argmin}} \, ||\mathbf{\Omega} - \widetilde{\mathbf{\Omega}}||_* \quad \text{s.t. } \mathbf{\Omega} = \mathbf{\Omega}^T, \tag{18}$$

where $||\cdot||_*$ can be the matrix ℓ_1, Frobenius, or max norm. More details about how to choose a suitable norm will be explained in the next section.

Remark 5: For the Frobenius and max norms, (18) has a closed form solution as follows,

$$\widehat{\mathbf{\Omega}} = \frac{1}{2}\left(\widetilde{\mathbf{\Omega}} + \widetilde{\mathbf{\Omega}}^T\right).$$

For the matrix ℓ_1 norm, see our proposed smoothed proximal gradient algorithm in Appendix C. More details about how to choose a suitable norm will be explained in the next section.

IV. Statistical Properties

To analyze the statistical properties of the EPIC estimator, we define the following class of sparse symmetric matrices,

$$\mathcal{U}(s, M, \kappa_u) = \Big\{\mathbf{\Gamma} \in \mathbb{R}^{d\times d} \;\Big|\; \mathbf{\Gamma} \succ 0, \Lambda_{\max}(\mathbf{\Gamma}) \le \kappa_u, \max_j \sum_k I\left(\mathbf{\Gamma}_{kj} \neq 0\right) \le s, ||\mathbf{\Gamma}||_1 \le M\Big\},$$

where κ_u is a constant, and (s, d, M) may scale with the sample size n. We assume that the following conditions hold:

(A.1) $\mathbf{\Gamma} \in \mathcal{U}(s, M, \kappa_u)$,
(A.2) $\theta_{\min} \le \min_j \theta_j \le \max_j \theta_j \le \theta_{\max}$,
(A.3) $\max_j |\mu_j| \le \mu_{\max}$, $\max_j \mathbb{E}X_j^4 \le K$,
(A.4) $s^2 \log d/n \to 0$,

where $\theta_{\max}$, $\theta_{\min}$, $\mu_{\max}$, and K are constants.

Remark 6: Condition (A.3) only requires the fourth moment of the distribution to be finite. In contrast, sample covariance-based estimation methods can not achieve such theoretical results. See more details in [5] and [7].

Remark 7: The bounded mean in Condition (A.3) is actually a mild assumption. Existing high dimensional theories (Cai et al. 2011; Yuan, 2010; Rothman et al. 2008) on sparse precision matrix estimation all require the distribution to be light-tailed. For example, there exists some constant K such that $\max_j \mathbb{E}|X_j|^r \le K < \infty$ for some $r \gg 4$. By Jessen's inequality, we have $(\mathbb{E}|X_j|)^r \le \mathbb{E}|X_j|^r \le K < \infty$, which implies that $\max_j \mathbb{E}|X_j| \le K^{1/r} < \infty$. In another word, they also require $\max_j |\mu_j|$ to be bounded.

Before we proceed with main results, we first present the following important lemma.

Lemma 1: We assume that $\boldsymbol{X} \sim EC(\boldsymbol{\mu}, \xi, \mathbf{\Sigma})$ and (A.2)-(A.4) hold. Let $\widehat{\mathbf{Z}}$ and $\widehat{\theta}_j$ be defined in (9) and (12). There exist universal constants κ_1 and κ_2 such that for large enough n,

$$\mathbb{P}\left(\max_{k,j} |\widehat{\mathbf{Z}}_{kj} - \mathbf{Z}_{kj}| \le \kappa_1\sqrt{\frac{\log d}{n}}\right) \ge 1 - \frac{1}{d}, \tag{19}$$

$$\mathbb{P}\left(\max_{j} |\widehat{\theta}_j^{-1} - \theta_j^{-1}| \le \kappa_2\sqrt{\frac{\log d}{n}}\right) \ge 1 - \frac{2}{d}. \tag{20}$$

The proof of Lemma 1 is provided in Appendix D.

Remark 8: Lemma 1 shows that the transformed Kendall's tau estimator and Catoni's M-estimator possess good concentration properties for heavy-tailed elliptical distributions. That enables us to obtain a consistent precision matrix estimator in high dimensions.

A. Parameter Estimation Consistency

Theorem IV.1 provides the rates of convergence for precision matrix estimation under the matrix ℓ_1, spectral, and Frobenius norms.

Theorem IV.1: Suppose that $\boldsymbol{X} \sim EC(\boldsymbol{\mu}, \xi, \mathbf{\Sigma})$ and (A.1)-(A.4) hold, if we take $\lambda = \kappa_1\sqrt{\log d/n}$ and choose the matrix ℓ_1 norm as $||\cdot||_*$ in (18), then for large enough n and $p = 1, 2$, there exists a universal constant C_1 such that

$$\mathbb{P}\left(||\widehat{\mathbf{\Omega}} - \mathbf{\Omega}||_p \le C_1 M s\sqrt{\frac{\log d}{n}}\right) \ge 1 - \frac{3}{d}. \tag{21}$$

Moreover, if we choose the Forbenius norm as $||\cdot||_*$ in (18), then for large enough n, there exists a universal constant C_2 such that

$$\mathbb{P}\left(\frac{1}{d}||\widehat{\mathbf{\Omega}} - \mathbf{\Omega}||_{\mathrm{F}}^2 \le C_2 M^2\frac{s\log d}{n}\right) \ge 1 - \frac{3}{d}. \tag{22}$$

The proof of Theorem IV.1 is provided in Appendix E. Note that the rates of convergence obtained in the above theorem are faster than those in [5].

B. Model Selection Consistency

Theorem IV.2 provides the rate of convergence under the elementwise max norm.

Theorem IV.2: Suppose that $\boldsymbol{X} \sim EC(\boldsymbol{\mu}, \xi, \mathbf{\Sigma})$ and (A.1)-(A.4) hold. If we take $\lambda = \kappa_1\sqrt{\log d/n}$ and choose the max norm for (18), then for large enough n, there exists a universal constant C_3 such that

$$\mathbb{P}\left(||\widehat{\mathbf{\Omega}} - \mathbf{\Omega}||_{\max} \le C_3 M^2\sqrt{\frac{\log d}{n}}\right) \ge 1 - \frac{3}{d}. \tag{23}$$

Moreover, let $E = \{(k, j) \mid \mathbf{\Omega}_{kj} \neq 0\}$, and $\widehat{E} = \{(k, j) \mid \widehat{\mathbf{\Omega}}_{kj} \neq 0\}$, if there exists large enough constant C_4 such that

$$\min_{(k,j)\in E} |\mathbf{\Omega}_{kj}| \ge C_4 M^2\sqrt{\frac{\log d}{n}},$$

then we have $\mathbb{P}\left(E \subseteq \widehat{E}\right) \to 1$.

The proof of Theorem IV.2 is provided in Appendix G. The obtained rate of convergence in Theorem IV.2 is comparable to that of [5].

Remark 9: Our selected regularization parameter $\lambda = \kappa_1\sqrt{\log d/n}$ in Theorems IV.1 and IV.2 does not contain any unknown parameter of the underlying distribution (e.g. $||\mathbf{\Gamma}||_1$). Note that κ_1 comes from (19) in Lemma 1. Theoretically we can choose κ_1 as a reasonably large constant without any additional tuning (e.g. $\sqrt{2}\pi$. See more details in [23]). In practice, we found that a fine tuning of κ_1 delivers better finite sample performance.

V. Numerical Results

In this section, we compare the EPIC estimator with several competing estimators including:

- CLIME.RC: We obtain the sparse precision matrix estimator by plugging the covariance matrix estimator $\widehat{\mathbf{S}}$ defined in (17) into (3).

TABLE I

TIMING PERFORMANCE OF DIFFERENT ESTIMATORS ON THE BAND, ERDÖS-RÉNYI, AND SCALE-FREE MODELS (IN SECONDS). THE BASELINE PERFORMANCE IS OBTAINED BY SOLVING THE CLIME.SC METHOD USING THE SIMPLEX METHOD

Model	d	EPIC	GLASSO.RC	CLIME.RC	CLIME.SC	BASELINE
Band	101	0.1561(0.0248)	0.3633(0.0070)	0.1233(0.0057)	0.1701(0.0119)	49.467(1.7862)
	201	1.6622(0.1253)	0.4417(0.0122)	1.5897(0.1249)	1.6085(0.0518)	687.57(23.720)
	401	23.061(0.5777)	1.0864(0.1403)	24.441(1.5344)	25.445(3.8066)	4756.4(170.25)
Erdös-Rényi	101	0.1414(0.0079)	0.3703(0.0072)	0.1309(0.0331)	0.2073(0.0925)	59.775(2.0521)
	201	1.6214(0.5175)	0.4448(0.0164)	1.5992(0.1840)	1.6155(0.2957)	803.51(29.835)
	401	21.722(0.5470)	1.1517(0.0959)	22.795(0.6999)	24.230(3.1871)	4531.7(151.46)
Scale-free	101	0.2245(0.0514)	0.4398(0.0843)	0.1509(0.0054)	0.1871(0.0149)	55.112(1.7109)
	201	1.8682(0.1078)	0.4632(0.0067)	1.5472(0.1350)	1.7235(0.1778)	865.98(31.399)
	401	21.926(0.7112)	1.0093(0.1140)	23.135(1.4318)	25.596(3.3401)	4991.2(202.44)

- CLIME.SC: We obtain the sparse precision matrix estimator by plugging the sample covariance matrix estimator $\mathbf{S}$ defined in (1) into (3).
- GLASSO.RC: We obtain the sparse precision matrix estimator by plugging the covariance matrix estimator $\widehat{\mathbf{S}}$ defined in (17) into (2).

Moreover, (3) is also solved by the parametric simplex method as our proposed EPIC method, and (2) is solved by the block coordinate descent algorithm. All experiments are conducted on a PC with Core i5 3.3GHz CPU and 16GB memory. All programs are coded using `C` using double precision, and further called from `R`.

A. Data Generation

We consider three different settings for comparison: (1) $d = 101$; (2) $d = 201$; (3) $d = 401$. We adopt the following three graph generation schemes, as illustrated in Figure 1, to obtain precision matrices:

- **Band**. Each node is assigned an index j with $j = 1, \ldots, d$. Two nodes are connected by an edge if the difference between their indices is no larger than 2.
- **Erdös-Rényi**. We set an edge between each pair of nodes with probability $4/d$, independently of the other edges.
- **Scale-free**. The degree distribution of the graph follows a power law. The graph is generated by the preferential attachment mechanism.

The graph begins with an initial chain graph of 10 nodes. New nodes are added to the graph one at a time. Each new node is connected to an existing node with a probability that is proportional to the number of degrees that the existing nodes already have. Formally, the probability p_i that the new node is connected to the i^{th} existing node is $p_i = \frac{k_i}{\sum_j k_j}$ where k_i is the degree of node i.

Let $\mathbf{G}$ be the adjacency matrix of the generated graph, we calculate $\widetilde{\mathbf{G}} = [\widetilde{\mathbf{G}}_{jk}]$ as

$$\widetilde{\mathbf{G}}_{jk} = \widetilde{\mathbf{G}}_{kj} = \begin{cases} \mathbf{U}_{kj} & \text{if } \mathbf{G}_{jk} = \mathbf{G}_{kj} = 1 \\ 0 & \text{if } \mathbf{G}_{jk} = \mathbf{G}_{kj} = 0 \end{cases}$$

where all $\mathbf{U}_{kj}$'s are independently sampled from the uniform distribution Uniform$(-1, +1)$. Let $\mathcal{C}_2$ be the rescaling operator that converts a symmetric positive definite matrix to the corresponding correlation matrix, we further calculate

$$\mathbf{\Sigma} = \mathbf{\Theta}\mathcal{C}_2[(\widetilde{\mathbf{G}} + (0.1 - \Lambda_{\min}(\widetilde{\mathbf{G}})) \cdot \mathbf{I})^{-1}]\mathbf{\Theta},$$

where $\mathbf{\Theta}$ is the diagonal standard deviation matrix with $\mathbf{\Theta}_{jj} = 2^{\frac{2j-d-1}{2(d-1)}}$ for $j = 1, \ldots, d$.

We then generate $n = \lceil 14\sqrt{d} \rceil$ independent samples from the t-distribution with 6 degrees of freedom, mean $\mathbf{0}$, and covariance $\mathbf{\Sigma}$. For the EPIC estimator, we set $c = 0.5$ in (13). For the Catoni's M-estimator, we set $K_{\max} = 10$ and $K_{\min} = 0.1$.

B. Timing Performance

We first evaluate the computational performance of the parametric simplex method. For each model, we choose a regularization parameter, which yields approximate $0.05 \cdot d(d-1)$ nonzero off-diagonal entries. The EPIC and CLIME methods are solved by the parametric simplex method, which is described in Appendix B. The GLASSO is solved by the dual block coordinate descent algorithm, which is described in [11]. Table I summarizes the timing performance averaged over 100 replications. To obtain the baseline performance, we solve the CLIME.SC method using the simplex method[1] as suggested in [5]. We see that all four methods greatly outperform the baseline. The EPIC, CLIME.RC, and CLIME.SC methods attain similar timing performance for all settings, and the GLASSO.RC method is more efficient than the others for $d = 201$ and $d = 401$.

C. Parameter Estimation

To select the regularization parameter, we independently generate a validation set of n samples from the same distribution. We tune λ over a refined grid, then the selected optimal regularization parameter is $\widehat{\lambda} = \text{argmin}_\lambda \, ||\widehat{\mathbf{\Omega}}^\lambda \widehat{\mathbf{\Sigma}} - \mathbf{I}||_{\max}$, where $\widehat{\mathbf{\Omega}}^\lambda$ denotes the estimated precision matrix of the training set using the regularization parameter λ, and $\widehat{\mathbf{\Sigma}}$ denotes the estimated covariance matrix of the validation set using either (1) or (17).

[1]The implementation of the simplex method is based on the `R` packages `linprog` and `lpSolve`.

TABLE II
QUANTITIVE COMPARISON OF DIFFERENT ESTIMATORS ON THE BAND, ERDÖS-RÉNYI, AND SCALE-FREE MODELS. THE EPIC ESTIMATOR OUTPERFORMS THE COMPETITORS IN ALL SETTINGS

Spectral Norm: $\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}\|_2$					
Model	d	EPIC	GLASSO.RC	CLIME.RC	CLIME.SC
Band	101	3.3748(0.2081)	4.4360(9,1445)	3.3961(0.4403)	3.6885(0.5850)
	201	3.3283(0.1114)	4.8616(0.0644)	3.4559(0.0979)	4.4789(0.3399)
	401	3.5933(0.5192)	5.1667(0.0354)	4.0623(0.2397)	5.7164(0.9666)
Erdös-Rényi	101	2.1849(0.2281)	2.6681(0.1293)	2.6787(0.8414)	2.3391(0.2976)
	201	1.8322(0.0769)	2.3753(0.0949)	2.0106(0.3943)	2.0528(0.1548)
	401	1.3322(0.1294)	2.4265(0.0564)	2.0051(0.4144)	4.0667(1.1174)
Scale-free	101	2.1113(0.3081)	2.9979(0.1654)	2.0401(0.3703)	2.6541(0.5882)
	201	2.3519(0.1779)	3.2394(0.1078)	2.3785(0.4186)	2.5789(0.5139)
	401	3.2273(0.1201)	4.0105(0.5812)	3.3139(0.5812)	3.9287(1.1750)

TABLE III
QUANTITIVE COMPARISON OF DIFFERENT ESTIMATORS ON THE BAND, ERDÖS-RÉNYI, AND SCALE-FREE MODELS. THE EPIC ESTIMATOR OUTPERFORMS THE COMPETITORS IN ALL SETTINGS

Frobenius Norm: $\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}\|_{\mathrm{F}}$					
Model	d	EPIC	GLASSO.RC	CLIME.RC	CLIME.SC
Band	101	9.4307(0.3245)	11.069(0.2618)	9.7538(0.3949)	11.392(0.8319)
	201	12.720(0.2282)	16.135(0.1399)	13.533(0.1898)	14.850(0.6167)
	401	18.298(1.0537)	23.177(0.1957)	20.412(0.2366)	25.254(1.0002)
Erdös-Rényi	101	6.0660(0.1552)	6.8777(0.2115)	6.7097(0.3672)	7.3789(0.4390)
	201	6.7794(0.1632)	8.1531(0.1828)	7.6175(0.2616)	8.3555(0.2844)
	401	7.3497(0.1743)	10.795(0.1323)	8.3869(0.4755)	11.104(0.6069)
Scale-free	101	4.6695(0.2435)	5.6689(0.2344)	4.9658(0.1762)	6.2264(0.3841)
	201	5.6732(0.1782)	7.2768(0.0940)	6.2343(0.2401)	7.2842(0.3310)
	401	7.2979(0.1094)	9.0940(0.0935)	7.3765(0.2328)	9.5396(0.5636)

Tables II and III summarize the numerical results averaged over 100 replications. We see that the EPIC estimator outperforms the GLASSO.RC and CLIME.RC estimators in all settings.

D. Model Selection

To evaluate the model selection performance, we calculate the ROC curve of each obtained regularization path using the false positive rate (FPR) and true positive rate (FNR) defined as follows,

$$\text{F.P.R.} = \frac{\sum_{k,j} I\left(\widehat{\mathbf{\Omega}}_{kj}^{(\lambda)} \neq 0,\ \mathbf{\Omega}_{kj} = 0\right)}{\sum_{k,j} I\left(\mathbf{\Omega}_{kj} = 0\right)},$$

$$\text{T.P.R.} = \frac{\sum_{k,j} I\left(\widehat{\mathbf{\Omega}}_{kj}^{(\lambda)} \neq 0,\ \mathbf{\Omega}_{kj} \neq 0\right)}{\sum_{k,j} I\left(\widehat{\mathbf{\Omega}}_{kj}^{(\lambda)} \neq 0\right)}.$$

Figure 1 summarizes ROC curves of all methods averaged over 100 replications.[2] We see that the EPIC estimator outperforms the competing estimators throughout all settings. Similarly, our method outperforms the sample covariance matrix-based CLIME estimator.

[2]The ROC curves from different replications are first aligned by regularization parameters. The averaged ROC curve shows the false positive and true positive rate averaged over all replications w.r.t. each regularization parameter

VI. REAL DATA EXAMPLE

To illustrate the effectiveness of the proposed EPIC method, we adopt the sonar dataset from UCI Machine Learning Repository[3] [13]. The dataset contains 101 patterns obtained by bouncing sonar signals off a metal cylinder at various angles and under various conditions, and 97 patterns obtained from rocks under similar conditions. Each pattern is a set of 60 features. Each feature represents the logarithm of the energy integrated over a certain period of time within a particular frequency band. Our goal is to discriminate between sonar signals bounced off a metal cylinder and those bounced off a roughly cylindrical rock.

We randomly split the data into two sets. The training set contains 80 metal and 77 rock patterns. The testing set contains 21 metal and 20 rock patterns. Let $\boldsymbol{\mu}^{(k)}$ be the class conditional means of the data where $k = 1$ represents the metal category and $k = 0$ represents the rock category. [5] assume that two classes share the same covariance matrix, and then adopt the sample mean for estimating $\boldsymbol{\mu}_k$'s and the sample covariance matrix-based CLIME estimator for estimating $\mathbf{\Omega}$.

[3]Available at http://http://archive.ics.uci.edu/ml/datasets.html.

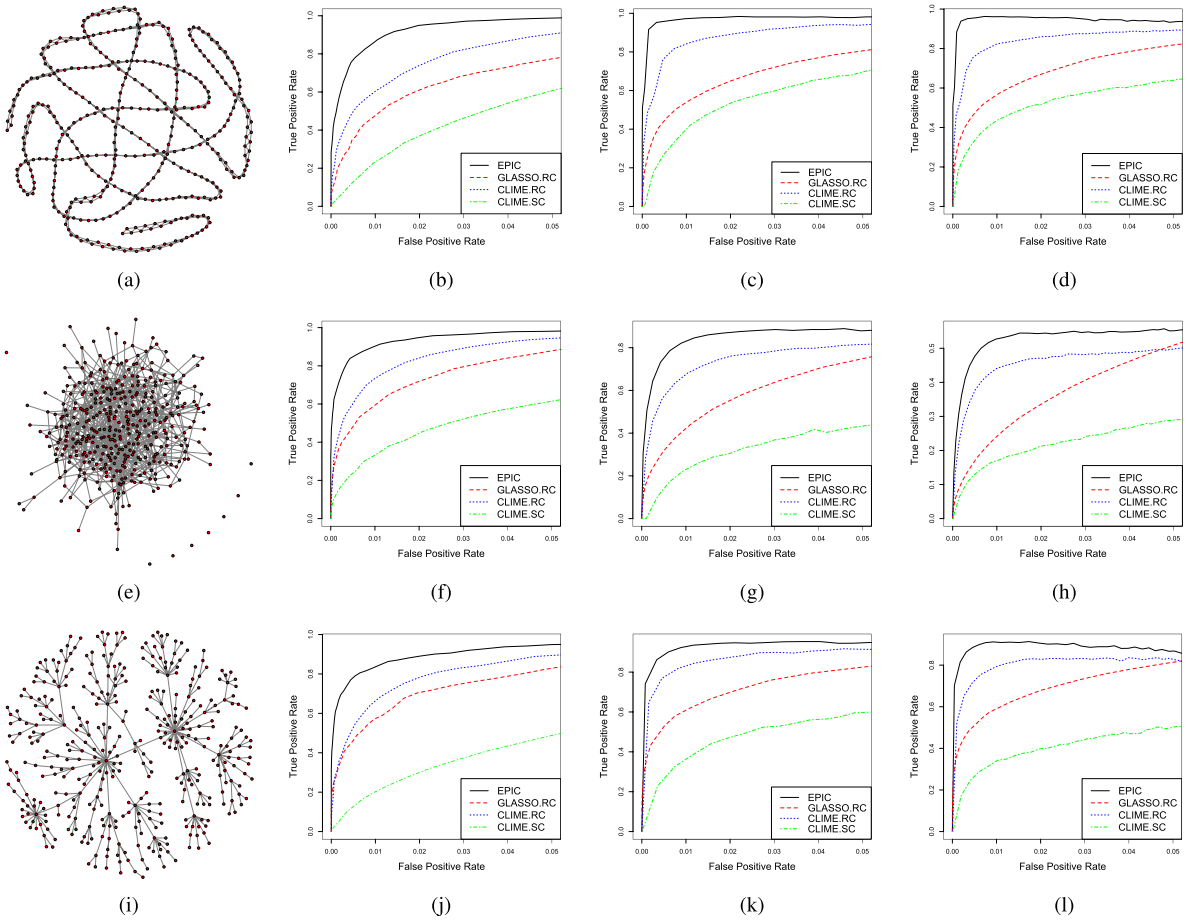

Fig. 1. Three different graph patterns and corresponding average ROC curves. EPIC outperforms the competitors throughout all settings. (a) Band ($d = 401$). (b) Band ($d = 101$). (c) Band ($d = 201$). (d) Band ($d = 401$). (e) Erdös-Rényi ($d = 401$). (f) Erdös-Rényi ($d = 101$). (g) Erdös-Rényi ($d = 201$). (h) Erdös-Rényi ($d = 401$). (i) Scale-free ($d = 401$). (j) Scale-free ($d = 101$). (k) Scale-free ($d = 201$). (l) Scale-free ($d = 401$).

In contrast, we adopt the Catoni's M-estimator for estimating $\boldsymbol{\mu}_k$'s and the EPIC estimator for estimating $\boldsymbol{\Omega}$. We classify a sample $\boldsymbol{x}$ to the metal category if

$$\left(\boldsymbol{x} - \frac{\widehat{\boldsymbol{\mu}}^{(1)} + \widehat{\boldsymbol{\mu}}^{(0)}}{2}\right)^T \widehat{\boldsymbol{\Omega}} \left(\widehat{\boldsymbol{\mu}}^{(1)} - \widehat{\boldsymbol{\mu}}^{(0)}\right) \geq 0,$$

and to the rock category otherwise. We use the testing set to evaluate the performance of the EPIC estimator. For tuning parameter selection, we use a 5-fold cross validation on the training set to pick the regularization parameter λ.

To evaluate the classification performance, we use the criteria of misclassification rate, specificity, sensitivity, and Mathews Correlation Coefficient (MCC). More specifically, let y_i's and $\widehat{y}_i$'s be true labels and predicted labels of the testing samples, we define

$$\text{Misclassification Rate} = \frac{\text{TP} + \text{TN}}{\text{TN} + \text{TP} + \text{FN} + \text{FP}},$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \quad \text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}},$$

where

$$\text{TP} = \sum_i I(\widehat{y}_i = y_i = 1), \ \text{FP} = \sum_i I(\widehat{y}_i = 1, \ y_i = 0),$$

$$\text{TN} = \sum_i I(\widehat{y}_i = y_i = 0), \ \text{FN} = \sum_i I(\widehat{y}_i = 0, \ y_i = 1).$$

Table IV summarizes the performance of both methods averaged over 100 replications (with standard errors in parentheses). We see that the EPIC estimator significantly outperforms the competitor on the sensitivity and misclassification rate, but slightly worse on the specificity. The overall classification performance measured by MCC shows that the EPIC estimator has about 8% improvement over the competitor.

VII. Discussion and Conclusion

In this paper, we propose a new sparse precision matrix estimation method for the elliptical family. Our method handles heavy-tailness, and conducts parameter estimation under a calibration framework. We show that the proposed method achieves improved rates of convergence and better finite sample performance than existing methods. The effectiveness of the proposed method is further illustrated by numerical experiments on both simulated and real datasets.

TABLE IV
QUANTITIVE COMPARISON OF THE EPIC AND SAMPLE COVARIANCE MATRIX-BASED CLIME ESTIMATORS IN THE SONAR DATA CLASSIFICATION

Method	Misclassification Rate	Specificity	Sensitivity	MCC
EPIC	0.1990(0.0285)	0.7288(0.0499)	0.8579(0.0301)	0.6023(0.0665)
CLIME.SC	0.2362(0.0317)	0.7460(0.0403)	0.7791(0.0429)	0.5288(0.0631)

[25] proposed another calibrated graph estimation method named TIGER for Gaussian family. However, unlike the EPIC estimator, the TIGER method can not handle the elliptical family due to two reasons: (1) The transformed Kendall's tau estimator cannot guarantee the positive semidefiniteness. If we directly plug it into the TIGER method, it makes the TIGER formulation nonconvex. Existing algorithms may not obtain a global solution in polynomial time. (2) The theoretical analysis in [25] is only applicable to the Gaussian family. Theoretical properties of the TIGER method for the elliptical family is unclear.

Another closely related method is the rank-based CLIME method for estimating inverse correlation matrix estimation for the elliptical family [24]. The rank-based CLIME method is based on the formulation in (3) and cannot calibrate the regularization. Furthermore, the rank-based CLIME method can only estimate the inverse correlation matrix. Thus for applications such as the linear discriminant analysis (as is demonstrated in §6) which requires the input to be a precision matrix [2], [30], [35], the rank-based CLIME method is not applicable.

APPENDIX A
PROOF OF PROPOSITION III.1

Proof: To show the equivalence between (14) and (15), we only need to verify that the optimal solution $(\widehat{\mathbf{\Gamma}}_{*j}, \widehat{\tau}_j)$ to (15) satisfies

$$||\widehat{\mathbf{Z}}\widehat{\mathbf{\Gamma}}_{*j} - \mathbf{I}_{*j}||_\infty = \frac{c}{\nu}\widehat{\tau}_j. \tag{A.1}$$

We then prove (A.1) by contradiction. Assuming that there exists some $\bar{\tau}_j \geq 0$ such that

$$||\widehat{\mathbf{Z}}\widehat{\mathbf{\Gamma}}_{*j} - \mathbf{I}_{*j}||_\infty = \frac{c}{\nu}\bar{\tau}_j < \frac{c}{\nu}\widehat{\tau}_j, \tag{A.2}$$

(A.2) implies that $(\widehat{\mathbf{\Gamma}}_{*j}, \bar{\tau}_j)$ is also a feasible solution to (15) and

$$||\widehat{\mathbf{\Gamma}}_{*j}||_1 + c\bar{\tau}_j < ||\widehat{\mathbf{\Gamma}}_{*j}||_1 + c\widehat{\tau}_j. \tag{A.3}$$

(A.3) contradicts with the fact that $(\widehat{\mathbf{\Gamma}}_{*j}, \widehat{\tau}_j)$ minimizes (15). Thus (A.1) must hold, and (15) is equivalent to (14). □

APPENDIX B
PARAMETRIC SIMPLEX METHOD

We provide a brief description of the parametric simplex method only for self-containedness. More details of the derivation can be found in [34]. We consider the following generic form of linear program,

$$\max_{\boldsymbol{x}\in\mathbb{R}^m} \boldsymbol{c}^T\boldsymbol{x} \quad \text{s.t. } \mathbf{A}\boldsymbol{x} \leq \boldsymbol{b},\ \boldsymbol{x} \geq \mathbf{0}, \tag{B.1}$$

where $\boldsymbol{c} \in \mathbb{R}^m$, $\mathbf{A} \in \mathbb{R}^{n\times m}$, and $\boldsymbol{b} \in \mathbb{R}^n$. It is well known that (B.1) has a dual formulation as follows,

$$\min_{\boldsymbol{y}\in\mathbb{R}^n} \boldsymbol{b}^T\boldsymbol{y} \quad \text{s.t. } \mathbf{A}^T\boldsymbol{y} \geq \boldsymbol{c},\ \boldsymbol{y} \geq \mathbf{0}, \tag{B.2}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ are dual variables. The simplex method usually solves either (B.1) or (B.2). It contains two phases: Phase I is to find a feasible initial solution for Phase II; Phase II is an iterative procedure to recover the optimal solution based on the given initial solution.

Different from the simplex method, the parametric simplex method adds some perturbation to (B.1) and (B.2) such that the optimal solutions can be trivially obtained. More specifically, the parametric simplex method solves the following pair of linear programs

$$\max_{\boldsymbol{x}\in\mathbb{R}^m} (\boldsymbol{c} + \boldsymbol{\beta q})^T\boldsymbol{x} \text{ s.t. } \mathbf{A}\boldsymbol{x} \leq \boldsymbol{b} + \beta\boldsymbol{p},\ \boldsymbol{x} \geq \mathbf{0}, \tag{B.3}$$

$$\min_{\boldsymbol{y}\in\mathbb{R}^n} (\boldsymbol{b} + \beta\boldsymbol{p})^T\boldsymbol{y} \text{ s.t. } \mathbf{A}^T\boldsymbol{y} \geq \boldsymbol{c} + \boldsymbol{\beta q},\ \boldsymbol{y} \geq \mathbf{0}, \tag{B.4}$$

where $\beta \geq 0$ is a perturbation parameter, $\boldsymbol{p} \in \mathbb{R}^n$ and $\boldsymbol{q} \in \mathbb{R}^m$ are perturbation vectors. When β, $\boldsymbol{p}$, and $\boldsymbol{q}$ are suitably chosen such that $\boldsymbol{b} + \beta\boldsymbol{p} \geq \mathbf{0}$ and $\boldsymbol{c} + \boldsymbol{\beta q} \leq \mathbf{0}$, $\boldsymbol{x} = \mathbf{0}$ and $\boldsymbol{y} = \mathbf{0}$ are the optimal solutions to (B.3) and (B.4) respectively. The parametric simplex method is an iterative procedure, which gradually reduces β to 0 (corresponding to no perturbation) and eventually recovers the optimal solution to (B.1).

To derive the iterative procedure, we first add slack variables $\boldsymbol{w} = (w_1, \ldots, w_n)^T \in \mathbb{R}^n$, and rewrite (B.3) as

$$\max_{\widetilde{\boldsymbol{x}}\in\mathbb{R}^{m+n}} (\widetilde{\boldsymbol{c}} + \beta\widetilde{\boldsymbol{q}})^T\widetilde{\boldsymbol{x}} \text{ s.t. } \mathbf{H}\widetilde{\boldsymbol{x}} = \boldsymbol{b} + \beta\boldsymbol{p},\ \widetilde{\boldsymbol{x}} \geq \mathbf{0}. \tag{B.5}$$

where $\mathbf{H} = [\mathbf{A}\ \ \mathbf{I}]$, $\widetilde{\boldsymbol{c}} = (\boldsymbol{c}^T,\ \mathbf{0}^T)^T$, $\widetilde{\boldsymbol{q}} = (\boldsymbol{q}^T,\ \mathbf{0}^T)^T$ and

$$\begin{aligned}\widetilde{\boldsymbol{x}} &= (\widetilde{x}_1, \ldots, \widetilde{x}_m, \widetilde{x}_{m+1}, \ldots, \widetilde{x}_{m+n})^T \\ &= (x_1, \ldots, x_m, w_1, \ldots, w_n)^T \in \mathbb{R}^{m+n}.\end{aligned}$$

Since $\boldsymbol{b} + \beta\boldsymbol{p} \geq \mathbf{0}$ and $\boldsymbol{c} + \boldsymbol{\beta q} \leq \mathbf{0}$, $\widetilde{\boldsymbol{x}} = (\mathbf{0},\ \boldsymbol{b} + \beta\boldsymbol{p})^T$ is the optimal solution to (B.5). We then divide all variables in $\widetilde{\boldsymbol{x}}$ into a nonbasic group $\mathcal{N}$ and a basic group $\mathcal{B}$. In particular, $\widetilde{x}_1, .., \widetilde{x}_m$ belong to the nonbasic group denoted by $\widetilde{\boldsymbol{x}}_\mathcal{N}$, and $\widetilde{x}_{m+1}, \ldots, \widetilde{x}_{m+n}$ belong to the basic group denoted by $\widetilde{\boldsymbol{x}}_\mathcal{B}$. We also divide $\mathbf{H}$ into two submatrices $\mathbf{H}_\mathcal{N}$ and $\mathbf{H}_\mathcal{B}$, where $\mathbf{H}_\mathcal{N}$ contains all columns of $\mathbf{H}$ corresponding to $\widetilde{\boldsymbol{x}}_\mathcal{N}$, and $\mathbf{H}_\mathcal{B}$ contains all columns of $\mathbf{H}$ corresponding to $\widetilde{\boldsymbol{x}}_\mathcal{N}$. We then rewrite the constraint in (B.5) as $\mathbf{H}_\mathcal{N}\widetilde{\boldsymbol{x}}_\mathcal{N} + \mathbf{H}_\mathcal{B}\widetilde{\boldsymbol{x}}_\mathcal{B} = \boldsymbol{b} + \beta\boldsymbol{p}$. Consequently, we obtain the primal dictionary associated with the basic group $\mathcal{B}$ by

$$\widetilde{\boldsymbol{x}}_\mathcal{B} = \widetilde{\boldsymbol{x}}^*_\mathcal{B} + \beta\bar{\boldsymbol{x}}_\mathcal{B} - \mathbf{H}_\mathcal{B}^{-1}\mathbf{H}_\mathcal{N}, \tag{B.6}$$

$$\phi_P = \phi^* - (\widetilde{\boldsymbol{z}}^* + \beta\bar{\boldsymbol{z}})^T\widetilde{\boldsymbol{x}}_\mathcal{N}., \tag{B.7}$$

where $\widetilde{\boldsymbol{x}}^*_{\mathcal{B}} = \mathbf{H}_{\mathcal{B}}^{-1}\boldsymbol{b}$, $\bar{\boldsymbol{x}}_{\mathcal{B}} = \mathbf{H}_{\mathcal{B}}^{-1}\boldsymbol{p}$, $\phi^*_P = \widetilde{\boldsymbol{c}}^T_{\mathcal{B}}\mathbf{H}_{\mathcal{B}}\boldsymbol{b}$, $\widetilde{\boldsymbol{z}}^* = (\mathbf{H}_{\mathcal{B}}^{-1}\mathbf{H}_{\mathcal{N}})^T\widetilde{\boldsymbol{c}}_{\mathcal{B}} - \widetilde{\boldsymbol{c}}_{\mathcal{N}}$, $\bar{\boldsymbol{z}}_{\mathcal{N}} = (\mathbf{H}_{\mathcal{B}}^{-1}\mathbf{H}_{\mathcal{N}})^T\widetilde{\boldsymbol{q}}_{\mathcal{B}} - \widetilde{\boldsymbol{q}}_{\mathcal{N}}$, and ϕ_P is the objective value of (B.5) at current iteration.

We then add slack variables $\boldsymbol{z} = (z_1, \ldots, z_m)^T$, and rewrite (16) as

$$\min_{\boldsymbol{y}\in\mathbb{R}^n} (\widetilde{\boldsymbol{b}} + \beta\widetilde{\boldsymbol{p}})^T\boldsymbol{y}$$
$$\text{s.t.} \quad \mathbf{A} - \boldsymbol{z}^T\widetilde{\boldsymbol{y}} \geq \boldsymbol{c} + \boldsymbol{\beta q}, \quad \widetilde{\boldsymbol{y}} \geq \mathbf{0}. \tag{B.8}$$

To make the notation consistent with the primal problem, we define

$$\widetilde{\boldsymbol{z}} = (\widetilde{z}_1, \ldots, \widetilde{z}_m, \widetilde{z}_{m+1}, \ldots, \widetilde{z}_{m+n}) = (z_1, \ldots, z_m, y_1, \ldots, y_n)^T \in \mathbb{R}^{m+n}.$$

Similarly we can obtain the dual dictionary associated with the nonbasic variable $\mathcal{N}$ by

$$\widetilde{\boldsymbol{z}}_{\mathcal{N}} = (\widetilde{\boldsymbol{z}}^*_{\mathcal{N}} + \beta\bar{\boldsymbol{z}}_{\mathcal{N}}) + (\mathbf{H}_{\mathcal{B}}^{-1}\mathbf{H}_{\mathcal{N}})^T\widetilde{\boldsymbol{z}}_{\mathcal{B}}, \tag{B.9}$$
$$-\phi_D = -\phi^*_D - (\boldsymbol{x}^*_{\mathcal{B}} + \beta\bar{\boldsymbol{x}}_{\mathcal{B}})^T\widetilde{\boldsymbol{z}}_{\mathcal{B}}, \tag{B.10}$$

where $\phi^*_D = \widetilde{\boldsymbol{c}}^T_{\mathcal{B}}\mathbf{H}_{\mathcal{B}}^{-1}\boldsymbol{b}$, and ϕ_D is the objective value of the dual problem at current iteration.

Once we obtain (B.6), (B.7), (B.9), and (B.10), we start to decrease β, and the smallest value of β at current iteration is obtained by

$$\beta^* = \min\{\beta \mid \widetilde{\boldsymbol{x}}^*_{\mathcal{B}} + \beta\bar{\boldsymbol{x}}_{\mathcal{B}} \geq 0,\ \widetilde{\boldsymbol{z}}^*_{\mathcal{N}} + \beta\bar{\boldsymbol{z}}_{\mathcal{N}} \geq 0\}.$$

we then swap a pair of basic and nonbasic variables in $\mathcal{B}$ and $\mathcal{N}$ and update the primal and dual dictionaries such that β can be decreased to β^*. See more details on updating the dictionaries in [34]. By repeating the above procedure, we eventually decrease β to 0. The parametric simplex method guarantees the feasibility and optimality for both (B.3) and (B.4) in each iteration, and eventually obtain the optimal solution to the original problem (B.3).

Since the parametric simplex method starts with all zero solutions, it can recover the optimal solution only in a few iterations when the optimal solution is very sparse. That naturally fits into the sparse estimation problems such as the EPIC method. Moreover, if we rewrite (16) in the same form as (B.3), we need to set $\boldsymbol{p} = (\mathbf{0}^T, \boldsymbol{e}^T, 0)^T$ and start with $\beta = 1$. Since $\boldsymbol{c} = (-\mathbf{1}^T, -c)^T$, we can set $\boldsymbol{q} = \mathbf{0}$, i.e., we do not need perturbation on $\boldsymbol{c}$. Thus the computation in each iteration can be further simplified due to the sparsity of $\boldsymbol{p}$ and $\boldsymbol{q}$.

Remark B.1: For sparse estimation problems, Phase I of the simplex method does not guarantee the sparseness of the initial solution. As a result, Phase II may start with a dense initial solution, and gradually reduce the sparsity of the solution. Thus the overall convergence of the simplex method often requires a large number of iterations when the optimal solution is very sparse.

Appendix C
Smoothed Proximal Gradient Algorithm

We first apply the smoothing approach in [28] to obtain a smooth surrogate of the matrix ℓ_1 norm based on the Fenchel dual representation,

$$||\boldsymbol{\Omega} - \widetilde{\boldsymbol{\Omega}}||_\eta = \min_{||\mathbf{U}||_\infty \leq 1} \text{tr}(\mathbf{U}^T(\boldsymbol{\Omega} - \widetilde{\boldsymbol{\Omega}})) + \tfrac{\eta}{2}||\mathbf{U}||^2_{\text{F}}, \tag{C.1}$$

where $\eta > 0$ is a smoothing parameter. (C.1) has a closed form solution $\widehat{\mathbf{U}}$ as follows,

$$\widehat{\mathbf{U}}^{\boldsymbol{\Omega}}_{kj} = \text{sign}(\widetilde{\mathbf{U}}^{\boldsymbol{\Omega}}_{kj}) \cdot \max\{|\widetilde{\mathbf{U}}^{\boldsymbol{\Omega}}_{kj}| - \gamma_k\}. \tag{C.2}$$

where $\widetilde{\mathbf{U}}^{\boldsymbol{\Omega}} = (\boldsymbol{\Omega} - \widetilde{\boldsymbol{\Omega}})/\eta$, and γ_k is the minimum positive value such that $||\widehat{\mathbf{U}}||_\infty = \max_k ||\widehat{\mathbf{U}}_{k*}||_1 \leq 1$. See [9] for an efficient algorithms to find γ_k with the average computational complexity of $O(d^2)$. As is shown in [28], the smooth surrogate $||\boldsymbol{\Omega} - \widetilde{\boldsymbol{\Omega}}||_\eta$ is smooth, convex, and has a simple form gradient as

$$\mathbf{G}(\boldsymbol{\Omega}) = \frac{\partial||\boldsymbol{\Omega} - \widetilde{\boldsymbol{\Omega}}||_\eta}{\partial\boldsymbol{\Omega}} = \widehat{\mathbf{U}}^{\boldsymbol{\Omega}}_{kj}.$$

Since $\widehat{\mathbf{U}}^{\boldsymbol{\Omega}}_{kj}$ is obtained by the soft-thresholding in (C.2), we have $\mathbf{G}(\boldsymbol{\Omega})$ continuous in $\boldsymbol{\Omega}$ with the Lipschitz constant η^{-1}. Motivated by these good computational properties, we consider the following optimization problem instead of (18),

$$\overline{\boldsymbol{\Omega}} = \underset{\boldsymbol{\Omega}=\boldsymbol{\Omega}^T}{\text{argmin}} ||\boldsymbol{\Omega} - \widetilde{\boldsymbol{\Omega}}||_\eta. \tag{C.3}$$

To solve (C.3), we adopt the accelerated projected gradient algorithm proposed in [27]. More specifically, we define two sequences of auxiliary variables $\{\mathbf{M}^{(t)}\}$ and $\{\mathbf{W}^{(t)}\}$ with $\mathbf{M}^{(0)} = \mathbf{W}^{(0)} = \boldsymbol{\Omega}^{(0)}$, and a sequence of weights $\{\theta_t = 2/(1+t)\}$. For the t^{th} iteration, we first calculate the auxiliary variable $\mathbf{M}^{(t)}$ as

$$\mathbf{M}^{(t)} = (1-\theta_t)\boldsymbol{\Omega}^{(t-1)} + \theta_t\mathbf{W}^{(t-1)}.$$

We then calculate the auxiliary variable $\mathbf{W}^{(t)}$ as

$$\mathbf{W}^{(t)} = \underset{\mathbf{W}=\mathbf{W}^T}{\text{argmin}} ||\mathbf{W}^{(t-1)} - \widetilde{\boldsymbol{\Omega}}||_\eta + \text{tr}\left((\mathbf{W} - \mathbf{W}^{(t-1)})^T\mathbf{G}(\mathbf{M}^{(t)})\right) + \frac{1}{2\eta_t\theta_t}||\mathbf{W} - \mathbf{W}^{(t-1)}||^2_{\text{F}} = \frac{1}{2}\left(\mathbf{W}^{(t-1)} - \frac{\eta_t}{\theta_t}\mathbf{G}(\mathbf{M}^{(t)})\right) + \frac{1}{2}\left(\mathbf{W}^{(t-1)} - \frac{\eta_t}{\theta_t}\mathbf{G}(\mathbf{M}^{(t)})\right)^T,$$

where η_t is the step size. We can either choose $\eta_t = \eta$ in all iterations or estimate η_t's by the back-tracking line search for better empirical performance [4]. At last, we calculate $\boldsymbol{\Omega}^{(t)}$ as,

$$\boldsymbol{\Omega}^{(t)} = (1-\theta_t)\boldsymbol{\Omega}^{(t-1)} + \theta_t\mathbf{W}^{(t)}.$$

The next theorem provides the convergence rate of the algorithm with respect to minimizing (18).

Theorem C.1: Given the desired accuracy ε such that $||\boldsymbol{\Omega}^{(t)} - \widetilde{\boldsymbol{\Omega}}||_1 - ||\widehat{\boldsymbol{\Omega}} - \widetilde{\boldsymbol{\Omega}}||_1 < \varepsilon$, let $\eta = d^{-1}\varepsilon/2$, we need the number of iterations to be at most

$$t = 2\sqrt{2d}||\boldsymbol{\Omega}^{(0)} - \overline{\boldsymbol{\Omega}}||_{\text{F}} \cdot \varepsilon^{-1} - 1 = O\left(\varepsilon^{-1}\right).$$

Proof: Due to the fact that $||\mathbf{A}||_{\text{F}} \leq d||\mathbf{A}||_\infty$, a direct consequence of (C.1) is the following uniform bound

$$||\boldsymbol{\Omega} - \widetilde{\boldsymbol{\Omega}}||_1 - d\eta \leq ||\boldsymbol{\Omega} - \widetilde{\boldsymbol{\Omega}}||_\eta \leq ||\boldsymbol{\Omega} - \widetilde{\boldsymbol{\Omega}}||_1.$$

Then we consider the following decomposition

$$\begin{aligned}
&||\mathbf{\Omega}^{(t)} - \widetilde{\mathbf{\Omega}}||_1 - ||\widehat{\mathbf{\Omega}} - \widetilde{\mathbf{\Omega}}||_1 \\
&= ||\mathbf{\Omega}^{(t)} - \widetilde{\mathbf{\Omega}}||_1 - ||\overline{\mathbf{\Omega}} - \widetilde{\mathbf{\Omega}}||_\eta + ||\overline{\mathbf{\Omega}} - \widetilde{\mathbf{\Omega}}||_\eta - ||\widehat{\mathbf{\Omega}} - \widetilde{\mathbf{\Omega}}||_1 \\
&\leq ||\mathbf{\Omega}^{(t)} - \widetilde{\mathbf{\Omega}}||_\eta - ||\overline{\mathbf{\Omega}} - \widetilde{\mathbf{\Omega}}||_\eta + d\eta \\
&\leq \frac{2||\mathbf{\Omega}^{(t)} - \widetilde{\mathbf{\Omega}}||_{\mathrm{F}}^2}{(t+1)^2\eta} + d\eta,
\end{aligned}$$

where the last inequality comes from the result established in [27],

$$||\mathbf{\Omega}^{(t)} - \widetilde{\mathbf{\Omega}}||_\eta - ||\overline{\mathbf{\Omega}} - \widetilde{\mathbf{\Omega}}||_\eta \leq \frac{2||\mathbf{\Omega}^{(0)} - \overline{\mathbf{\Omega}}||_{\mathrm{F}}^2}{(t+1)^2\eta}.$$

Thus given $d\eta = \epsilon/2$, we only need

$$\frac{4d||\mathbf{\Omega}^{(0)} - \overline{\mathbf{\Omega}}||_{\mathrm{F}}^2}{(t+1)^2} \leq \frac{\epsilon^2}{2}. \tag{C.4}$$

By solving (C.4), we obtain

$$t \leq \frac{2\sqrt{2d}||\mathbf{\Omega}^{(0)} - \overline{\mathbf{\Omega}}||_{\mathrm{F}}}{\epsilon} - 1. \qquad \square$$

Theorem C.1 guarantees that the above algorithm achieves the optimal rate of convergence for minimizing (18) over the class of all first-order computational algorithms.

Appendix D
Proof of Lemma 1

Proof: [7] shows that there exist universal constants κ_3 and κ_4 such that

$$\mathbb{P}\left(|\widehat{\mu}_j - \mu_j| \leq \kappa_3\theta_{\max}\epsilon\right) \geq 1 - \exp(-n\epsilon^2), \tag{D.1}$$

$$\mathbb{P}\left(|\widehat{m}_j - \mathbb{E}X_j^2| \leq \kappa_4\sqrt{K}\epsilon\right) \geq 1 - \exp(-n\epsilon^2). \tag{D.2}$$

We then define the following events

$$\begin{aligned}
\mathcal{C}_1 &= \{|\widehat{\mu}_j - \mu_j| \leq \mu_{\max}\}, \\
\mathcal{C}_2 &= \{|\widehat{\mu}_j - \mu_j| \leq \kappa_3\theta_{\max}\epsilon\}, \\
\mathcal{C}_3 &= \{|\widehat{m}_j - \mathbb{E}X_j^2| \leq \kappa_4\sqrt{K}\epsilon\}, \\
\mathcal{C}_4 &= \{|\widehat{\theta}_j - \theta_j| \leq \theta_{\min}\}.
\end{aligned}$$

Conditioning on $\mathcal{C}_1$, we have

$$\begin{aligned}
|\widehat{\mu}_j^2 - \mu_j^2| &= |\widehat{\mu}_j - \mu_j| \cdot |\widehat{\mu}_j + \mu_j| \\
&\leq |\widehat{\mu}_j - \mu_j| \cdot |\widehat{\mu}_j - \mu_j| + |\widehat{\mu}_j - \mu_j| \cdot |2\mu_j| \\
&\leq (2\mu_{\max} + |\widehat{\mu}_j - \mu_j|)|\widehat{\mu}_j - \mu_j| \\
&\leq 3\mu_{\max}|\widehat{\mu}_j - \mu_j|.
\end{aligned} \tag{D.3}$$

Conditioning on $\mathcal{C}_2$ and $\mathcal{C}_3$, (D.3) implies

$$\begin{aligned}
|\widehat{\theta}_j^2 - \theta_j^2| &= |\widehat{m}_j - \widehat{\mu}_j^2 - \mathbb{E}X_j^2 + \mu_j^2| \\
&\leq |\widehat{m}_j - \mathbb{E}X_j^2| + |\widehat{\mu}_j^2 - \mu_j^2| \\
&\leq (3\mu_{\max}\kappa_3\theta_{\max} + \kappa_4\sqrt{K})\epsilon.
\end{aligned} \tag{D.4}$$

(D.4) further implies

$$\begin{aligned}
|\widehat{\theta}_j - \theta_j| &\leq \frac{|\widehat{\theta}_j^2 - \theta_j^2|}{\widehat{\theta}_j + \theta_j} \leq \frac{|\widehat{\theta}_j^2 - \theta_j^2|}{\theta_j} \\
&\leq \frac{(3\mu_{\max}\kappa_3\theta_{\max} + \kappa_4\sqrt{K})\epsilon}{\theta_{\min}}.
\end{aligned} \tag{D.5}$$

Conditioning $\mathcal{C}_4$, (D.5) implies

$$\begin{aligned}
|\widehat{\theta}_j^{-1} - \theta_j^{-1}| &= \frac{|\widehat{\theta}_j - \theta_j|}{\widehat{\theta}_j\theta_j} = \frac{|\widehat{\theta}_j - \theta_j|}{(\widehat{\theta}_j - \theta_j)\theta_j + 2\theta_j^2} \\
&\leq \frac{(3\mu_{\max}\kappa_3\theta_{\max} + \kappa_4\sqrt{K})\epsilon}{(\widehat{\theta}_j - \theta_j)\theta_j\theta_{\min} + 2\theta_j^2\theta_{\min}} \\
&\leq \frac{(3\mu_{\max}\kappa_3\theta_{\max} + \kappa_4\sqrt{K})\epsilon}{2\theta_j^2\theta_{\min}} \\
&\leq \frac{(3\mu_{\max}\kappa_3\theta_{\max} + \kappa_4\sqrt{K})\epsilon}{2\theta_{\min}^3}.
\end{aligned} \tag{D.6}$$

Combining (D.1), (D.2), and (D.6), for small enough ϵ such that

$$\epsilon \leq \min\left\{\frac{\mu_{\max}}{\kappa_3\theta_{\max}}, \frac{\theta_{\min}^2}{3\mu_{\max}\kappa_3\theta_{\max} + \kappa_4\sqrt{K}}\right\}, \tag{D.7}$$

we have

$$\mathbb{P}\left(|\widehat{\theta}_j^{-1} - \theta_j^{-1}| \leq \frac{(3\mu_{\max}\kappa_3\theta_{\max} + \kappa_4\sqrt{K})\epsilon}{\theta_{\min}^3}\right) \geq 1 - 2\exp(-4n\epsilon^2). \tag{D.8}$$

By taking the union bound of (D.8), we have

$$\begin{aligned}
&\mathbb{P}\left(\max_{1\leq j\leq d}|\widehat{\theta}_j^{-1} - \theta_j^{-1}| \leq \frac{(3\mu_{\max}\kappa_3\theta_{\max} + \kappa_4\sqrt{K})\epsilon}{\theta_{\min}^3}\right) \\
&\qquad \geq 1 - 2\exp(-4n\epsilon^2 + \log d).
\end{aligned}$$

If we take $\epsilon = \sqrt{\log d/n}$, then (D.7) implies that we need n large enough such that

$$n \geq \max\left\{\frac{\kappa_3^2\theta_{\max}^2}{\mu_{\max}^2}, \frac{(3\mu_{\max}\kappa_3\theta_{\max} + \kappa_4\sqrt{K})^2}{\theta_{\min}^4}\right\} \cdot \log d.$$

Taking $\kappa_2 = (3\mu_{\max}\kappa_3\theta_{\max} + \kappa_4\sqrt{K})/\theta_{\min}^3$, we then have

$$\mathbb{P}\left(\max_{1\leq j\leq d}|\widehat{\theta}_j^{-1} - \theta_j^{-1}| \leq \kappa_2\sqrt{\frac{\log d}{n}}\right) \geq 1 - \frac{2}{d}.$$

(19) is a direct result in [24], therefore its proof is omitted. $\square$

Appendix E
Proof of Theorem IV.1

Proof: We first define the following pair of orthogonal subspaces $(\mathcal{S}_j,\ \mathcal{S}_j^\perp)$,

$$\begin{aligned}
\mathcal{S}_j &= \left\{\boldsymbol{v} \in \mathbb{R}^d \mid v_k = 0 \text{ for all } \mathbf{\Gamma}_{kj} = 0\right\}, \\
\mathcal{S}_j^\perp &= \left\{\boldsymbol{v} \in \mathbb{R}^d \mid v_k = 0 \text{ for all } \mathbf{\Gamma}_{kj} \neq 0\right\}.
\end{aligned}$$

We will use $(\mathcal{S}_j,\ \mathcal{S}_j^\perp)$ to exploit the sparseness of $\mathbf{\Gamma}_{*j}$. We then define the following event

$$\mathcal{D}_1 = \left\{||\widehat{\mathbf{Z}} - \mathbf{Z}||_{\max} \leq \lambda\right\}.$$

Conditioning on $\mathcal{D}_1$, we have

$$\begin{aligned}
||\widehat{\mathbf{Z}}\mathbf{\Gamma}_{*j} - \mathbf{I}_{*j}||_\infty &= ||(\widehat{\mathbf{Z}} - \mathbf{Z})\mathbf{\Gamma}_{*j}||_\infty \\
&\leq ||\mathbf{\Gamma}_{*j}||_1||\widehat{\mathbf{Z}} - \mathbf{Z}||_{\max} \leq \lambda||\mathbf{\Gamma}_{*j}||_1.
\end{aligned} \tag{E.1}$$

Now let $\tau_j = ||\mathbf{\Gamma}_{*j}||_1$, (E.1) implies that $\left(\mathbf{\Gamma}_{*j},\ \tau_j\right)$ is a feasible solution to (13). Since $\left(\widehat{\mathbf{\Gamma}}_{*j},\ \widehat{\tau}_j\right)$ is the empirical minimizer, we have

$$\begin{aligned}||\widehat{\mathbf{\Gamma}}_{\mathcal{S}_j j}||_1 + ||\widehat{\mathbf{\Gamma}}_{\mathcal{S}_j^\perp j}||_1 + c\widehat{\tau}_j &= ||\widehat{\mathbf{\Gamma}}_{*j}||_1 + c\widehat{\tau}_j \\ &\le ||\mathbf{\Gamma}_{*j}||_1 + c\tau_j = ||\mathbf{\Gamma}_{\mathcal{S}_j j}||_1 + c\tau_j, \end{aligned} \tag{E.2}$$

where the last equality comes from the fact that $\mathbf{\Gamma}_{\mathcal{S}_j^\perp j} = \mathbf{0}$. Let $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}$ be the estimation error, (E.2) implies

$$\begin{aligned}||\widehat{\mathbf{\Gamma}}_{\mathcal{S}_j^\perp j}||_1 &\le ||\mathbf{\Gamma}_{\mathcal{S}_j j}||_1 - ||\widehat{\mathbf{\Gamma}}_{\mathcal{S}_j j}||_1 + c(\tau_j - \widehat{\tau}_j) \\ &\le ||\widehat{\mathbf{\Delta}}_{\mathcal{S}_j j}||_1 + c(\tau_j - \widehat{\tau}_j) \\ &\le ||\widehat{\mathbf{\Delta}}_{\mathcal{S}_j j}||_1 + c(||\mathbf{\Gamma}_{*j}||_1 - ||\widehat{\mathbf{\Gamma}}_{*j}||_1) \\ &\overset{\text{(i)}}{\le} ||\widehat{\mathbf{\Delta}}_{\mathcal{S}_j j}||_1 + c||\widehat{\mathbf{\Delta}}_{*j}||_1 \\ &\overset{\text{(ii)}}{\le} (1+c)||\widehat{\mathbf{\Delta}}_{\mathcal{S}_j j}||_1 + c||\widehat{\mathbf{\Delta}}_{\mathcal{S}_j^\perp j}||_1, \end{aligned} \tag{E.3}$$

where (i) comes from the constraint in (13):$||\widehat{\mathbf{\Gamma}}_{*j}||_1 \le \widehat{\tau}_j$, and (ii) comes from the fact $||\widehat{\mathbf{\Delta}}_{*j}||_1 = ||\widehat{\mathbf{\Delta}}_{\mathcal{S}_j j}||_1 + ||\widehat{\mathbf{\Delta}}_{\mathcal{S}_j^\perp j}||_1$. Combining the fact $||\widehat{\mathbf{\Delta}}_{\mathcal{S}_j^\perp j}||_1 = ||\widehat{\mathbf{\Gamma}}_{\mathcal{S}_j^\perp j} - \mathbf{0}||_1 = ||\widehat{\mathbf{\Gamma}}_{\mathcal{S}_j^\perp j}||_1$ with (E.3), we have

$$||\widehat{\mathbf{\Delta}}_{\mathcal{S}_j^\perp j}||_1 \le \bar{c}||\widehat{\mathbf{\Delta}}_{\mathcal{S}_j j}||_1, \tag{E.4}$$

where $\bar{c} = (1+c)/(1-c)$. (E.4) implies that $\widehat{\mathbf{\Delta}}_{*j}$ belongs to the following cone shape set

$$\mathcal{M}_j^{\bar{c}} = \left\{ \boldsymbol{v} \in \mathbb{R}^d \setminus \{\mathbf{0}\} \;\middle|\; ||\boldsymbol{v}_{\mathcal{S}_j^\perp}||_1 \le \bar{c}||\boldsymbol{v}_{\mathcal{S}_j}||_1 \right\}.$$

The following lemma characterizes an important property of $\mathcal{M}_j^{\bar{c}}$ when $\mathcal{D}_1$ holds.

Lemma E.1: Suppose that $\boldsymbol{X} \sim EC(\boldsymbol{\mu}, \xi, \mathbf{\Sigma})$, and (A.1) and $\mathcal{D}_1$ hold. Given any $\boldsymbol{v} \in \mathcal{M}_j^{\bar{c}}$, for small enough λ such that $2(1+\bar{c})^2 s\lambda\kappa_u \le 1$, we have

$$\min_{\boldsymbol{v} \in \mathcal{M}_j^{\bar{c}}} \boldsymbol{v}^T \widehat{\mathbf{Z}} \boldsymbol{v} \ge \frac{||\boldsymbol{v}||_2^2}{2\kappa_u}. \tag{E.5}$$

The proof of Lemma E.1 is provided in Appendix E.1. Since $\mathbf{\Delta}_{*j}$ exactly belongs to $\mathcal{M}_j^{\bar{c}}$, we have a simple variant of (E.5) as

$$\begin{aligned}||\widehat{\mathbf{\Delta}}_{*j}||_1 ||\widehat{\mathbf{Z}}\widehat{\mathbf{\Delta}}_{*j}||_\infty \ge \widehat{\mathbf{\Delta}}_{*j}^T \widehat{\mathbf{Z}}\widehat{\mathbf{\Delta}}_{*j} &\ge \frac{||\widehat{\mathbf{\Delta}}_{*j}||_2^2}{2\kappa_u} \\ &\ge \frac{||\widehat{\mathbf{\Delta}}_{\mathcal{S}_j j}||_2^2}{2\kappa_u} \ge \frac{||\widehat{\mathbf{\Delta}}_{\mathcal{S}_j j}||_1^2}{2s\kappa_u}, \end{aligned} \tag{E.6}$$

where the last inequality comes from the fact that $\widehat{\mathbf{\Delta}}_{\mathcal{S}_j j}$ has at most s nonzero entries. Since

$$\begin{aligned}||\widehat{\mathbf{Z}}\widehat{\mathbf{\Delta}}_{*j}||_\infty &\le ||\widehat{\mathbf{Z}}\widehat{\mathbf{\Gamma}}_{*j} - \mathbf{I}_{*j}||_\infty + ||\widehat{\mathbf{Z}}\mathbf{\Gamma}_{*j} - \mathbf{I}_{*j}||_\infty \\ &\le \lambda\left(\widehat{\tau}_j + \tau_j\right) \\ &\le \lambda\left(2\widehat{\tau}_j + \tau_j - \widehat{\tau}_j\right) \\ &\le \lambda\left(2\widehat{\tau}_j + ||\widehat{\mathbf{\Delta}}_{*j}||_1\right) \\ &\le \lambda\left(2\widehat{\tau}_j + (1+\bar{c})||\widehat{\mathbf{\Delta}}_{\mathcal{S}_j j}||_1\right), \end{aligned} \tag{E.7}$$

where the last inequality comes from (E.4). Combining (E.6) and (E.7), we have

$$||\widehat{\mathbf{Z}}\widehat{\mathbf{\Delta}}_{*j}||_\infty \le \lambda\left(2\widehat{\tau}_j + 2(1+\bar{c})\kappa_u s||\widehat{\mathbf{Z}}\widehat{\mathbf{\Delta}}_{*j}||_\infty\right). \tag{E.8}$$

Assuming that $1 - 2(1+\bar{c})\kappa_u s\lambda = \delta_1 > 0$, (E.8) implies

$$||\widehat{\mathbf{Z}}\widehat{\mathbf{\Delta}}_{*j}||_\infty \le 2\delta_1^{-1}\lambda\widehat{\tau}_j. \tag{E.9}$$

Combining (E.6) and (E.9), we have

$$\begin{aligned}c\widehat{\tau}_j &\le ||\widehat{\mathbf{\Delta}}_{\mathcal{S}_j j}||_1 + c\tau_j \le 2\kappa_u s||\widehat{\mathbf{Z}}\widehat{\mathbf{\Delta}}_{*j}||_\infty + c\tau_j \\ &\le 4\kappa_u s\delta_1^{-1}\lambda\widehat{\tau}_j + c\tau_j. \end{aligned} \tag{E.10}$$

Assuming that $1 - 4\kappa_u s\delta_1^{-1}c^{-1}\lambda = \delta_2 > 0$, (E.10) implies

$$\widehat{\tau}_j \le \delta_2^{-1}\tau_j. \tag{E.11}$$

Recall $\lambda = \kappa_1\sqrt{\log d/n}$, in order to secure

$$\begin{aligned}1 - 2(1+\bar{c})\kappa_u s\lambda &= \delta_1 > 0, \\ c - 4\kappa_u s\delta_1^{-1}\lambda &= \delta_2 > 0, \\ 2(1+\bar{c})^2 s\lambda\kappa_u &\le 1, \end{aligned}$$

we need large enough n such that

$$\begin{aligned}n \ge \max\Big\{&4(1-\delta_1)^{-2}(1+\bar{c})^2\kappa_u^2, 16(c-\delta_2)^{-2}\kappa_u^2\delta_1^{-2}, \\ &4(1+\bar{c})^4\kappa_u\kappa_1\Big\} \cdot s^2 \log d. \end{aligned}$$

Combining (E.9) and (E.11), we have

$$||\widehat{\mathbf{Z}}\widehat{\mathbf{\Delta}}_{*j}||_\infty \le 2\delta_1^{-1}\delta_2^{-1}\lambda\tau_j. \tag{E.12}$$

Combining (E.4), (E.6), and (E.12), we obtain

$$||\widehat{\mathbf{\Delta}}_{*j}||_1 \le 4(1+\bar{c})\kappa_u\delta_1^{-1}\delta_2^{-1}s\lambda\tau_j. \tag{E.13}$$

Combining (E.12) and (E.13), we have

$$\begin{aligned}\widehat{\mathbf{\Delta}}_{*j}^T\widehat{\mathbf{Z}}\widehat{\mathbf{\Delta}}_{*j} &\le ||\widehat{\mathbf{\Delta}}_{*j}||_1 \cdot ||\widehat{\mathbf{Z}}\widehat{\mathbf{\Delta}}_{*j}||_\infty \\ &\le 8(1+\bar{c})\kappa_u\delta_1^{-2}\delta_2^{-2}s\lambda^2\tau_j^2. \end{aligned} \tag{E.14}$$

By Lemma E.1 again, (E.14) implies

$$||\widehat{\mathbf{\Delta}}_{*j}||_2^2 \le 16(1+\bar{c})\kappa_u^2\delta_1^{-2}\delta_2^{-2}s\lambda^2\tau_j^2. \tag{E.15}$$

Let $\kappa_5 = 4(1+\bar{c})\kappa_u\delta_1^{-1}\delta_2^{-1}\kappa_1$ and $\kappa_6 = 16(1+\bar{c})\kappa_u^2\delta_1^{-2}\delta_2^{-2}\kappa_1^2$. Recall $\lambda = \kappa_1\sqrt{\log d/n}$, by definition of the matrix ℓ_1 and Frobenius norms, (E.13) and (E.15) imply

$$\begin{aligned}||\widehat{\mathbf{\Delta}}||_1 &= \max_j ||\widehat{\mathbf{\Delta}}_{*j}||_1 \le 4(1+\bar{c})\kappa_u\delta_1^{-1}\delta_2^{-1}s\lambda\max_j\tau_j \\ &\le \kappa_5 \cdot M \cdot s\sqrt{\frac{\log d}{n}}, \end{aligned} \tag{E.16}$$

and

$$\begin{aligned}\frac{1}{d}||\widehat{\mathbf{\Delta}}||_{\mathrm{F}}^2 = \frac{1}{d}\sum_j ||\widehat{\mathbf{\Delta}}_{*j}||_2^2 &\le 16(1+\bar{c})\kappa_u^2\delta_1^{-2}\delta_2^{-2}s\lambda^2\tau_j^2 \\ &\le \kappa_6 \cdot M^2 \cdot \frac{s\log d}{n}. \end{aligned} \tag{E.17}$$

Now we start to derive the error bound of $\widetilde{\mathbf{\Omega}}$ obtained by the ensemble rule. We have the following decomposition

$$\begin{aligned}\widetilde{\mathbf{\Omega}}-\mathbf{\Omega}&=\widehat{\mathbf{\Theta}}^{-1}\widehat{\mathbf{\Gamma}}\widehat{\mathbf{\Theta}}^{-1}-\mathbf{\Theta}^{-1}\mathbf{\Gamma}\mathbf{\Theta}^{-1}\\&=(\widehat{\mathbf{\Theta}}^{-1}-\mathbf{\Theta}^{-1}+\mathbf{\Theta}^{-1})(\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}+\mathbf{\Gamma})\\&\quad\cdot(\widehat{\mathbf{\Theta}}^{-1}-\mathbf{\Theta}^{-1}+\mathbf{\Theta}^{-1})-\mathbf{\Theta}^{-1}\mathbf{\Gamma}\mathbf{\Theta}^{-1}\\&=(\widehat{\mathbf{\Theta}}^{-1}-\mathbf{\Theta}^{-1})(\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma})(\widehat{\mathbf{\Theta}}^{-1}-\mathbf{\Theta}^{-1})\\&\quad+(\widehat{\mathbf{\Theta}}^{-1}-\mathbf{\Theta}^{-1})(\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma})\mathbf{\Theta}^{-1}+(\widehat{\mathbf{\Theta}}^{-1}-\mathbf{\Theta}^{-1})\mathbf{\Gamma}\mathbf{\Theta}^{-1}\\&\quad+(\widehat{\mathbf{\Theta}}^{-1}-\mathbf{\Theta}^{-1})\mathbf{\Gamma}(\widehat{\mathbf{\Theta}}^{-1}-\mathbf{\Theta}^{-1})+\mathbf{\Theta}^{-1}(\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma})\mathbf{\Theta}^{-1}\\&\quad+\mathbf{\Theta}^{-1}(\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma})(\widehat{\mathbf{\Theta}}^{-1}-\mathbf{\Theta}^{-1})\\&\quad+\mathbf{\Theta}^{-1}\mathbf{\Gamma}(\widehat{\mathbf{\Theta}}^{-1}-\mathbf{\Theta}^{-1}).\end{aligned}\tag{E.18}$$

Moreover, for any $\mathbf{A},\mathbf{B},\mathbf{C}\in\mathbb{R}^{d\times d}$, where $\mathbf{A}$ and $\mathbf{C}$ are diagonal matrices, we have

$$||\mathbf{ABC}||_1\le||\mathbf{A}||_{\max}\cdot||\mathbf{B}||_1\cdot||\mathbf{C}||_{\max},\tag{E.19}$$

$$||\mathbf{ABC}||_{\mathrm{F}}\le||\mathbf{A}||_{\max}\cdot||\mathbf{B}||_{\mathrm{F}}\cdot||\mathbf{C}||_{\max}.\tag{E.20}$$

Here we define the following event

$$\mathcal{D}_2=\left\{||\widehat{\mathbf{\Theta}}^{-1}-\mathbf{\Theta}^{-1}||_{\max}\le\kappa_2\sqrt{\frac{\log d}{n}}\right\}.$$

Thus conditioning $\mathcal{D}_2$, (E.16), (E.18), and (E.19) imply

$$\begin{aligned}||\widetilde{\mathbf{\Omega}}-\mathbf{\Omega}||_1\le&\kappa_2^2\kappa_5\cdot\frac{\log d}{n}\cdot M\cdot s\sqrt{\frac{\log d}{n}}\\&+\frac{\kappa_2\kappa_5}{\theta_{\min}}\cdot\sqrt{\frac{\log d}{n}}\cdot M\cdot s\sqrt{\frac{\log d}{n}}\\&+\kappa_2^2\cdot M\cdot\frac{\log d}{n}+\frac{\kappa_2}{\theta_{\min}}\cdot M\cdot\sqrt{\frac{\log d}{n}}\\&+\frac{\kappa_2\kappa_5}{\theta_{\min}}\cdot\sqrt{\frac{\log d}{n}}\cdot M\cdot s\sqrt{\frac{\log d}{n}}\\&+\frac{\kappa_5}{\theta_{\min}^2}M\cdot s\sqrt{\frac{\log d}{n}}+\frac{\kappa_2}{\theta_{\min}}\cdot M\cdot\sqrt{\frac{\log d}{n}}.\end{aligned}\tag{E.21}$$

If (A.4): $s^2\log d/n\to 0$ holds, then (E.21) is determined by the slowest rate $Ms\sqrt{\log d/n}$. Thus for large enough n, there exists a universal constant C_4 such that

$$||\widetilde{\mathbf{\Omega}}-\mathbf{\Omega}||_1\le C_4\cdot M\cdot s\sqrt{\frac{\log d}{n}}.\tag{E.22}$$

Similarly, conditioning on $\mathcal{D}_2$, (E.17), (E.18), (E.20) and the fact $||\mathbf{\Gamma}||_{\mathrm{F}}\le M\sqrt{d}$ imply

$$\begin{aligned}||\widetilde{\mathbf{\Omega}}-\mathbf{\Omega}||_{\mathrm{F}}\le&\kappa_2^2\kappa_6\cdot\frac{\log d}{n}\cdot M\cdot\sqrt{\frac{ds\log d}{n}}\\&+\frac{\kappa_2\kappa_6}{\theta_{\min}}\cdot\sqrt{\frac{\log d}{n}}\cdot M\cdot\sqrt{\frac{ds\log d}{n}}\\&+\kappa_2^2\cdot M\sqrt{d}\cdot\frac{\log d}{n}+\frac{\kappa_2}{\theta_{\min}}\cdot M\sqrt{d}\cdot\sqrt{\frac{\log d}{n}}\\&+\frac{\kappa_2\kappa_6}{\theta_{\min}}\cdot\sqrt{\frac{\log d}{n}}\cdot M\cdot\sqrt{\frac{ds\log d}{n}}\\&+\frac{\kappa_6}{\theta_{\min}^2}M\cdot\sqrt{\frac{ds\log d}{n}}+\frac{\kappa_2}{\theta_{\min}}\cdot M\sqrt{d}\cdot\sqrt{\frac{\log d}{n}}.\end{aligned}\tag{E.23}$$

Again if (A.4) holds, then (E.23) is determined by the slowest rate $M\sqrt{ds\log d/n}$. Thus for large enough n, there exists a universal constant C_2 such that

$$\frac{1}{d}||\widetilde{\mathbf{\Omega}}-\mathbf{\Omega}||_{\mathrm{F}}^2\le C_2\cdot M^2\cdot\frac{s\log d}{n}.\tag{E.24}$$

We then proceed to prove the error bound of $\widehat{\mathbf{\Omega}}$ obtained by the symmetrization procedure (18). Let $C_1=2C_4$, if we choose the matrix ℓ_1 norm as $||\cdot||_*$ in (18), we have

$$\begin{aligned}||\widehat{\mathbf{\Omega}}-\mathbf{\Omega}||_1&\le||\widetilde{\mathbf{\Omega}}-\widehat{\mathbf{\Omega}}||_1+||\widetilde{\mathbf{\Omega}}-\mathbf{\Omega}||_1\\&\le 2||\widetilde{\mathbf{\Omega}}-\mathbf{\Omega}||_1\le C_1\cdot M\cdot s\sqrt{\frac{\log d}{n}},\end{aligned}\tag{E.25}$$

where the second inequality comes from the fact that $\mathbf{\Omega}$ is a feasible solution to (18), and $\widehat{\mathbf{\Omega}}$ is the empirical minimizer. If we choose the Frobenius norm as $||\cdot||_*$ in (18), using the fact that the Frobenius norm projection is contractive, we have

$$\frac{1}{d}||\widehat{\mathbf{\Omega}}-\mathbf{\Omega}||_{\mathrm{F}}^2\le\frac{1}{d}||\widetilde{\mathbf{\Omega}}-\mathbf{\Omega}||_{\mathrm{F}}^2\le C_2M^2\frac{s\log d}{n}.\tag{E.26}$$

All above analysis are conditioned on $\mathcal{D}_1$ and $\mathcal{D}_2$. Thus combining Lemma 1 with (E.25) and (E.26), we have

$$\mathbb{P}\left(||\widehat{\mathbf{\Omega}}-\mathbf{\Omega}||_p\le C_1Ms\sqrt{\frac{\log d}{n}}\right)\ge 1-\frac{3}{d},\tag{E.27}$$

$$\mathbb{P}\left(\frac{1}{d}||\widehat{\mathbf{\Omega}}-\mathbf{\Omega}||_{\mathrm{F}}^2\le C_2M^2\frac{s\log d}{n}\right)\ge 1-\frac{3}{d}.\tag{E.28}$$

where $p=1,2$, and (E.27) comes from the fact that $||\mathbf{A}||_2\le||\mathbf{A}||_1$ for any symmetric matrix $\mathbf{A}$. □

Appendix F
Proof of Lemma E.1

Proof: Since $\Lambda_{\min}(\mathbf{Z})=1/\Lambda_{\max}(\mathbf{\Gamma})\ge\kappa_u^{-1}$, we have

$$\begin{aligned}\boldsymbol{v}^T\widehat{\mathbf{Z}}\boldsymbol{v}&=\boldsymbol{v}^T\mathbf{Z}\boldsymbol{v}-\boldsymbol{v}^T(\mathbf{Z}-\widehat{\mathbf{Z}})\boldsymbol{v}\\&\ge\kappa_u^{-1}||\boldsymbol{v}||_2^2-||\boldsymbol{v}||_1^2\cdot||\mathbf{Z}-\widehat{\mathbf{Z}}||_{\max}.\end{aligned}\tag{F.1}$$

Since $\boldsymbol{v}\in\mathcal{M}_j^{\bar{c}}$, we have $||\boldsymbol{v}_{\mathcal{S}_j^\perp}||_1\le\bar{c}||\boldsymbol{v}_{\mathcal{S}_j}||_1$, which implies

$$\begin{aligned}||\boldsymbol{v}||_1&=||\boldsymbol{v}_{\mathcal{S}_j}||_1+||\boldsymbol{v}_{\mathcal{S}_j^\perp}||_1\le(1+\bar{c})||\boldsymbol{v}_{\mathcal{S}_j}||_1\\&\le(1+\bar{c})\sqrt{s}||\boldsymbol{v}_{\mathcal{S}_j}||_2,\end{aligned}\tag{F.2}$$

where the last inequality comes from the fact that there are at most s nonzero entries in $\boldsymbol{v}_{\mathcal{S}_j}$. Then combining (F.1) and (F.2), we have

$$\begin{aligned}\boldsymbol{v}^T\widehat{\mathbf{Z}}\boldsymbol{v}&\ge\kappa_u^{-1}||\boldsymbol{v}||_2^2-(1+\bar{c})^2||\boldsymbol{v}_{\mathcal{S}}||_1^2\cdot||\mathbf{Z}-\widehat{\mathbf{Z}}||_{\max}\\&\ge\kappa_u^{-1}||\boldsymbol{v}||_2^2-(1+\bar{c})^2s\lambda||\boldsymbol{v}_{\mathcal{S}}||_2^2.\end{aligned}\tag{F.3}$$

Since we have $2(1+\bar{c})^2s\lambda\kappa_u\le 1$, (F.3) implies

$$\boldsymbol{v}^T\widehat{\mathbf{Z}}\boldsymbol{v}\ge\frac{||\boldsymbol{v}||_2^2}{2\kappa_u}.$$ □

Appendix G
Proof of Theorem IV.2

Proof: Our following analysis also assumes that $\mathcal{D}_1=\left\{||\widehat{\mathbf{Z}}-\mathbf{Z}||_{\max}\le\lambda\right\}$ holds. Since $\mathcal{D}_1$ implies (E.1),

$$||\widehat{\mathbf{Z}}\mathbf{\Gamma}_{*j}-\mathbf{I}_{*j}||_\infty\le\lambda\tau_j,\quad\forall j=1,\ldots,d,$$

where $\tau_j = ||\mathbf{\Gamma}_{*j}||_1$. Then $(\mathbf{\Gamma}_{*j}, \tau_j)$ is a feasible solution to (13), which implies

$$||\widehat{\mathbf{Z}}(\widehat{\mathbf{\Gamma}}_{*j} - \mathbf{\Gamma}_{*j})||_\infty \leq ||\widehat{\mathbf{Z}}\widehat{\mathbf{\Gamma}}_{*j} - \mathbf{I}_{*j}||_\infty + ||\widehat{\mathbf{Z}}\mathbf{\Gamma}_{*j} - \mathbf{I}_{*j}||_\infty \leq \lambda\widehat{\tau}_j + \lambda\tau_j. \quad \text{(G.1)}$$

Moreover, we have

$$(1+c)||\widehat{\mathbf{\Gamma}}_{*j}||_1 \leq ||\widehat{\mathbf{\Gamma}}_{*j}||_1 + c\widehat{\tau}_j \leq ||\mathbf{\Gamma}_{*j}||_1 + c\tau_j = (1+c)||\mathbf{\Gamma}_{*j}||_1 = (1+c)\tau_j,$$

which further implies

$$||\widehat{\mathbf{\Gamma}}_{*j}||_1 \leq ||\mathbf{\Gamma}_{*j}||_1, \ ||\widehat{\mathbf{\Gamma}}_{*j} - \mathbf{\Gamma}_{*j}||_1 \leq 2||\mathbf{\Gamma}_{*j}||_1. \quad \text{(G.2)}$$

Combing (G.1) and (G.2), we have

$$\begin{aligned}&||\mathbf{Z}(\widehat{\mathbf{\Gamma}}_{*j} - \mathbf{\Gamma}_{*j})||_\infty \\ &\leq ||\widehat{\mathbf{Z}}(\widehat{\mathbf{\Gamma}}_{*j} - \mathbf{\Gamma}_{*j})||_\infty + ||(\widehat{\mathbf{Z}} - \mathbf{Z})(\widehat{\mathbf{\Gamma}}_{*j} - \mathbf{\Gamma}_{*j})||_\infty \\ &\leq \lambda\tau + \lambda\widehat{\tau}_j + ||\widehat{\mathbf{Z}} - \mathbf{Z}||_{\max}||\widehat{\mathbf{\Gamma}}_{*j} - \mathbf{\Gamma}_{*j}||_1 \\ &\leq 3\lambda\tau + \lambda\widehat{\tau}_j \leq \frac{(1+4c)\lambda\tau_j}{c}, \end{aligned} \quad \text{(G.3)}$$

where the last inequality comes from

$$c\widehat{\tau}_j \leq ||\widehat{\mathbf{\Gamma}}_{*j}||_1 + c\widehat{\tau}_j \leq ||\mathbf{\Gamma}_{*j}||_1 + c\tau_j = (1+c)\tau_j.$$

By (G.3), we have

$$||\widehat{\mathbf{\Gamma}}_{*j} - \mathbf{\Gamma}_{*j}||_\infty \leq ||\mathbf{\Gamma}_{*j}||_1||\mathbf{Z}(\widehat{\mathbf{\Gamma}}_{*j} - \mathbf{\Gamma}_{*j})||_\infty \leq \frac{\lambda(1+4c)\tau_j}{c}\cdot||\mathbf{\Gamma}_{*j}||_1 \leq \frac{\lambda(1+4c)\tau_j^2}{c}. \quad \text{(G.4)}$$

Recall $\lambda = \kappa_1\sqrt{\log d/n}$, by the definition of the max norm and (G.4), we have

$$||\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}||_{\max} \leq \kappa_6 \cdot M^2 \cdot \sqrt{\frac{\log d}{n}}, \quad \text{(G.5)}$$

where $\kappa_7 = \kappa_1(1+4c)/c$. Since for any $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{R}^{d\times d}$, where $\mathbf{A}$ and $\mathbf{C}$ are diagonal matrices, we have

$$||\mathbf{ABC}||_{\max} \leq ||\mathbf{A}||_{\max} \cdot ||\mathbf{B}||_{\max} \cdot ||\mathbf{C}||_{\max}. \quad \text{(G.6)}$$

Conditioning on

$$\mathcal{D}_2 = \left\{||\widehat{\mathbf{\Theta}}^{-1} - \mathbf{\Theta}^{-1}||_{\max} \leq \kappa_2\sqrt{\frac{\log d}{n}}\right\}, \quad \text{(G.7)}$$

(G.6), (E.18) and the fact $||\mathbf{\Gamma}||_{\max} \leq M$ imply

$$\begin{aligned}||\widetilde{\mathbf{\Omega}} - \mathbf{\Omega}||_{\max} \leq{}& \kappa_2^2\kappa_7 \cdot \frac{\log d}{n} \cdot M^2 \cdot \sqrt{\frac{\log d}{n}} \\ &+ \frac{\kappa_2\kappa_7}{\theta_{\min}} \cdot \sqrt{\frac{\log d}{n}} \cdot M^2 \cdot \sqrt{\frac{\log d}{n}} \\ &+ \kappa_2^2 \cdot M \cdot \frac{\log d}{n} + \frac{\kappa_2}{\theta_{\min}} \cdot M \cdot \sqrt{\frac{\log d}{n}} \\ &+ \frac{\kappa_2\kappa_7}{\theta_{\min}} \cdot \sqrt{\frac{\log d}{n}} \cdot M^2 \cdot \sqrt{\frac{\log d}{n}} \\ &+ \frac{\kappa_7}{\theta_{\min}^2} M^2 \cdot \sqrt{\frac{\log d}{n}} + \frac{\kappa_2}{\theta_{\min}} \cdot M \cdot \sqrt{\frac{\log d}{n}}. \end{aligned} \quad \text{(G.8)}$$

Again if (A.4): $s^2 \log d/n \to 0$ holds, then (G.8) is determined by the slowest rate $M^2\sqrt{\log d/n}$. Thus for large enough n, if we choose the max norm as $||\cdot||_*$ in (18), we have

$$||\widehat{\mathbf{\Omega}} - \mathbf{\Omega}||_{\max} \leq ||\widetilde{\mathbf{\Omega}} - \widehat{\mathbf{\Omega}}||_{\max} + ||\widetilde{\mathbf{\Omega}} - \mathbf{\Omega}||_{\max} \leq 2||\widetilde{\mathbf{\Omega}} - \mathbf{\Omega}||_{\max} \leq 2\kappa_7 \cdot M^2 \cdot \sqrt{\frac{\log d}{n}}, \quad \text{(G.9)}$$

where the second inequality comes from the fact that $\mathbf{\Omega}$ is a feasible solution to (18), and $\widehat{\mathbf{\Omega}}$ is the empirical minimizer.

Note that the results obtained here only depend on $\mathcal{D}_1$ and $\mathcal{D}_2$. Thus by Lemma 1 and (G.9), let $C_3 = 2\kappa_7$, we have

$$\mathbb{P}\left(||\widehat{\mathbf{\Omega}} - \mathbf{\Omega}||_{\max} \leq C_3 \cdot M^2 \cdot \sqrt{\frac{\log d}{n}}\right) \geq 1 - \frac{3}{d}.$$

To show the partial consistency in graph estimation $\mathbb{P}\left(E \subseteq \widehat{E}\right) \to 1$, we follow a similar argument to Theorem 4 in [25]. Therefore the proof is omitted. □

REFERENCES

[1] O. Banerjee, L. El Ghaoui, and A. d'Aspremont, "Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data," *J. Mach. Learn. Res.*, vol. 9, pp. 485–516, Jun. 2008.

[2] P. J. Bickel and E. Levina, "Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations," *Bernoulli*, vol. 10, no. 6, pp. 989–1010, 2004.

[3] D. M. Blei and J. D. Lafferty, "A correlated topic model of science," *Ann. Appl. Statist.*, vol. 1, no. 1, pp. 17–35, 2007.

[4] S. Boyd and L. Vandenberghe, *Convex Optimization*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[5] T. Cai, W. Liu, and X. Luo, "A constrained ℓ_1 minimization approach to sparse precision matrix estimation," *J. Amer. Statist. Assoc.*, vol. 106, no. 494, pp. 594–607, 2011.

[6] S. Cambanis, S. Huang, and G. Simons, "On the theory of elliptically contoured distributions," *J. Multivariate Anal.*, vol. 11, no. 3, pp. 368–385, 1981.

[7] O. Catoni, "Challenging the empirical mean and empirical variance: A deviation study," *Ann. Inst. Henri Poincaré Probab. Statist.*, vol. 48, no. 4, pp. 1148–1185, 2012.

[8] A. P. Dempster, "Covariance selection," *Biometrics*, vol. 28, no. 1, pp. 157–175, 1972.

[9] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, "Efficient projections onto the ℓ_1-ball for learning in high dimensions," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 272–279.

[10] K.-T. Fang, S. Kotz, and K. W. Ng, *Symmetric Multivariate and Related Distributions. Monographs on Statistics and Applied Probability*, vol. 36. London, U.K.: Chapman & Hall, 1990.

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

[12] E. Gautier and A. B. Tsybakov, "High-dimensional instrumental variables regression and confidence sets," ENSAE ParisTech, Malakoff, France, Tech. Rep. arxiv.org, 2011.

[13] R. P. Gorman and T. J. Sejnowski, "Analysis of hidden units in a layered network trained to classify sonar targets," *Neural Netw.*, vol. 1, no. 1, pp. 75–89, 1988.

[14] A. K. Gupta, T. Varga, and T. Bodnar, *Elliptically Contoured Models in Statistics and Portfolio Theory*. New York, NY, USA: Springer-Verlag, 2013.

[15] J. Honorio, L. Ortiz, D. Samaras, N. Paragios, and R. Goldstein, "Sparse and locally constant Gaussian graphical models," in *Advances in Neural Information Processing Systems 22*. Red Hook, NY, USA: Curran Associates, 2009.

[16] C.-J. Hsieh, I. S. Dhillon, P. K. Ravikumar, and M. A. Sustik, "Sparse inverse covariance matrix estimation using quadratic approximation," in *Advances in Neural Information Processing Systems*, vol. 24. Red Hook, NY, USA: Curran Associates, 2011, pp. 2330–2338.

[17] H. Hult and F. Lindskog, "Multivariate extremes, aggregation and dependence in elliptical distributions," *Adv. Appl. Probab.*, vol. 34, no. 3, pp. 587–608, 2002.
[18] W. H. Kruskal, "Ordinal measures of association," *J. Amer. Statist. Assoc.*, vol. 53, no. 284, pp. 814–861, 1958.
[19] W. J. Krzanowski, *Principles of Multivariate Analysis*. Oxford, U.K.: Clarendon, 2000.
[20] 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.
[21] S. L. Lauritzen, *Graphical Models*, vol. 17. London, U.K.: Oxford Univ. Press, 1996.
[22] X. Li, T. Zhao, X. Yuan, and H. Liu, "The flare Package for High-dimensional Sparse Linear Regression in R," *J. Mach. Learn. Res.*, 2014.
[23] H. Liu, F. Han, M. Yuan, J. Lafferty, and L. Wasserman, "High-dimensional semiparametric Gaussian copula graphical models," *Ann. Statist.*, vol. 40, no. 4, pp. 2293–2326, 2012.
[24] H. Liu, F. Han, and C.-H. Zhang, "Transelliptical graphical models," in *Advances in Neural Information Processing Systems 25*. Red Hook, NY, USA: Curran Associates, 2012.
[25] H. Liu and L. Wang, "TIGER: A tuning-insensitive approach for optimally estimating Gaussian graphical models," Massachusett Inst. Technol., Cambridge, MA, USA, Tech. Rep., 2012.
[26] H. Liu, L. Wang, and T. Zhao, "Sparse covariance matrix estimation with eigenvalue constraints," *J. Comput. Graph. Statist.*, vol. 23, no. 2, pp. 439–459, 2014.
[27] Y. E. Nesterov, "An approach to constructing optimal methods for minimization of smooth convex functions," *Èkonomika Matematicheskie Metody*, vol. 24, no. 3, pp. 509–517, 1988.
[28] Y. Nesterov, "Smooth minimization of non-smooth functions," *Math. Program.*, vol. 103, no. 1, pp. 127–152, 2005.
[29] A. J. Rothman, P. J. Bickel, E. Levina, and J. Zhu, "Sparse permutation invariant covariance estimation," *Electron. J. Statist.*, vol. 2, pp. 494–515, 2008.
[30] J. Shao, Y. Wang, X. Deng, and S. Wang, "Sparse linear discriminant analysis by thresholding for high dimensional data," *Ann. Statist.*, vol. 39, no. 2, pp. 1241–1265, 2011.
[31] J. Stoer, R. Bulirsch, R. Bartels, W. Gautschi, and C. Witzgall, *Introduction to Numerical Analysis*, vol. 2. New York, NY, USA: Springer-Verlag, 1993.
[32] T. Sun and C. Zhang, "Scaled sparse linear regression," *Biometrika*, vol. 99, no. 4, p. 879, 2012.
[33] T. Tokuda, B. Goodrich, I. Van Mechelen, A. Gelman, and F. Tuerlinckx, "Visualizing distributions of covariance matrices," Columbia Univ., New York, NY, USA, Tech. Rep., 2011.
[34] R. J. Vanderbei, *Linear Programming: Foundations and Extensions*. New York, NY, USA: Springer-Verlag, 2008.
[35] H. Wakaki, "Discriminant analysis under elliptical populations," *Hiroshima Math. J.*, vol. 24, no. 2, pp. 257–298, 1994.
[36] A. Wille *et al.*, "Sparse graphical Gaussian modeling of the isoprenoid gene network in arabidopsis thaliana," *Genome Biol.*, vol. 25, no. 5, p. R92, 2004.
[37] M. Yuan, "High dimensional inverse covariance matrix estimation via linear programming," *J. Mach. Learn. Res.*, vol. 11, pp. 2261–2286, Mar. 2010.
[38] M. Yuan and Y. Lin, "Model selection and estimation in the Gaussian graphical model," *Biometrika*, vol. 94, no. 1, pp. 19–35, 2007.
[39] T. Zhao and H. Liu, "Sparse inverse covariance estimation with calibration," in *Advances in Neural Information Processing Systems 26*. Red Hook, NY, USA: Curran Associates, 2013.
[40] T. Zhao, H. Liu, K. Roeder, J. Lafferty, and L. Wasserman, "The huge package for high-dimensional undirected graph estimation in R," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 1059–1062, 2012.

Tuo Zhao received his B.S. and M.S. degrees in Computer Science from Harbin Institute of Technology, and his second M.S. degree in Applied Math from University of Minnesota.

He is currently a Ph.D. Candidate in Department of Computer Science at Johns Hopkins University. He is also a visiting student in Department of Operations Research and Financial Engineering at Princeton University. His research focuses on large-scale semiparametric and nonparametric learning and applications to high throughput genomics and neuroimaging.

Han Liu received a joint Ph.D. degree in Machine Learning and Statistics from the Carnegie Mellon University, Pittsburgh, PA, USA in 2011.

He is currently an Assistant Professor of Statistical Machine Learning in the Department of Operations Research and Financial Engineering at Princeton University, Princeton, NJ. He is also an adjunct Professor in the Department of Biostatistics and Department of Computer Science at Johns Hopkins University. He built and is serving as the principal investigator of the Statistical Machine Learning (SMiLe) lab at Princeton University. His research interests include high dimensional semiparametric inference, statistical optimization, Big Data inferential analysis.