Robust Sparse Large Covariance Matrix Estimation Under Huber Loss

XXX

***Abstract*—XXX**

Keywords: Heavy-tailed, fourth moment, non-adaptive,

I. Introduction

[Covariance matrix estimation]

Modern multivariate data analysis poses a fundamental problem: the estimation of covariance matrices. This estimation finds widespread application across numerous fields, including statistics, biology, finance, signal processing, and machine learning. As an instance, numerous dimension reduction techniques rely on the prior estimation of a covariance matrix from a given set of data points. These techniques include principal component analysis [1] as well as linear and quadratic discriminant analysis [2]. A classical way to estimate covariance matrices is to use the sample covariance matrix. However, the sample covariance matrix suffers from poor finite sample performance when data is heavy-tailed as illustrated in [papers]. [Should we introduce this later, when we have explained what is "heavy-tailedness"?]

[why robustness/outliers] However, theoretical properties of large covariance estimators discussed in the literature often hinge heavily on the Gaussian or sub-Gaussian assumption. Given that data from fields including genomic studies and quantative finance usually do not follow the assumed Gaussian or sub-Gaussian shape, such an assumption is typically very restrictive in practice. It is therefore imperative to develop robust inferential procedures that are less sensitive to the distributional assumptions. Heavy-tailed distribution[1] is a viable model for data contaminated by outliers that are typically encountered in applications. [3] demonstrates the concept of tail-robustness: Due to heavy-tailedness, the probability that some observations are sampled far away from the "true" parameter of the population is nonnegligible. These outlying data points are referred as stochastic outliers. A procedure that is robust against the stochastic outliers, evidenced by its better finite-sample performance than a nonrobust method, is called a tail-robust procedure. The tail robustness is different from the classical notion of robustness that is often characterized by the breakdown point [4]. There are estimators that can be applied to more general robust scenarios, where tail-robustness and breakdown properties are simlutaneously taken into account, see [5]. Nevertheless, we will follow the approach of tail-robustness, which is a concept that combines robustness and finite-sample (nonasymptotic) error bounds.

[Robust covariance: existing methods]

The work of [6] triggered a trend of tail-robust estimators, which are featured by tight nonasymptotic deviation analysis, rather than mean squared errors. Lam (2016) [which is a manuscript] further generalizes the method in [6] to accomodate even weaker moment assumptions.

[7] combined robust estimates of the first and second moments to obtain covariance estimators. To avoid the accumulated error from estimating the first and second moments, [3] proposed a pairwise-difference-based elementwise truncation method that adopts the idea of using U-statistics in robust estimation from [8]. Another spectrumwise truncation method is also proposed by Minsker (2018), yielding deviation bounds in operator norm.

Pilot estimators:

By exploiting a bijective mapping between Pearson correlation and Kendall's tau or Spearman's rho dependence measures that hold for elliptical distributions, [9] and [10] both proposed rank-based estimation of correlation matrices, which can be combined with marginal standard deviations estimated via the method in [6] to obtain a covariance estimate as illustrated in [11]. Huber's M-Estimator (Huber1968) with a diverging robustification parameter as specified in [12] has achieved the optimal deviation bound in ℓ_∞-norm, assuming only fourth moment exists for the distribution. Another method based on the median-of-means (MOM) technique [13][14][15] avoids the tuning of robustification parameters and can be tuning-free by fixing the number of groups, but requires stronger assumptions, e.g. existence of moments of order six.

[why sparsity/high-dimensional]

In covariance estimation problems, the number of parameters to be estimated grows quadratically with the dimension of the covariance matrix. To reduce the number of parameters to be estimated, one of the most popular assumptions is sparsity. [7], [16]

In short: to deal with high-dimensionality, there are ways including the effective rank for sparsity in the spectral [papers] and the canonical definition of sparsity in coefficients. Large dimensional, thresholding covariance matrix estimator, PD covariance estimator...Another line of work follows a slightly different assumption on the so-called weak sparsity, which imposes a uniform ℓ_q norm bound on each row or column of the covariance matrix...?

[in this paper, we solve the robust XXX]

The square-loss ℓ_1-regularized sparse covariance estimator based on the sample covariance matrix (see [17], [18] and [19]) has been extensively studied for estimating sparse covariance matrices and is proved to achieve the minimax optimal statistical rate under subgaussian data. Given that the unconstrained

[1] The distribution of a random variable X is said to be heavy-tailed if the moment generating function of X is infinite for all $t > 0$, that is, $\mathbb{E}e^{tX} = \infty$ for all $t > 0$. Hence we can only assume weaker moment conditions (for instance, $\mathbb{E}X^4 < \infty$) for heavy-tailed distributions.

square-loss ℓ_1-regularized sparse covariance estimator can be obtained using the soft thresholding operator over the sample covariance matrix [20], a common method for estimating sparse covariance matrix under heavy-tailed data is to first introduce a pilot estimator[2] as a robust substitution of the sample covariance matrix, then apply thresholding to the pilot estimator as in [7](adaptive thresholding over kendall's tau, Huber's M, MoM), [21](adaptive thresholding over MoM) and [22](hard thresholding over Maronna's), or compute square-loss ℓ_1-regularized sparse covariance estimator based on the pilot estimator as in [23] (a quaitile-based pilot estimator). Note that directly applying thresholding to the sample covariance matrix ([16]) or using the square-loss ℓ_1-regularized sparse covariance estimator based on the sample covariance matrix ([17]) results in a suboptimal statistical rate for heavy-tailed data.

[contribution of this paper can be summarized as follows:]

By looking into the tail-robust sparse estimation procedures above, we can see that they proceed in two separate steps bridged by an intermediate pilot estimator. In the following discussions, we will combine the two separate steps into a single-step Huber-loss ℓ_1-regularized sparse covariance estimator that not only is a pilot estimator, but also achieves the minimax optimal rate under high-dimensional heavy-tailed data.

Meanwhile, we will point out a simple yet general approach that turns any pilot estimator into a positive-definite sparse covariance estimator, with the idea inspired by [18] and used for a specific pilot estimator in [23]. In other words, we will demonstrate that there is little gap between a pilot estimator and a positive-definite sparse covariance estimator. Therefore, if we want our proposed estimator to further obtain positive-definiteness, we can easily convert it into a positive-definite sparse covariance estimator while retaining other desirable properties.

[the structure]

XXX

Finally, we will evident via simulation that our estimator achieves the desired rate.

A. Notations

The following notation is adopted. Standard lower-case or upper-case letters stand for scalars and boldface lower-case (upper-case) letters denote vectors (matrices). Both X_{ij} and $[\mathbf{X}]_{ij}$ denote the (i,j)-th entry of the matrix $\mathbf{X}$. $\mathbb{R}_+$ denotes the set of non-negative real numbers, $\mathbb{R}^{m\times n}$ denotes the set of real $m \times n$ matrices. $\mathbf{0}$ and $\mathbf{1}$ stand for the all-zero and all-one vector/matrix, respectively. $\mathbf{I}$ stands for the identity matrix. $\mathbf{X} \succ \mathbf{0}$ ($\mathbf{X} \succeq \mathbf{0}$) means $\mathbf{X}$ is positive definite (semidefinite). $\mathbf{x} \geq \mathbf{0}$ denotes each element of $\mathbf{x}$ is non-negative.

Let $\|\mathbf{X}\|_\infty = \max_{k,l} |X_{kl}|$ and $\|\mathbf{X}\|_{\min} = \min_{k,l} |X_{kl}|$. Let $\|\mathbf{X}\|_{1,\text{off}} = \sum_{k\neq l} |X_{kl}|$ denote the sum-absolute-value norm for all entries and for off-diagonals. We write $[d]$ for the set $\{1, 2, \dots, d\}$ and $\lfloor x \rfloor$ for the largest integer not exceeding x. For an index set $\mathcal{E}$, we use $|\mathcal{E}|$ to denote its cardinality, $\overline{\mathcal{E}}$ to denote its complement. Use $\mathbf{X}_\mathcal{E}$ to denote the matrix whose (i,j)-th entry is equal to X_{ij} if $(i,j) \in \mathcal{E}$, and zero otherwise. Let $\mathbf{A} \circ \mathbf{B}$ denote the Hadamard product between matrix $\mathbf{A}$ and $\mathbf{B}$. Let $\partial f(\cdot)$ denote the subdifferential of a multivariate function f.

Let $\operatorname{sgn}(x)$ denote the sign of variable x, i.e., $\operatorname{sgn}(x) = x/|x|$. For functions $f(n)$ and $g(n)$, we denote $f(n) \lesssim g(n)$ if $f(n) \leq Cg(n)$, $f(n) \gtrsim g(n)$ if $f(n) \geq cg(n)$ and $f(n) \asymp g(n)$ if $cg(n) \leq f(n) \leq Cg(n)$ for some positive constants c and C.

II. Problem Formulation

A. Thresholding Estimator

XXX

B. Problem Formulation

Given zero-mean samples $\mathbf{x}_i$, $i = 1, \dots, n$ from a heavy-tailed distribution, define

$$L_\alpha(\mathbf{\Sigma}) := \sum_{k,\ell} \frac{1}{n} \sum_{i=1}^{n} \rho_\alpha(\Sigma_{k\ell} - x_{ik}x_{i\ell})$$

with $\rho_\alpha : \mathbb{R} \to \mathbb{R}_+$ a Huber loss function defined as

$$\rho_\alpha(x) = \begin{cases} x^2/2 & \text{if } |x| \leq \alpha, \\ \alpha|x| - \alpha^2/2 & \text{if } |x| > \alpha. \end{cases}$$

Further, define

$$\widehat{\mathbf{\Sigma}} \in \arg\min_{\mathbf{\Sigma}} \left\{ L_\alpha(\mathbf{\Sigma}) + \lambda \|\mathbf{\Sigma}\|_{1,\text{off}} \right\} \tag{1}$$

In this paper, we want to show $\widehat{\mathbf{\Sigma}}$ achieves the minimax optimal statistical rate for robust sparse covariance estimation.

Next, we introduce the pairwise difference approach:XXX The pairwise difference approach is typical in literature, with XXX

III. Theoretical Results

We denote the underlying true covariance matrix by $\mathbf{\Sigma}^*$. Let $\mathcal{S} = \{(i,j) \mid \Sigma^*_{ij} \neq 0\}$ be the support set of $\mathbf{\Sigma}^*$ and s be its cardinality, i.e., $s = |\mathcal{S}|$. In the following, we impose some mild conditions on the true covariance matrix $\mathbf{\Sigma}^*$ and the distribution of the i.i.d. samples $\mathbf{x}_i$, $i = 1, \dots, n$.

Assumption 1. *$\mathbf{x}_i \in \mathbb{R}^d$ is a heavy-tailed random variable with zero mean, i.e. $\mathbb{E}[x_{ij}] = 0$ and $\mathbb{E}\left[|x_{ij}|^4\right] \leq \sigma^2$ for all $1 \leq j \leq d$ with some positive σ.*

Remark 2. Assumption 1 immediately implies that there exists constant $K > 0$ that only depends on σ, such that $\mathbb{E}\left[(\Sigma^*_{kl} - x_{ik}x_{il})^2\right] \leq K$ for all $k, l \in [d]$. Also note that a scaling scheme of K with respect to d is explicitly assumed. In other words, K also depends on d.

A typical assumption on the heavy-tailed distribution is the fourth moment condition in assumption 1, which is adopted by robust pilot estimators, including the truncation methods and their M-Estimation counterparts from [3], the (adaptive) generalized thresholding methods from [24][7][21]. Other variants of the fourth moment condition include the

[2] Here we refer to the pilot estimator as a tail-robust estimator that achieves a certain deviation bound in ℓ_∞-norm, which follows the terminology in [7].

polynomial-tail condition in ℓ_1-regularized estimators [19][18] and the finite kurtoses condition in [3][25][8][5]. The fourth moment assumption is justified in scenarios where the data is subject to heavy-tailed and asymmetric errors. For instance, it is widely known that financial returns typically exhibit heavy tails, and [26] provides further evidence showing that a Student's t-distribution with **four degrees of freedom** displays a tail behavior similar to many asset returns. Another method based on the median-of-means (MOM) technique [13][14][15] do not need the tuning of thresholding parameters but requires stronger assumptions, namely, existence of moments of order six. (should I further explain that results obtained under <4 moment assumptions do not achieve the minimax optimal rate?)

The estimator proposed by [23] overcomes heavy-tailed high-dimensional data and achieves the minimax optimal statistical rate, but their result hinges on an elliptical-shape assumption that is only known to hold for pair-elliptically distributed random data.

Lemma 3. *Assume $\left\|\nabla L_\alpha(\widehat{\boldsymbol{\Sigma}})\right\|_\infty < \sqrt{K}\epsilon_{n,d}$ holds with $\epsilon_{n,d}$ to be a deterministic bounded sequence. Let $\alpha \asymp \sqrt{Kn/\log d}$. Take the sample size $n \gtrsim \log d$, then*

$$\left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\right\|_\infty \lesssim \sqrt{K\log d/n} + \sqrt{K}\epsilon_{n,d} \tag{2}$$

holds with high probability.

Proof: For fixed k, l, let $\widehat{\theta} := (\widehat{\boldsymbol{\Sigma}})_{kl}$ and define

$$\Psi(\theta) := \frac{1}{n}\sum_{i=1}^n \rho_\alpha'(\theta - x_{ik}x_{il}), \qquad \theta \in \mathbb{R}.$$

Note that $\left|\Psi(\widehat{\theta})\right| = \left|\left(\nabla L_\alpha(\widehat{\boldsymbol{\Sigma}})\right)_{kl}\right| < \sqrt{K}\epsilon_{n,d}$ always hold. In addition, it is easy to verify the inequality that

$$-\log(1 - x + x^2) \le \rho_1'(x) \le \log(1 + x + x^2) \tag{3}$$

By (3) and the fact that $\alpha^{-1}\rho_\alpha'(t) = \rho_1'(t/\alpha)$,

$$\begin{aligned}
&\mathbb{E}e^{(n/\alpha)\cdot\Psi(\theta)} = \prod_{i=1}^n \mathbb{E}e^{\rho_1'((\theta - x_{ik}x_{il})/\alpha)} \\
&\le \prod_{i=1}^n \mathbb{E}\left\{1 + \alpha^{-1}(\theta - x_{ik}x_{il}) + \alpha^{-2}(\theta - x_{ik}x_{il})^2\right\} \\
&\le \prod_{i=1}^n \left[1 + \alpha^{-1}(\theta - \Sigma_{kl}^*) + \alpha^{-2}\left\{(\theta - \Sigma_{kl}^*)^2 + K\right\}\right] \\
&\le \exp\left[n\alpha^{-1}(\theta - \Sigma_{kl}^*) + n\alpha^{-2}\left\{(\theta - \Sigma_{kl}^*)^2 + K\right\}\right].
\end{aligned} \tag{4}$$

Similarly, it can be shown that

$$\begin{aligned}
&\mathbb{E}e^{-(n/\alpha)\cdot\Psi(\theta)} \\
&\le \exp\left[-n\alpha^{-1}(\theta - \Sigma_{kl}^*) + n\alpha^{-2}\left\{(\theta - \Sigma_{kl}^*)^2 + K\right\}\right].
\end{aligned} \tag{5}$$

For $\eta \in (0,1)$, define

$$B_-(\theta) = (\theta - \Sigma_{kl}^*) + \left\{(\theta - \Sigma_{kl}^*)^2 + K\right\}/\alpha - (\alpha/n)\log\eta$$

$$B_+(\theta) = -(\theta - \Sigma_{kl}^*) + \left\{(\theta - \Sigma_{kl}^*)^2 + K\right\}/\alpha + (\alpha/n)\log\eta$$

Together, (4), (5) and Markov's inequality imply

$$\Pr\left(\Psi(\theta) > B_-(\theta)\right) \le e^{-nB_-(\theta)/\alpha}\cdot\mathbb{E}e^{(n/\alpha)\cdot\Psi(\theta)} \le \eta,$$

and $\Pr\left(\Psi(\theta) < B_+(\theta)\right) \le e^{-nB_+(\theta)/\alpha}\cdot\mathbb{E}e^{-(n/\alpha)\cdot\Psi(\theta)} \le \eta.$

Let θ_+ be the smallest solution of the quadratic equation $B_+(\theta_+) = \sqrt{K}\epsilon_{n,d}$, and θ_- be the largest solution of the quadratic equation $B_-(\theta_-) = -\sqrt{K}\epsilon_{n,d}$. We need to check that θ_- and θ_+ are well-defined. Let Δ_- and Δ_+ denote the discriminant of equation $B_-(\theta) = -\sqrt{K}\epsilon_{n,d}$ and $B_+(\theta) = \sqrt{K}\epsilon_{n,d}$, respectively. Since $\alpha \asymp \sqrt{Kn/\log d}$, $\epsilon_{n,d} = O(1)$ and by taking $n \gtrsim \log d$, $\eta = 1/d^3$, we have

$$\begin{aligned}
B_-(\Sigma_{kl}^* - \alpha/2) &= -\alpha/4 + K/\alpha - (\alpha/n)\log\eta < -\sqrt{K}\epsilon_{n,d} \\
B_-(\Sigma_{kl}^*) &= K/\alpha - (\alpha/n)\log\eta > -\sqrt{K}\epsilon_{n,d}
\end{aligned}$$

which implies that θ_- is well-defined as a solution to $B_-(\theta) = -\sqrt{K}\epsilon_{n,d}$ on $(\Sigma_{kl}^* - \alpha/2, \Sigma_{kl}^*)$. Similarly, θ_+ is also well-defined. Then, with at least $1 - 2\eta$ probability,

$$\Psi(\theta_+) \ge B_+(\theta_+) = \sqrt{K}\epsilon_{n,d} \quad \text{and} \quad \Psi(\theta_-) \le B_-(\theta_-) = -\sqrt{K}\epsilon_{n,d}.$$

Recall that $\left|\Psi(\widehat{\theta})\right| < \sqrt{K}\epsilon_{n,d}$ always hold, and given that $\Psi(\theta)$ is nondecreasing, $\Psi(\theta_-) < \Psi(\widehat{\theta}) < \Psi(\theta_+)$ immediately implies $\theta_- \le \widehat{\theta} \le \theta_+$.

Now we estimate θ_-. Notice that by convexity, the following holds for all $\theta \in (\Sigma_{kl}^* - \alpha/2, \Sigma_{kl}^*)$:

$$B_-(\theta) \le (1/2)\cdot(\theta - \Sigma_{kl}^*) + B_-(\Sigma_{kl}^*),$$

which immediately implies that

$$\theta_- - \Sigma_{kl}^* \ge -2\left(K/\alpha - (\alpha/n)\log\eta + \sqrt{K}\epsilon_{n,d}\right).$$

To estimate θ_+, it can be seen that assuming $B_+(\theta_+) - \sqrt{K}\epsilon_{n,d} = K/\alpha + (\alpha/n)\log\eta - \sqrt{K}\epsilon_{n,d} > 0$, we have $\theta_+ \in (\Sigma_{kl}^*, \Sigma_{kl}^* + \alpha/2)$, and similarly

$$\theta_+ - \Sigma_{kl}^* \le 2\left(K/\alpha + (\alpha/n)\log\eta - \sqrt{K}\epsilon_{n,d}\right). \tag{6}$$

Otherwise if $B_+(\theta_+) - \sqrt{K}\epsilon_{n,d} \le 0$, then $\theta_+ \le 0$. Combining this with (6), we have

$$\theta_+ - \Sigma_{kl}^* \le \max\left\{2\left(K/\alpha + (\alpha/n)\log\eta - \sqrt{K}\epsilon_{n,d}\right), 0\right\}.$$

Therefore, with $\theta_- \le \widehat{\theta} \le \theta_+$,

$$\left|\widehat{\theta} - \Sigma_{kl}^*\right| \le 2\left(K/\alpha - (\alpha/n)\log\eta + \sqrt{K}\epsilon_{n,d}\right).$$

With $\eta = 1/d^3$ and the union bound, we have that with at least $1 - 2/d$ probability, $\left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\right\|_\infty \lesssim \sqrt{K\log d/n} + \sqrt{K}\epsilon_{n,d}$. ■

Proposition 4. *Let $\widetilde{\boldsymbol{\Sigma}}$ denote any solution to (1). Then, $\widetilde{\boldsymbol{\Sigma}} \in \boldsymbol{\Sigma}^* + \mathbb{C}(l)$, where $l = 4s^{1/2}$. Further, assume $\left\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\right\|_\infty \le \alpha/2$. Conditioned on the event $\mathcal{E}_1(\alpha/2, 1/2) \cap \{\|\nabla L_\alpha(\boldsymbol{\Sigma}^*)\|_\infty \le 0.5\lambda\}$,*

$$\left\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\right\|_F \le 3\lambda s^{1/2} \quad \textit{and} \quad \left\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\right\|_1 \le 12\lambda s.$$

Proposition 4 gives the deterministic interpretation of Theorem 7. In the following propositions we will analyze the probability of the conditioned event $\mathcal{E}_1(\alpha/2, 1/2) \cap \{\|\nabla L_\alpha(\mathbf{\Sigma}^*)\|_\infty \leq 0.5\lambda\}$ mentioned in Proposition 4.

Proposition 5. *Suppose that Assumption 1 hold. Recall that K is the constant defined in Remark 2. Assume $\alpha \asymp \sqrt{Kn/\log d}$ and take $n \gtrsim \log d$. Then, for any $\kappa \in (0,1)$ and $C > 0$,*

$$\langle \nabla L_\alpha(\mathbf{\Sigma}) - \nabla L_\alpha(\mathbf{\Sigma}^*), \mathbf{\Sigma} - \mathbf{\Sigma}^* \rangle \geq \min\{\kappa, \kappa/2C\} \|\mathbf{\Sigma} - \mathbf{\Sigma}^*\|_F^2$$

holds uniformly for all $\mathbf{\Sigma} \in \mathbf{\Sigma}^* + \mathbb{B}^\infty(C\alpha)$ with high probability.

Proof: Let $D_{kl} = (1/n)\sum_{i=1}^n 1(|\Sigma_{kl}^* - x_{ik}x_{il}| \leq \alpha/2)$. By Chebyshev's inequality,

$$\mathrm{E}[D_{kl}] = \Pr(|\Sigma_{kl}^* - x_{ik}x_{il}| \leq \alpha/2) \geq 1 - 4K/\alpha^2 > (1+\kappa)/2.$$

The last inequality holds because $4K/\alpha^2 < (1-\kappa)/2$, which follows from $\alpha \asymp \sqrt{Kn/\log d}$ and by taking $n \gtrsim \log d$.

For each fixed $k, l \in [d]$, let $X_i = 1(|\Sigma_{kl}^* - x_{ik}x_{il}| \leq \alpha/2)$. With Hoeffding's inequality,

$$\begin{aligned}
&\Pr\left(\left|\sum_{i=1}^n \{X_i - \mathrm{E}[X_i]\}\right| \geq (1-\kappa)n/2\right) \\
\leq &2 \cdot \exp\left(-(1-\kappa)^2 n^2/(2n)\right) = 2 \cdot \exp\left(-(1-\kappa)^2 n/2\right)
\end{aligned}$$

and

$$\begin{aligned}
&\Pr\{D_{kl} < \kappa\} \\
\leq &\Pr\{|D_{kl} - \mathrm{E}[D_{kl}]| \geq (1-\kappa)/2\} \\
= &\Pr\left\{\left|(1/n)\sum_{i=1}^n \{X_i - \mathrm{E}[X_i]\}\right| \geq (1-\kappa)/2\right\} \\
\leq &2 \cdot \exp\left(-(1-\kappa)^2 n/2\right).
\end{aligned}$$

With union bound we have

$$\Pr\left[\min_{k,l} D_{kl} < \kappa\right] \leq 2d^2 \cdot \exp\left(-(1-\kappa)^2 n/2\right) < 2/d,$$

where the last inequality follows by taking $n \geq 6\log d/(1-\kappa)^2$. Let $\mathcal{G}_{kl} := \{i \in [n] : |\Sigma_{kl}^* - x_{ik}x_{il}| \leq \alpha/2\}$. Under the event that $\min_{k,l} D_{kl} \geq \kappa$,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \{\rho_\alpha'(\Sigma_{kl} - x_{ik}x_{il}) - \rho_\alpha'(\Sigma_{kl}^* - x_{ik}x_{il})\} \cdot (\Sigma_{kl} - \Sigma_{kl}^*) \\
\geq &\frac{1}{n}\sum_{i \in \mathcal{G}_{kl}}^n \{\rho_\alpha'(\Sigma_{kl} - x_{ik}x_{il}) - \rho_\alpha'(\Sigma_{kl}^* - x_{ik}x_{il})\} \cdot (\Sigma_{kl} - \Sigma_{kl}^*) \\
\geq &\frac{1}{n}\sum_{i \in \mathcal{G}_{kl}}^n \min\{|\Sigma_{kl} - \Sigma_{kl}^*|, \alpha/2\} \cdot |\Sigma_{kl} - \Sigma_{kl}^*| \\
\geq &\frac{1}{n}\sum_{i \in \mathcal{G}_{kl}}^n \min\{1, 1/2C\}(\Sigma_{kl} - \Sigma_{kl}^*)^2 \\
\geq &\kappa \min\{1, 1/2C\}(\Sigma_{kl} - \Sigma_{kl}^*)^2
\end{aligned}$$

The second last inequality holds since $\mathbf{\Sigma} \in \mathbf{\Sigma}^* + \mathbb{B}^\infty(C\alpha)$ implies $\alpha/2 \geq |\Sigma_{kl} - \Sigma_{kl}^*|/2C$, and the last inequality follows from $|\mathcal{G}_{kl}|/n = D_{kl}$. Therefore

$$\begin{aligned}
&\langle \nabla L_\alpha(\mathbf{\Sigma}) - \nabla L_\alpha(\mathbf{\Sigma}^*), \mathbf{\Sigma} - \mathbf{\Sigma}^* \rangle \\
= &\sum_{k,l} \frac{1}{n}\sum_{i=1}^n \{\rho_\alpha'(\Sigma_{kl} - x_{ik}x_{il}) - \rho_\alpha'(\Sigma_{kl}^* - x_{ik}x_{il})\} \cdot (\Sigma_{kl} - \Sigma_{kl}^*) \\
\geq &\kappa \cdot \min\{1, 1/2C\} \cdot \|\mathbf{\Sigma} - \mathbf{\Sigma}^*\|_{\mathrm{F}}^2
\end{aligned}$$

with at least $1 - 2/d$ probability. ■

Proposition 5 implies that for any $\kappa \in (0,1)$ and $C > 0$, with $n \gtrsim \log d$, event $\mathcal{E}_1(C, \min\{\kappa, \kappa/2C\})$ happens with high probability.

Proposition 6. *Suppose that Assumption 1 hold. Let K be the constant defined in Remark 2. Then,*

$$\|\nabla L_\alpha(\mathbf{\Sigma}^*)\|_\infty \leq \sqrt{6K\log d/n} + 6\alpha\log d/n + K/\alpha \quad (7)$$

with at least $1 - 2/d$ probability.

In Proposition 6, (7) indicates that event $\{\|\nabla L_\alpha(\mathbf{\Sigma}^*)\|_\infty \leq 0.5\lambda\}$ happens with high probability if we take $\alpha \asymp \sqrt{Kn/\log d}$ and $\lambda \asymp \sqrt{K\log d/n}$.

Theorem 7. *(minimax-optimal rate) Suppose that Assumption 1 holds. Take $\lambda \asymp \sqrt{K\log d/n}$ and let $\alpha \asymp \sqrt{Kn/\log d}$. If the sample size satisfies $n \gtrsim \log d$, then*

$$\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right\|_F \lesssim \sqrt{\frac{Ks\log d}{n}} \quad \textit{and} \quad \left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right\|_1 \lesssim s\sqrt{\frac{K\log d}{n}}$$

hold simultaneously with high probability (w.h.p.).

Proof: The proof combines Proposition 4 with Lemma 3, Proposition 5 and Proposition 6. By Proposition 6, $\|\nabla L_\alpha(\mathbf{\Sigma}^*)\|_\infty \lesssim \sqrt{K\log d/n}$ given $\alpha \asymp \sqrt{Kn/\log d}$.

By taking $\lambda \asymp \sqrt{K\log d/n}$, event $\{\|\nabla L_\alpha(\mathbf{\Sigma}^*)\|_\infty \leq 0.5\lambda\}$ happens with at least $1 - 2/d$ probability.

To invoke Proposition 4, we first notice that given $\nabla L_\alpha(\widehat{\mathbf{\Sigma}}) + \lambda\mathbf{\Xi} = \mathbf{0}$ for some $\mathbf{\Xi} \in \partial\left\|\widehat{\mathbf{\Sigma}}\right\|_{1,\text{off}}$, we must have $\left\|\nabla L_\alpha(\widehat{\mathbf{\Sigma}})\right\|_\infty < 2\lambda$ always hold. Taking the deterministic sequence in Lemma 3 to be $\epsilon_{n,d} := \lambda_{n,d}/\sqrt{K} \lesssim \sqrt{\log d/n}$, we conclude that

$$\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right\|_\infty \lesssim \sqrt{K\log d/n} + 2\lambda \asymp \sqrt{K\log d/n} \leq \alpha/2,$$

where the last inequality holds by taking $n \gtrsim \log d$.

With $n \gtrsim \log d$, Proposition 5 indicates that $\mathcal{E}_1(\alpha/2, 1/2)$ happens with high probability. With union bound, event $\mathcal{E}_1(\alpha/2, 1/2) \cap \{\|\nabla L_\alpha(\mathbf{\Sigma}^*)\|_\infty \leq 0.5\lambda\}$ holds with high probability. Under this event and by Proposition 4,

$$\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right\|_{\mathrm{F}} \leq 3\lambda s^{1/2} \quad \text{and} \quad \left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right\|_1 \leq 12\lambda s.$$

Then it suffices to recall $\lambda \asymp \sqrt{K\log d/n}$. ■

IV. Numerical Simulation

XX

V. Conclusion

XX

Appendix

Lemma 8. *For any $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ satisfying $\boldsymbol{\Sigma}_{\overline{\mathcal{S}}} = \mathbf{0}$ and $\epsilon > 0$, provided $\lambda > \|\nabla L_\alpha(\boldsymbol{\Sigma})_{\overline{\mathcal{S}}}\|_\infty$, any solution $\widetilde{\boldsymbol{\Sigma}}$ to (1) satisfies*

$$\begin{aligned}&\left\|(\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{\overline{\mathcal{S}}}\right\|_1\\ \le&(\lambda - \|\nabla L_\alpha(\boldsymbol{\Sigma})_{\overline{\mathcal{S}}}\|_\infty)^{-1}\\ &\cdot(\lambda + \|\nabla L_\alpha(\boldsymbol{\Sigma})_{\mathcal{S}}\|_\infty)\cdot\left\|(\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{\mathcal{S}}\right\|_1\end{aligned}$$

Proof: For any $\Xi \in \partial\left\|\widetilde{\boldsymbol{\Sigma}}\right\|_{1,\text{off}}$, define $\boldsymbol{U}(\Xi) = \nabla L_\alpha(\widetilde{\boldsymbol{\Sigma}}) + \lambda\Xi \in \mathbb{R}^{d\times d}$. Optimality condition of (1) implies $\inf_{\Xi}\boldsymbol{U}(\Xi) = 0$. By convexity of $L_\alpha(\boldsymbol{\Sigma})$:

$$\langle\nabla L_\alpha(\widetilde{\boldsymbol{\Sigma}}) - \nabla L_\alpha(\boldsymbol{\Sigma}), \widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\rangle \ge 0.$$

Therefore,

$$\begin{aligned}&\|\boldsymbol{U}(\Xi)\|_\infty\left\|\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\right\|_1 \ge \langle\boldsymbol{U}(\Xi),\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\rangle\\ =&\langle\nabla L_\alpha(\widetilde{\boldsymbol{\Sigma}}) - \nabla L_\alpha(\boldsymbol{\Sigma}), \widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\rangle + \langle\nabla L_\alpha(\boldsymbol{\Sigma}), \widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\rangle\\ &+\langle\lambda\Xi,\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\rangle\\ \ge&0 - \|\nabla L_\alpha(\boldsymbol{\Sigma})_{\mathcal{S}}\|_\infty\left\|(\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{\mathcal{S}}\right\|_1\\ &-\|\nabla L_\alpha(\boldsymbol{\Sigma})_{\overline{\mathcal{S}}}\|_\infty\left\|(\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{\overline{\mathcal{S}}}\right\|_1 + \langle\lambda\Xi,\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\rangle\end{aligned}$$

Moreover, we have

$$\begin{aligned}&\langle\lambda\Xi,\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\rangle\\ =&\lambda\langle\Xi_{\overline{\mathcal{S}}},(\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{\overline{\mathcal{S}}}\rangle + \lambda\langle\Xi_{\mathcal{S}},(\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{\mathcal{S}}\rangle\\ \ge&\lambda\left\|(\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{\overline{\mathcal{S}}}\right\|_1 - \lambda\left\|(\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{\mathcal{S}}\right\|_1\end{aligned}$$

Together, the last two displays imply

$$\begin{aligned}&\|\boldsymbol{U}(\Xi)\|_\infty\left\|\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\right\|_1\\ \ge&-\|\nabla L_\alpha(\boldsymbol{\Sigma})_{\mathcal{S}}\|_\infty\left\|(\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{\mathcal{S}}\right\|_1 - \|\nabla L_\alpha(\boldsymbol{\Sigma})_{\overline{\mathcal{S}}}\|_\infty\left\|(\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{\overline{\mathcal{S}}}\right\|_1\\ &+\lambda\left\|(\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{\overline{\mathcal{S}}}\right\|_1 - \lambda\left\|(\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{\mathcal{S}}\right\|_1\end{aligned}$$

Since the right-hand side of this inequality does not depend on Ξ, taking the infimum with respect to $\Xi \in \partial\left\|\widetilde{\boldsymbol{\Sigma}}\right\|_{1,\text{off}}$ on both sides to reach

$$\begin{aligned}0\ge&-\|\nabla L_\alpha(\boldsymbol{\Sigma})_{\mathcal{S}}\|_\infty\left\|(\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{\mathcal{S}}\right\|_1 - \|\nabla L_\alpha(\boldsymbol{\Sigma})_{\overline{\mathcal{S}}}\|_\infty\left\|(\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{\overline{\mathcal{S}}}\right\|_1\\ &+\lambda\left\|(\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{\overline{\mathcal{S}}}\right\|_1 - \lambda\left\|(\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{\mathcal{S}}\right\|_1\end{aligned}$$

Decompose $\left\|\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\right\|_1$ as $\left\|(\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{\mathcal{S}}\right\|_1 + \left\|(\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{\overline{\mathcal{S}}}\right\|_1$, the stated result follows immediately. ■

Lemma 9. *Conditioned on event $\{\|\nabla L_\alpha(\boldsymbol{\Sigma})\|_\infty \le 0.5\lambda\}$, any solution $\widetilde{\boldsymbol{\Sigma}}$ to (1) satisfies $\widetilde{\boldsymbol{\Sigma}} \in \boldsymbol{\Sigma} + \mathbb{C}(l)$, where $l = 4s^{1/2}$. Moreover, assume $\widetilde{\boldsymbol{\Sigma}} \in \boldsymbol{\Sigma} + \mathbb{B}^\infty(C\alpha)$. Then, conditioned on the event $\mathcal{E}_1(C\alpha,\kappa) \cap \{\|\nabla L_\alpha(\boldsymbol{\Sigma})\|_\infty \le 0.5\lambda\}$,*

$$\begin{aligned}\left\|\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\right\|_F &\le \kappa^{-1}\left(\lambda s^{1/2} + \|\nabla L_\alpha(\boldsymbol{\Sigma})_{\mathcal{S}}\|_F\right)\\ &\le 1.5\kappa^{-1}\lambda s^{1/2}.\end{aligned}$$

Proof: Conditioned on the stated event, Lemma 8 indicates

$$\left\|(\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{\overline{\mathcal{S}}}\right\|_1 \le 3\left\|(\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{\mathcal{S}}\right\|_1.$$

Therefore,

$$\left\|\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\right\|_1 \le 4s^{1/2}\left\|\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\right\|_{\text{F}},$$

which implies that $\widetilde{\boldsymbol{\Sigma}} \in \boldsymbol{\Sigma} + \mathbb{C}(l)$.

Now we prove the second statement. Since $\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} \in \mathbb{B}^\infty(C\alpha)$, conditioned on event $\mathcal{E}_1(C\alpha,\kappa)$, we have

$$\langle\nabla L_\alpha(\widetilde{\boldsymbol{\Sigma}}) - \nabla L_\alpha(\boldsymbol{\Sigma}), \widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\rangle \ge \kappa\left\|\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\right\|_{\text{F}}^2 \tag{8}$$

Now we upper bound the right-hand side of (8). For any $\Xi \in \partial\left\|\widetilde{\boldsymbol{\Sigma}}\right\|_{1,\text{off}}$, write

$$\begin{aligned}&\langle\nabla L_\alpha(\widetilde{\boldsymbol{\Sigma}}) - \nabla L_\alpha(\boldsymbol{\Sigma}), \widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\rangle\\ =&\underbrace{\langle\boldsymbol{U}(\Xi),\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\rangle}_{:=\Pi_1} - \underbrace{\langle\nabla L_\alpha(\boldsymbol{\Sigma}),\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\rangle}_{:=\Pi_2} - \underbrace{\langle\lambda\Xi,\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\rangle}_{:=\Pi_3}\end{aligned} \tag{9}$$

where $\boldsymbol{U}(\Xi) := \nabla L_\alpha(\widetilde{\boldsymbol{\Sigma}}) + \lambda\Xi \in \mathbb{R}^{d\times d}$. We have

$$\begin{aligned}|\Pi_1| &\le \|\boldsymbol{U}(\Xi)\|_\infty\left\|(\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{\overline{\mathcal{S}}}\right\|_1 + \|(\boldsymbol{U}(\Xi))_{\mathcal{S}}\|_{\text{F}}\left\|(\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{\mathcal{S}}\right\|_{\text{F}}\\ |\Pi_2| &\le \|\nabla L_\alpha(\boldsymbol{\Sigma})_{\mathcal{S}}\|_{\text{F}}\left\|(\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{\mathcal{S}}\right\|_{\text{F}}\\ &\quad+\|\nabla L_\alpha(\boldsymbol{\Sigma})_{\overline{\mathcal{S}}}\|_\infty\left\|(\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{\overline{\mathcal{S}}}\right\|_1\end{aligned}$$

Turning to Π_3, decompose $\lambda\Xi$ and $\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}$ according to $\mathcal{S}\cup\overline{\mathcal{S}}$ to reach

$$\Pi_3 = \langle(\lambda\Xi)_{\mathcal{S}},(\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{\mathcal{S}}\rangle + \langle(\lambda\Xi)_{\overline{\mathcal{S}}},(\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{\overline{\mathcal{S}}}\rangle$$

Since $\boldsymbol{\Sigma}_{\overline{\mathcal{S}}} = \mathbf{0}$ and $\Xi \in \partial\left\|\widetilde{\boldsymbol{\Sigma}}\right\|_{1,\text{off}}$, we have $\langle(\lambda\Xi)_{\overline{\mathcal{S}}},(\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{\overline{\mathcal{S}}}\rangle = \langle(\lambda\Xi)_{\overline{\mathcal{S}}},\widetilde{\boldsymbol{\Sigma}}_{\overline{\mathcal{S}}}\rangle = \lambda\left\|\widetilde{\boldsymbol{\Sigma}}_{\overline{\mathcal{S}}}\right\|_1 = \lambda\left\|(\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{\overline{\mathcal{S}}}\right\|_1$. Therefore,

$$\Pi_3 \ge \lambda\left\|(\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{\overline{\mathcal{S}}}\right\|_1 - \lambda s^{1/2}\left\|(\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{\mathcal{S}}\right\|_{\text{F}}$$

Combining (9) with our estimation for Π_1, Π_2 and Π_3, we have

$$\begin{aligned}&\langle\nabla L_\alpha(\widetilde{\boldsymbol{\Sigma}}) - \nabla L_\alpha(\boldsymbol{\Sigma}), \widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\rangle\\ \le&-\{\lambda - \|\nabla L_\alpha(\boldsymbol{\Sigma})\|_\infty - \|\boldsymbol{U}(\Xi)\|_\infty\}\left\|(\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{\overline{\mathcal{S}}}\right\|_1\\ &+\|\nabla L_\alpha(\boldsymbol{\Sigma})_{\mathcal{S}}\|_{\text{F}}\left\|(\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{\mathcal{S}}\right\|_{\text{F}} + \|(\boldsymbol{U}(\Xi))_{\mathcal{S}}\|_{\text{F}}\left\|(\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{\mathcal{S}}\right\|_{\text{F}}\\ &+\lambda s^{1/2}\left\|(\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{\mathcal{S}}\right\|_{\text{F}}\end{aligned}$$

Taking the infimum with respect to $\Xi \in \partial\left\|\widetilde{\boldsymbol{\Sigma}}\right\|_{1,\text{off}}$ on both sides, it follows that

$$\begin{aligned}&\langle\nabla L_\alpha(\widetilde{\boldsymbol{\Sigma}}) - \nabla L_\alpha(\boldsymbol{\Sigma}), \widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\rangle\\ \le&-\{\lambda - \|\nabla L_\alpha(\boldsymbol{\Sigma})\|_\infty\}\left\|(\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{\overline{\mathcal{S}}}\right\|_1\\ &+\|\nabla L_\alpha(\boldsymbol{\Sigma})_{\mathcal{S}}\|_{\text{F}}\left\|(\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{\mathcal{S}}\right\|_{\text{F}}\\ &+\lambda s^{1/2}\left\|(\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{\mathcal{S}}\right\|_{\text{F}}\end{aligned} \tag{10}$$

It follows from $\widetilde{\boldsymbol{\Sigma}} \in \boldsymbol{\Sigma} + \mathbb{B}^\infty(C\alpha)$, (8) and (10) that conditioned on $\mathcal{E}_1(C\alpha,\kappa) \cap \{\|\nabla L_\alpha(\boldsymbol{\Sigma})\|_\infty \le 0.5\lambda\}$,

$$\begin{aligned}&\kappa\left\|\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\right\|_{\text{F}}^2 \le\\ &\left\{\lambda s^{1/2} + \|\nabla L_\alpha(\boldsymbol{\Sigma})_{\mathcal{S}}\|_{\text{F}}\right\}\left\|\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\right\|_{\text{F}}\end{aligned}$$

Therefore,

$$\begin{aligned}&\left\|\widetilde{\mathbf{\Sigma}}-\mathbf{\Sigma}\right\|_{\mathrm{F}}\\ &\le \kappa^{-1}\left\{\lambda s^{1/2}+\|\nabla L_\alpha(\mathbf{\Sigma})_{\mathcal{S}}\|_{\mathrm{F}}\right\}\\ &\le \kappa^{-1}\{\lambda s^{1/2}+0.5\lambda s^{1/2}\}=1.5\kappa^{-1}\lambda s^{1/2}\end{aligned} \tag{11}$$

■

A. Proof of Proposition 4

Proof: $\left\|\widetilde{\mathbf{\Sigma}}-\mathbf{\Sigma}^*\right\|_{\mathrm{F}} \le 3\lambda s^{1/2}$ follows immediately from Lemma 9 with $\mathbf{\Sigma}=\mathbf{\Sigma}^*$ and $C=\kappa=1/2$. Combining this with $\widetilde{\mathbf{\Sigma}}\in\mathbf{\Sigma}^*+\mathbb{C}(l)$, where $l=4s^{1/2}$, yields $\left\|\widetilde{\mathbf{\Sigma}}-\mathbf{\Sigma}^*\right\|_1 \le 12\lambda s$. ■

B. Proof of Proposition 5

We adopt the following notations for the next stage of proof. Recall that $L_\alpha(\mathbf{\Sigma})=\sum_{k,\ell}\frac{1}{n}\sum_{i=1}^n\rho_\alpha(\Sigma_{k\ell}-x_{ik}x_{i\ell})$. Define $\mathbf{B}^*:=\mathbb{E}[\nabla L_\alpha(\mathbf{\Sigma}^*)]$, and $\mathbf{W}^*:=\nabla L_\alpha(\mathbf{\Sigma}^*)-\mathbb{E}[\nabla L_\alpha(\mathbf{\Sigma}^*)]$.

Lemma 10. *Recall that K is the constant defined in Remark 2. We have $|(\mathbf{B}^*)_{kl}|=|\mathrm{E}[\rho'_\alpha(\epsilon_{kl})]|<\frac{K}{\alpha}$ for all $k,l\in[d]$.*

Proof: For fixed $k,l\in[d]$, let $\epsilon_{kl}:=\Sigma^*_{k\ell}-x_{ik}x_{i\ell}$, then

$$\begin{aligned}|\mathrm{E}[\rho'_\alpha(\epsilon_{kl})]| &= |\mathrm{E}[\epsilon_{kl}I(|\epsilon_{kl}|\le\alpha)+\alpha\mathrm{sgn}(\epsilon_{kl})I(|\epsilon_{kl}|>\alpha)]|\\ &=|\mathrm{E}[\epsilon_{kl}+(\alpha\mathrm{sgn}(\epsilon_{kl})-\epsilon_{kl})I(|\epsilon_{kl}|>\alpha)]|\\ &=|\mathrm{E}\{[\epsilon_{kl}-\alpha\mathrm{sgn}(\epsilon_{kl})]I(|\epsilon_{kl}|>\alpha)\}|\\ &\le|\mathrm{E}[(|\epsilon_{kl}|-\alpha\mathrm{sgn}(\epsilon_{kl}))I(|\epsilon_{kl}|>\alpha)]|\\ &\le\frac{\left|\mathrm{E}[(\epsilon_{kl}^2-\alpha^2)I(|\epsilon_{kl}|>\alpha)]\right|}{\alpha}\\ &<\frac{K}{\alpha}.\end{aligned}$$

Therefore, for all k,l

$$|(\mathbf{B}^*)_{kl}|=\frac{1}{n}\left|\sum_{i=1}^n\mathrm{E}[\rho'_\alpha(\Sigma^*_{k\ell}-x_{ik}x_{i\ell})]\right|<\frac{K}{\alpha}.$$

■

C. Proof of Proposition 6

Proof: For each $k,l\in[d]$, recall:

$$W^*_{kl}=\frac{1}{n}\sum_{i=1}^n\left\{\rho'_\alpha(\Sigma^*_{k\ell}-x_{ik}x_{i\ell})-\mathrm{E}\left[\rho'_\alpha(\Sigma^*_{k\ell}-x_{ik}x_{i\ell})\right]\right\}.$$

Given that $|\rho'_\alpha(\Sigma^*_{k\ell}-x_{ik}x_{i\ell})|\le\alpha$, for all $m\ge2$:

$$\begin{aligned}&\mathrm{E}[\rho'_\alpha(\Sigma^*_{k\ell}-x_{ik}x_{i\ell})]^m\\ &\le\alpha^{m-2}\cdot\mathrm{Var}[\rho'_\alpha(\Sigma^*_{k\ell}-x_{ik}x_{i\ell})]\\ &\le\alpha^{m-2}\cdot\mathrm{Var}[\Sigma^*_{k\ell}-x_{ik}x_{i\ell}]\\ &\le\alpha^{m-2}K\le\alpha^{m-2}K\cdot m!/2\end{aligned}$$

The second inequality follows given $\rho'_\alpha(\cdot)$ is 1-Lipschitz. With Bernstein's inequality, for any $t\ge0$,

$$\begin{aligned}&\Pr\left(\left|\sum_{i=1}^n\{\rho'_\alpha(\Sigma^*_{k\ell}-x_{ik}x_{i\ell})-\mathrm{E}[\rho'_\alpha(\Sigma^*_{k\ell}-x_{ik}x_{i\ell})]\}\right|\right.\\ &\qquad\qquad\left.\ge\sqrt{2Knt}+2\alpha t\right)\\ &\le2\cdot\exp\left(-\frac{(\sqrt{2Knt}+2\alpha t)^2/2}{Kn+\alpha\cdot\sqrt{2Knt}+2\alpha^2t}\right)\\ &=2\cdot\exp\left(-\frac{Kn+2\alpha\cdot\sqrt{2Knt}+2\alpha^2t}{Kn+\alpha\cdot\sqrt{2Knt}+2\alpha^2t}\cdot t\right)<e^{-t}.\end{aligned}$$

Taking $t=3\log d$ and in conjunction with the union bound,

$$\Pr\left(\|\mathbf{W}^*\|_\infty\ge\sqrt{6K\log d/n}+6\alpha\log d/n\right)<d^{-1}.$$

Recall that $\nabla L_\alpha(\mathbf{\Sigma}^*)=\mathbf{B}^*+\mathbf{W}^*$. With Lemma 10, we have $\|\mathbf{B}^*\|_\infty<K/\alpha$. Combing the two parts together and with the union bound, we have

$$\|\nabla L_\alpha(\mathbf{\Sigma}^*)\|_\infty<\sqrt{6K\log d/n}+6\alpha\log d/n+K/\alpha$$

with at least $1-d^{-1}$ probability. ■

References

[1] I. T. Jolliffe, *Principal Component Analysis*. Springer, 2002.

[2] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

[3] Y. Ke, S. Minsker, Z. Ren, Q. Sun, and W.-X. Zhou, "User-friendly covariance estimation for heavy-tailed distributions," *Statistical Science*, vol. 34, no. 3, pp. 454–471, 2019.

[4] F. R. Hampel, "A General Qualitative Definition of Robustness," *The Annals of Mathematical Statistics*, vol. 42, no. 6, pp. 1887 – 1896, 1971. [Online]. Available: https://doi.org/10.1214/aoms/1177693054

[5] S. Minsker and L. Wang, "Robust estimation of covariance matrices: Adversarial contamination and beyond," 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:247292226

[6] O. Catoni, "Challenging the empirical mean and empirical variance: A deviation study," *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, vol. 48, no. 4, pp. 1148 – 1185, 2012. [Online]. Available: https://doi.org/10.1214/11-AIHP454

[7] M. Avella-Medina, H. S. Battey, J. Fan, and Q. Li, "Robust estimation of high-dimensional covariance and precision matrices." *Biometrika*, vol. 105 2, pp. 271–284, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:49237226

[8] S. Minsker and X. Wei, "Robust modifications of U-statistics and applications to covariance estimation problems," *Bernoulli*, vol. 26, no. 1, pp. 694 – 727, 2020. [Online]. Available: https://doi.org/10.3150/19-BEJ1149

[9] H. Liu, F. Han, M. Yuan, J. Lafferty, and L. Wasserman, "High-dimensional semiparametric Gaussian copula graphical models," *The Annals of Statistics*, vol. 40, no. 4, pp. 2293 – 2326, 2012. [Online]. Available: https://doi.org/10.1214/12-AOS1037

[10] L. Xue and H. Zou, "Regularized rank-based estimation of high-dimensional nonparanormal graphical models," *The Annals of Statistics*, vol. 40, no. 5, pp. 2541 – 2571, 2012. [Online]. Available: https://doi.org/10.1214/12-AOS1041

[11] J. Fan, Y. Liao, and H. Liu, "An overview of the estimation of large covariance and precision matrices," *Econom. J.*, vol. 19, no. 1, pp. C1–C32, 2016.

[12] J. Fan, Q. Li, and Y. Wang, "Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 79, no. 1, pp. 247–265, 2017. [Online]. Available: http://www.jstor.org/stable/44681770

[13] C. Blair, "Problem complexity and method efficiency in optimization (a. s. nemirovsky and d. b. yudin)," *SIAM Review*, vol. 27, no. 2, pp. 264–265, 1985. [Online]. Available: https://doi.org/10.1137/1027074

[14] L. Devroye, M. Lerasle, G. Lugosi, and R. I. Oliveira, "Sub-gaussian mean estimators," *arXiv: Statistics Theory*, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:26805883

[15] S. Minsker and N. Strawn, "Distributed statistical estimation and rates of convergence in normal approximation," *ArXiv*, vol. abs/1704.02658, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:13393096

[16] P. J. Bickel and E. Levina, "Covariance regularization by thresholding," *Ann. Statist.*, vol. 36, no. 6, pp. 2577–2604, 2008.

[17] A. J. Rothman, "Positive definite estimators of large covariance matrices," *Biometrika*, vol. 99, no. 3, pp. 733–740, 2012.

[18] L. Xue, S. Ma, and H. Zou, "Positive-definite ℓ_1-penalized estimation of large covariance matrices," *J. Am. Stat. Assoc.*, vol. 107, no. 500, pp. 1480–1491, 2012.

[19] Y. Cui, C. Leng, and D. Sun, "Sparse estimation of high-dimensional correlation matrices," *Comput. Stat. Data Anal.*, vol. 93, pp. 390–403, 2016.

[20] Q. Wei and Z. Zhao, "Large covariance matrix estimation with oracle statistical rate," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, 2023.

[21] Y. He, P. Liu, X. Zhang, and W. Zhou, "Robust covariance estimation for high-dimensional compositional data with application to microbial communities analysis," *Statistics in Medicine*, vol. 40, no. 15, pp. 3499–3515, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8979

[22] E. Romanov, G. Kur, and B. Nadler, "Tyler's and maronna's m-estimators: Non-asymptotic concentration results," *Journal of Multivariate Analysis*, vol. 196, p. 105184, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0047259X23000301

[23] J. Lu, F. Han, and H. Liu, "Robust scatter matrix estimation for high dimensional distributions with heavy tail," *IEEE Transactions on Information Theory*, vol. 67, no. 8, pp. 5283–5304, 2021.

[24] A. J. Rothman, E. Levina, and J. Zhu, "Generalized thresholding of large covariance matrices," *J. Am. Stat. Assoc.*, vol. 104, no. 485, pp. 177–186, 2009.

[25] S. Mendelson and N. Zhivotovskiy, "Robust covariance estimation under $L_4 - L_2$ norm equivalence," *The Annals of Statistics*, vol. 48, no. 3, pp. 1648 – 1664, 2020. [Online]. Available: https://doi.org/10.1214/19-AOS1862

[26] R. Cont, "Empirical properties of asset returns: stylized facts and statistical issues," *Quantitative Finance*, vol. 1, no. 2, pp. 223–236, 2001. [Online]. Available: https://doi.org/10.1080/713665670