Statistica Sinica

Robust Estimation of Covariance Matrices: Adversarial Contamination and Beyond

Stanislav Minsker and Lang Wang

University of Southern California

Abstract: We consider the problem of estimating the covariance structure of a random vector $Y \in \mathbb{R}^d$ from a sample $Y_1, \ldots, Y_n$. We are interested in the situation when d is large compared to n but the covariance matrix Σ of interest has (exactly or approximately) low rank. We assume that the given sample is (a) ε-adversarially corrupted, meaning that ε fraction of the observations could have been replaced by arbitrary vectors, or that (b) the sample is i.i.d. but the underlying distribution is heavy-tailed, meaning that the norm of Y possesses only 4 finite moments. We propose an estimator that is adaptive to the potential low-rank structure of the covariance matrix as well as to the proportion of contaminated data, and admits tight deviation guarantees despite rather weak assumptions on the underlying distribution. Finally, we discuss the algorithms that allow to approximate the proposed estimator in a numerically efficient way.

Key words and phrases: Adversarial contamination, covariance estimation, heavy-tailed distribution, low-rank recovery, U-statistics.

1. Introduction

In this paper, we consider the problem of estimating covariance matrices from ε-*corrupted samples*: we are given n independent observations from some unknown distribution $\mathcal{D}$ over $\mathbb{R}^d$ with mean μ and covariance matrix Σ, where an ε-fraction of them are adversarially corrupted. More specifically, we assume that our observations $Y_1, \ldots, Y_n$ satisfy the equation

$$Y_j = Z_j + V_j, j = 1, \ldots, n,$$

where Z_j's come from the target distribution $\mathcal{D}$ and V_j's are arbitrary (possibly random) vectors such that only an ε-fraction of them are non-zeros. Our goal is to construct a robust estimator for the covariance matrix Σ in this framework.

As attested by some early references such as the works Tukey (1960); Huber (1992), robust estimation has a long history. During the past two decades, increasing amount of practical applications created a high demand for the tools to recover high-dimensional parameters of interest from grossly corrupted measurements. Robust covariance estimators in particular have been studied extensively, for instance, see the papers by Huber (1992, 2011); Maronna et al. (2019). Although some of the proposed estimators admit theoretically optimal error bounds, they are hard to compute in general when the dimension is high because the running time is exponential in the dimension (Bernholt (2006)).

Recent work by Lai et al. (2016); Diakonikolas et al. (2019) introduced the first robust estimators for the covariance matrix Σ that are computationally efficient in the high-dimensional case, i.e. the running time is only polynomial in the dimension, assuming that the distribution $\mathcal{D}$ is Gaussian or an affine transformation of a product distribution with a bounded 4^{th} moment. Since the publication of these initial papers, a growing body of subsequent works has appeared. For instance, Cheng et al. (2019) developed fast algorithms that nearly match the best-known running time to compute the empirical covariance matrix, assuming that the distribution of Z is Gaussian with zero mean. Chen et al. (2018) developed efficient algorithms under significantly weaker conditions on the unknown distribution $\mathcal{D}$, i.e. $\mathcal{D}$ does not have to be an affine transformation of a product distribution. However, these algorithms can only achieve a theoretically suboptimal error bound in the Frobenius norm.

The present paper continues this line of research. We design a double-penalized estimator for the covariance matrix Σ, which will be shown to admit theoretically optimal error bounds when the "effective rank" of Σ (to be defined later) is small, and can be efficiently calculated using traditional numerical methods.

The rest of the paper is organized as follows. Section 2 explains the main notations and background material. Section 3 introduces the main results. Section 4 displays applications of the main results to heavy-tailed data. Section 5 presents numerical experiments. Finally, the proofs of the main results are contained in the supplementary material.

2. Preliminaries

In this section, we introduce the main notations and recall some useful facts that we rely on in the subsequent exposition. Given two real numbers $a, b \in \mathbb{R}$, we define $a \vee b := \max\{a, b\}$, $a \wedge b := \min\{a, b\}$. Also, given $x \in \mathbb{R}$, we will denote $\lfloor x \rfloor := \max\{n \in \mathbb{Z} : n \leq x\}$ to be the largest integer less than or equal to x. We will separately introduce important results of matrix algebra and sub-Gaussian distributions in the following two subsections.

2.1 Matrix algebra

Assume that $A \in \mathbb{R}^{d_1 \times d_2}$ is a $d_1 \times d_2$ matrix with real-valued entries. Let A^T denote the transpose of A. A square matrix $A \in \mathbb{R}^{d \times d}$ is called an orthogonal matrix if $AA^T = A^T A = I_d$, where I_d is the identity matrix in $\mathbb{R}^{d \times d}$.

Given a square matrix $A \in \mathbb{R}^{d\times d}$, we define the trace of A to be the sum of elements on the main diagonal, namely,

$$\operatorname{tr}(A) := \sum_{i=1}^{d} A_{i,i},$$

where $A_{i,i}$ represents the element on the i^{th} row and i^{th} column of A. We introduce three types of matrix norms and the Frobenius (or Hilbert-Schmidt) inner product as follows:

Definition 1 (Matrix norms)**.** Given $A \in \mathbb{R}^{d_1\times d_2}$, we define the following three types of matrix norms.

1. Operator norm:

$$\|A\| := \sigma_1(A) = \sqrt{\lambda_{max}(A^T A)},$$

   where $\lambda_{max}(A^T A)$ stands for the largest eigenvalue of $A^T A$.

2. Frobenius norm:

$$\|A\|_F := \sqrt{\sum_{i=1}^{\operatorname{rank}(A)} \sigma_i^2(A)} = \sqrt{\operatorname{tr}(A^T A)}.$$

3. Nuclear norm:

$$\|A\|_1 := \sum_{i=1}^{\operatorname{rank}(A)} \sigma_i(A) = \operatorname{tr}(\sqrt{A^T A}),$$

   where $\sqrt{A^T A}$ is a nonnegative definite matrix such that $(\sqrt{A^T A})^2 = A^T A$.

Definition 2. Given $A, B \in \mathbb{R}^{d_1 \times d_2}$, we define the Frobenius inner product as

$$\langle A, B \rangle := \langle A, B \rangle_F = \mathrm{tr}\,(A^T B) = \mathrm{tr}\,(AB^T).$$

It is clear that $\|A\|_F^2 = \langle A, A \rangle$.

We will now introduce matrix functions. Denote

$$S^d(\mathbb{R}) := \left\{ A \in \mathbb{R}^{d \times d} : A^T = A \right\}$$

to be the set of all symmetric matrices. The eigenvalues of A will be denoted $\lambda_1, \ldots, \lambda_d$, all of which are real numbers. Next, we define functions acting on $S^d(\mathbb{R})$ as follows:

Definition 3. Given a real-valued function f defined on an interval $\mathbb{T} \subseteq \mathbb{R}$ and a real symmetric matrix $A \in S^d(\mathbb{R})$ with the spectral decomposition $A = U \Lambda U^T$ such that $\lambda_j(A) \in \mathbb{T},\ j = 1, \ldots, d$, define $f(A)$ as $f(A) = U f(\Lambda) U^T$, where

$$f(\Lambda) = f\left(\begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{pmatrix} \right) = \begin{pmatrix} f(\lambda_1) & & \\ & \ddots & \\ & & f(\lambda_d) \end{pmatrix}.$$

Finally, the effective rank of a matrix $A \in S^d(\mathbb{R}) \setminus \{0\}$ is defined as

$$\mathrm{rk}(A) := \frac{\mathrm{tr}\,(A)}{\|A\|}.$$

Note that $1 \leq \mathrm{rk}(A) \leq \mathrm{rank}(A)$ is always true, and it is possible that $\mathrm{rk}(A) \ll \mathrm{rank}(A)$ for approximately low-rank matrices A. For instance, consider $A \in S^d(\mathbb{R})$ with eigenvalues $\lambda_1 = 1, \lambda_2 = \cdots = \lambda_d = 1/d$, whence we have $\mathrm{rk}(A) = 2 - 1/d \ll d = \mathrm{rank}(A)$.

2.2 Sub-Gaussian distributions

Given a random variable X on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, and a convex nondecreasing function $\psi : \mathbb{R}_+ \mapsto \mathbb{R}_+$ with $\psi(0) = 0$, we define the ψ-norm of X as (Vershynin (2018, Section 2.7.1))

$$\|X\|_\psi := \inf \left\{ C > 0 : \mathbb{E}\left[\psi\left(\frac{|X|}{C} \right) \right] \leq 1 \right\}.$$

In particular, in what follows we will mainly consider $\psi_1(u) := \exp\{u\} - 1, u \geq 0$ and $\psi_2(u) := \exp\{u^2\} - 1, u \geq 0$, which correspond to the sub-exponential norm and sub-Gaussian norm respectively. We will say that a random variable X is sub-Gaussian (or sub-exponential) if $\|X\|_{\psi_2} < \infty$ (or $\|X\|_{\psi_1} < \infty$). Also, let $\|X\|_{L_2} := (\mathbb{E}[|X|^2])^{1/2}$ be the L_2 norm of a random variable X. The sub-Gaussian (or sub-exponential) random vector is defined as follows:

Definition 4. A random vector Z in $\mathbb{R}^d$ with mean $\mu = \mathbb{E}[Z]$ is called L-sub-Gaussian if for every $v \in \mathbb{R}^d$, there exists an absolute constant $L > 0$

such that

$$\|\langle Z-\mu, v\rangle\|_{\psi_2} \leq L \|\langle Z-\mu, v\rangle\|_{L_2}. \tag{2.1}$$

Moreover, Z is called L-sub-exponential if ψ_2-norm in (2.1) is replaced by ψ_1-norm.

It is clear that if Z is L-sub-Gaussian, then $(-Z)$ is also L-sub-Gaussian. We introduce some important results for sub-Gaussian distributions.

Proposition 1. *(Vershynin (2018, pp.24)) A mean zero random variable Z is L-sub-Gaussian if and only if there exists an absolute constant $K(L) > 0$ depending only on L such that*

$$P(|Z| \geq t) \leq 2\exp\{-t^2/K(L)^2\}, \quad \text{for all } t \geq 0.$$

Theorem 1. *(Vershynin (2018, Theorem 2.6.3)) Let $Z_1, \ldots, Z_n$ be i.i.d L-sub-Gaussian random variables with mean zero, and $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$. Then for any $t \geq 0$, there exists a constant $K(L) > 0$ depending only on L such that*

$$P\left(\left|\sum_{i=1}^n a_i Z_i\right| \geq t\right) \leq 2\exp\left\{-\frac{t^2}{K(L)^2 \|a\|_2^2}\right\},$$

where $\|a\|_2^2 = a_1^2 + \cdots + a_n^2$.

Corollary 1. *Let $Z_1, \ldots, Z_n$ be i.i.d L-sub-Gaussian random variables with common mean $\mathbb{E}[Z_1] = \mu$ and sub-Gaussian norm $\|Z_1\|_{\psi_2} = K$. Let $a = (a_1, \ldots, a_n)$ be a vector in $\mathbb{R}^d$ such that $\|a\|_2 \leq 1$. Then*

1. *$Y := \sum_{i=1}^n a_i(Z_i - \mu)$ is still L-sub-Gaussian.*

2. *There exists an absolute constant $c > 0$, such that $\|Y\|_{\psi_2} \le cK$.*

Proof. This corollary immediately follows by a combination of Theorem 1 and Proposition 1. □

3. Problem formulation and main results

Let $Z_1, \ldots, Z_n \in \mathbb{R}^d$ be i.i.d. copies of an L-sub-Gaussian random vector Z such that $\mathbb{E}Z = \mu$ and $\mathbb{E}(Z-\mu)(Z-\mu)^T = \Sigma$. Assume that we observe a sequence

$$Y_j = Z_j + V_j,\ j = 1, \ldots, n, \tag{3.2}$$

where V_j's are arbitrary (possibly random) vectors such that only a small portion of them are not equal to zero. Namely, we assume that there exists a set of indices $J \subseteq \{1, \ldots, n\}$ such that $|J| \ll n$ and $V_j = 0$ for $j \notin J$. In what follows, the sample points with $j \in J$ will be called *outliers* and $\varepsilon := |J|/n$ will denote the proportion of such points. In this case,

$$Y_j Y_j^T = Z_j Z_j^T + \underbrace{V_j V_j^T + V_j Z_j^T + Z_j V_j^T}_{:=\sqrt{n}U_j^*} := X_j + \sqrt{n} U_j^*,$$

where $\operatorname{rank}(U_j^*) \le 2$ and the $\sqrt{n}$ factor is added for technical convenience. Our main goal is to construct an estimator for the covariance matrix Σ in the presence of outliers V_j. In practice, we usually do not know the true mean μ of Z. To address this problem, we first recall the definition of U-statistics.

Definition 5 (Hoeffding (1992)). Let $Y_1, \ldots, Y_n$ $(n \geq 2)$ be a sequence of random variables taking values in a measurable space $(\mathcal{S}, \mathcal{B})$. Assume that $H : \mathcal{S}^m \mapsto \mathbb{S}^d(\mathbb{R})$ $(2 \leq m \leq n)$ is an $\mathcal{S}^m$-measurable permutation-symmetric kernel, i.e. $H(y_1, \ldots, y_m) = H(y_{\pi_1}, \ldots, y_{\pi_m})$ for any $(y_1, \ldots, y_m) \in \mathcal{S}^m$ and any permutation π. The U-statistic with kernel H is defined as

$$U_n := \frac{(n-m)!}{n!} \sum_{(i_1,\ldots,i_m) \in I_n^m} H(Y_{i_1}, \ldots, Y_{i_m}),$$

where $I_n^m := \{(i_1, \ldots, i_m) : 1 \leq i_j \leq n, i_j \neq i_k \text{ if } j \neq k\}$.

A particular example of U-statistics is the sample covariance matrix defined as

$$\widetilde{\Sigma}_s := \frac{1}{n-1} \sum_{j=1}^{n} (Y_j - \bar{Y})(Y_j - \bar{Y})^T, \tag{3.3}$$

where $\bar{Y} := \frac{1}{n} \sum_{j=1}^{n} Y_j$. Indeed, it is easy to verify that

$$\widetilde{\Sigma}_s = \frac{1}{n(n-1)} \sum_{(i,j) \in I_n^2} \frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2}, \tag{3.4}$$

hence the sample covariance matrix is a U-statistic with kernel

$$H(x, y) := \frac{(x-y)(x-y)^T}{2} \text{ for any } x, y \in \mathbb{R}^d. \tag{3.5}$$

Note that $\mathbb{E}\left[\frac{Y_i - Y_j}{\sqrt{2}}\right] = 0$ and $\mathbb{E}\left[\frac{(Y_i-Y_j)}{\sqrt{2}}\frac{(Y_i-Y_j)}{\sqrt{2}}^T\right] = \Sigma$ for all $(i,j) \in I_n^2$. Namely, by expressing the sample covariance matrix as a U-statistic in (3.4), the explicit estimation of the unknown mean μ can be avoided. Therefore, we consider the following settings:

$$\widetilde{Y}_{i,j} := \frac{Y_i - Y_j}{\sqrt{2}}, \quad \widetilde{Z}_{i,j} := \frac{Z_i - Z_j}{\sqrt{2}}, \quad \widetilde{V}_{i,j} := \frac{V_i - V_j}{\sqrt{2}},$$
$$\forall (i,j) \in I_n^2. \quad (3.6)$$

Then

$$\widetilde{Y}_{i,j}\widetilde{Y}_{i,j}^T = \widetilde{Z}_{i,j}\widetilde{Z}_{i,j}^T + \underbrace{\widetilde{V}_{i,j}\widetilde{V}_{i,j}^T + \widetilde{V}_{i,j}\widetilde{Z}_{i,j}^T + \widetilde{Z}_{i,j}\widetilde{V}_{i,j}^T}_{:=\sqrt{n(n-1)}\widetilde{U}_{i,j}^*}$$
$$:= \widetilde{X}_{i,j} + \sqrt{n(n-1)}\widetilde{U}_{i,j}^*, \quad (3.7)$$

where the $n(n-1) = |I_n^2|$ factor equals the total number of $\widetilde{Y}_{i,j}$'s, and is added for technical convenience. It is easy to check that the following claims hold:

1. $\widetilde{Y}_{i,j} = \widetilde{Z}_{i,j} + \widetilde{V}_{i,j}$, with $\mathbb{E}\left[\widetilde{Z}_{i,j}\right] = 0$ and $\mathbb{E}\left[\widetilde{Z}_{i,j}\widetilde{Z}_{i,j}^T\right] = \Sigma$, for any $(i,j) \in I_n^2$. Moreover, Corollary 1 shows that $\widetilde{Z}_{i,j}, (i,j) \in I_n^2$ has sub-Gaussian distribution.

2. $\widetilde{Z}_{i,j}$'s are identically distributed, but not independent.

3. Denote $\widetilde{J} = \left\{(i,j) \in I_n^2 : \widetilde{V}_{i,j} \neq 0\right\}$ to be the set of indices such that $\widetilde{V}_{i,j} = 0$, $\forall (i,j) \notin \widetilde{J}$. Then $|\widetilde{J}|$ represents the number of outliers in $\left\{\widetilde{Y}_{i,j} : (i,j) \in I_n^2\right\}$, and we have that

$$|\widetilde{J}| = 2|J|(n - |J|) + |J|(|J| - 1) = |J|(2n - |J| - 1). \tag{3.8}$$

4. $\mathrm{Rank}(\widetilde{U}^*_{i,j}) \leq 2$. This follows from the fact that for any vector $v \in \mathbb{R}^d$, $\widetilde{U}^*_{i,j} v \in \mathrm{span}\left\{\widetilde{V}_{i,j}, \widetilde{Z}_{i,j}\right\}$.

In what follows, we will use the notation

$$\mathbf{U}_{\mathbf{I_n^2}} := (U_{1,2}, \ldots, U_{n,n-1})$$

to represent the $n(n-1)$-dimensional sequence with subscripts taking from I_n^2. Similarly, the notation $(S, \mathbf{U}_{\mathbf{I_n^2}})$ will represent the (n^2-n+1)-dimensional sequence $(S, U_{1,2}, \ldots, U_{n,n-1})$.

Now we are ready to define our estimator. Given $\lambda_1, \lambda_2 > 0$, set

$$\begin{aligned}(\widehat{S}_\lambda, \widehat{\mathbf{U}}_{\mathbf{I_n^2}}) = \underset{S, U_{1,2}, \ldots, U_{n,n-1}}{\operatorname{argmin}} \Bigg[& \lambda_1 \left\|S\right\|_1 + \lambda_2 \sum_{i \neq j} \left\|U_{i,j}\right\|_1 \\ & + \frac{1}{n(n-1)} \sum_{i \neq j} \left\| \widetilde{Y}_{i,j} \widetilde{Y}_{i,j}^T - S - \sqrt{n(n-1)}\, U_{i,j} \right\|_{\mathrm{F}}^2 \Bigg],\end{aligned} \tag{3.9}$$

where the minimization is over $S, U_{i,j} \in S^d(\mathbb{R})$, $\forall (i,j) \in I_n^2$.

Remark 1. The double penalized least-squares estimator in (3.9) is in fact a penalized Huber's estimator (previously observed by Donoho and

Montanari (2016, Section 6) in the context of linear regression). To see this, we express (3.9) as

$$(\widehat{S}_\lambda, \widehat{\mathbf{U}}_{\mathbf{I_n^2}}) = \arg\min_{S}\min_{\mathbf{U}_{\mathbf{I_n^2}}} \Bigg[\lambda_1 \|S\|_1 + \lambda_2 \sum_{i\neq j} \|U_{i,j}\|_1 + \frac{1}{n(n-1)} \sum_{i\neq j} \left\| \widetilde{Y}_{i,j}\widetilde{Y}_{i,j}^T - S - \sqrt{n(n-1)}U_{i,j} \right\|_{\mathrm{F}}^2 \Bigg], \quad (3.10)$$

and observe that the minimization with respect to $\mathbf{U}_{\mathbf{I_n^2}}$ in (3.10) can be done explicitly. It yields that

$$\widehat{S}_\lambda = \underset{S}{\operatorname{argmin}} \left\{ \frac{2}{n(n-1)} \operatorname{tr}\left[\sum_{i\neq j} \rho_{\frac{\sqrt{n(n-1)}\lambda_2}{2}} (\widetilde{Y}_{i,j}\widetilde{Y}_{i,j}^T - S) \right] + \lambda_1 \|S\|_1 \right\}, \quad (3.11)$$

where

$$\rho_\lambda(u) := \begin{cases} \frac{u^2}{2}, & |u| \le \lambda \\ \lambda |u| - \frac{\lambda^2}{2}, & |u| > \lambda \end{cases} \qquad \forall u \in \mathbb{R}, \lambda \in \mathbb{R}^+$$

is the Huber's loss function. Details of the derivation are presented in section A.6.1 of the supplementary material.

3.1 Main results

We are ready to state the main results related to the error bounds for the estimator in (3.9). We will compare performance of our estimator to that of the sample covariance matrix $\widetilde{\Sigma}_s$ defined in (3.3). When there are no outliers, it is well-known that $\widetilde{\Sigma}_s$ is a consistent estimator of Σ with expected error at most $\mathcal{O}(d/\sqrt{n})$ in the Frobenius norm, namely, $\left\|\widetilde{\Sigma}_s - \Sigma\right\|_F \leq Cd/\sqrt{n}$ with probability 99% for some absolute constant $C > 0$ (see for example, Cai et al. (2010)). However, in the presence of outliers, the error for $\widetilde{\Sigma}_s$ can be large (see Section 5 for some examples). On the other hand, recall that $\widetilde{X}_{i,j} = \widetilde{Z}_{i,j}\widetilde{Z}_{i,j}^T$, and our estimator in (3.9) admits the following bound.

Theorem 2. *Let $\delta > 0$ be an absolute constant. Assume that $n \geq 2$ and $|J| \leq c_1(\delta)n$, where $c_1(\delta)$ is a constant depending only on δ. Then on the event*

$$\mathcal{E} = \Bigg\{\lambda_1 \geq \frac{140\,\|\Sigma\|}{\sqrt{n(n-1)}}\sqrt{\mathrm{rk}(\Sigma)} + 4\left\|\frac{1}{n(n-1)}\sum_{(i,j)\in I_n^2}\widetilde{X}_{i,j} - \Sigma\right\|, \\ \lambda_2 \geq \frac{140\,\|\Sigma\|}{n(n-1)}\sqrt{\mathrm{rk}(\Sigma)} + \frac{4}{\sqrt{n(n-1)}}\max_{(i,j)\in I_n^2}\left\|\widetilde{X}_{i,j} - \Sigma\right\|\Bigg\},$$

the following inequality holds:

$$\left\|\widehat{S}_\lambda - \Sigma\right\|_{\mathrm{F}}^2 \leq \inf_{S:\,\mathrm{rank}(S)\leq \frac{c_2 n^2\lambda_2^2}{\lambda_1^2}} \Big\{(1+\delta)\,\|S-\Sigma\|_{\mathrm{F}}^2 \\ + c(\delta)\big(\lambda_1^2\,\mathrm{rank}(S) + \lambda_2^2|J|^2\big)\Big\},$$

where c_2 is an absolute constant and $c(\delta)$ is a constant depending only on δ.

The proof of Theorem 2 is presented in section A.2 of the supplementary material.

Remark 2. The bound in Theorem 2 contains two terms:

1. The first term, $(1+\delta)\|S-\Sigma\|_{\mathrm{F}}^2 + c(\delta)\lambda_1^2 \operatorname{rank}(S)$, does not depend on the number of outliers. When there are no outliers, i.e. $|J| = 0$, the bound will only contain this part, and in such a scenario Lounici (2014) proved that the theoretically optimal bound is

$$\left\|\widehat{S}_\lambda - \Sigma\right\|_{\mathrm{F}}^2 \le \inf_S \left\{ \|\Sigma - S\|_F^2 + C\|\Sigma\|^2 \frac{(\mathrm{rk}(\Sigma)+t)}{n} \operatorname{rank}(S) \right\}$$

with probability at least $1 - e^{-t}$. By making the smallest choice of λ_1 as specified in (3.12), one sees that the first term of our bound coincides with the theoretically optimal bound.

2. The second term, $c(\delta)\lambda_2^2|J|^2$, controls the worst possible effect due to the presence of outliers. When more conditions on the outliers are imposed (for example, independence), this bound can be improved. Moreover, Diakonikolas et al. (2017) proved that when Z is Gaussian with zero mean, there exists an estimator $\widehat{\Sigma}$ achieving theoretically optimal bound $\left\|\widehat{\Sigma} - \Sigma\right\|_F \le \mathcal{O}(\varepsilon)\|\Sigma\|$, which is independent of the dimension d. In our case, by making the smallest choice of λ_2 as specified in (3.13), we can show that the error bound scales like $C(\log(n) + \mathrm{rk}(\Sigma))\,\varepsilon\,\|\Sigma\|$, where $C > 0$ is an absolute constant. The

additional factor $(\log(n)+\mathrm{rk}(\Sigma))$ shows that our bound is sub-optimal in general. However, if $\mathrm{rk}(\Sigma)$ is small, our bound is essentially optimal up to a logarithmic factor.

Note that in Theorem 2 the regularization parameters λ_1, λ_2 should be chosen sufficiently large such that the event $\mathcal{E}$ happens with high probability. Under the assumption that $Z_j, j = 1, \ldots, n$ are independent, identically distributed L-sub-Gaussian vectors, we can prove the following result which gives an explicit lower bound on the choice of λ_1.

Theorem 3. *Assume that Z is L-sub-Gaussian with mean μ and covariance matrix Σ. Let $Z_1, \ldots, Z_n$ be independent copies of Z, and define $\widetilde{Z}_{i,j} := (Z_i - Z_j)/\sqrt{2}$ for any $(i,j) \in I_n^2$. Then $\widetilde{Z}_{i,j}, (i,j) \in I_n^2$ are mean zero L-sub-Gaussian random vectors with the same covariance matrix Σ. Moreover, for any $t \geq 1$, there exists $c(L) > 0$ depending only on L such that*

$$\left\| \frac{1}{n(n-1)} \sum_{i \neq j} \widetilde{Z}_{i,j} \widetilde{Z}_{i,j}^T - \Sigma \right\| \leq c(L) \, \|\Sigma\| \left(\sqrt{\frac{\mathrm{rk}(\Sigma) + t}{n}} + \frac{\mathrm{rk}(\Sigma) + t}{n} \right)$$

with probability at least $1 - 2e^{-t}$.

Theorem 3 along with the definition of event $\mathcal{E}$ indicates that it suffices to choose λ_1 satisfying

$$\lambda_1 \geq c(L) \|\Sigma\| \sqrt{\frac{\mathrm{rk}(\Sigma)+t}{n}}, \tag{3.12}$$

given that $n \geq \mathrm{rk}(\Sigma)+t$. The next theorem provides a lower bound for the choice of λ_2:

Theorem 4. *Assume that Z is L-sub-Gaussian with mean zero and $Z_1, \ldots, Z_n$ are samples of Z (not necessarily independent), then there exists $c(L) > 0$ depending only on L, such that for any $t \geq 1$,*

$$\max_{j=1,\ldots,n} \|Z_j Z_j^T - \Sigma\| \leq c(L) \|\Sigma\| (\mathrm{rk}(\Sigma) + \log(n) + t)$$

with probability at least $1 - e^{-t}$.

Note that Theorem 4 does not require independence of samples, so it can be applied to the mean zero, L-sub-Gaussian vectors $\widetilde{Z}_{i,j}, (i,j) \in I_n^2$ to deduce that

$$\max_{i \neq j} \left\| \widetilde{Z}_{i,j} \widetilde{Z}_{i,j}^T - \Sigma \right\| \leq c(L) \|\Sigma\| [\mathrm{rk}(\Sigma) + \log(n(n-1)) + t]$$

with probability at least $1 - e^{-t}$. Combining this bound with the definition of event $\mathcal{E}$, we conclude that it suffices to choose λ_2 satisfying

$$\lambda_2 \geq c(L) \|\Sigma\| \frac{(\mathrm{rk}(\Sigma) + \log(n) + t)}{n}. \tag{3.13}$$

By choosing the smallest possible λ_1, λ_2 as indicated in (3.12)(3.13), we deduce the following corollary:

Corollary 2. *Let $\delta > 0$ be an absolute constant. Assume that $n \geq \mathrm{rk}(\Sigma) + \log(n)$ and $|J| \leq c_1(\delta)n$, where $c_1(\delta)$ is a constant depending only on δ. Then we have that*

$$\left\|\widehat{S}_\lambda - \Sigma\right\|_F^2 \leq \inf_{S:\,\mathrm{rank}(S) \leq c_2' n \left(\mathrm{rk}(\Sigma)+\log(n)\right)} \Bigg\{ (1+\delta)\left\|S-\Sigma\right\|_{\mathrm{F}}^2 + c(L,\delta)\left\|\Sigma\right\|^2 \left[\frac{\mathrm{rk}(\Sigma)+\log(n)}{n}\,\mathrm{rank}(S) + \frac{\left(\mathrm{rk}(\Sigma)+\log(n)\right)^2}{n^2}|J|^2\right]\Bigg\} \tag{3.14}$$

with probability at least $1-3/n$, where c_2' is an absolute constant and $c(L,\delta)$ is a constant depending only on L and δ.

Note that the last term in (3.14) can be equivalently written in terms of ε, the proportion of outliers, as

$$c(L,\delta)\left\|\Sigma\right\|^2\left(\mathrm{rk}(\Sigma)+\log(n)\right)^2 \varepsilon^2. \tag{3.15}$$

4. The case of heavy-tailed data

In this section, we consider the application as well as possible improvements of the previously discussed results to heavy-tailed data. Let $Y \in \mathbb{R}^d$ be a random vector with mean $\mathbb{E}[Y] = \mu$, covariance matrix $\Sigma = \mathbb{E}\left[(Y-\mu)(Y-\mu)^T\right]$, and such that $\mathbb{E}\left[\|Y-\mu\|_2^4\right] < \infty$. Assume that $Y_1, \ldots, Y_n$ are i.i.d copies of Y, and as before our goal is to estimate Σ. Since μ is unknown and the estimation of μ is non-trivial for the heavy tailed-data, we consider the setting $\widetilde{Y}_{i,j} = (Y_i - Y_j)/\sqrt{2}$ and denote, for brevity,

$$H_{i,j} := \widetilde{Y}_{i,j}\widetilde{Y}_{i,j}^T.$$

We have previously shown that $\mathbb{E}\left[\widetilde{Y}_{i,j}\right] = 0$ and $\mathbb{E}[H_{i,j}] = \Sigma$, so the mean estimation is no longer needed for $\widetilde{Y}_{i,j}$. Given $\lambda_1, \lambda_2 > 0$, we propose the following estimator for Σ:

$$\widehat{S}_\lambda = \operatorname*{argmin}_S \left\{ \frac{1}{n(n-1)} \text{tr} \left[\sum_{i \neq j} \rho_{\frac{\sqrt{n(n-1)}\lambda_2}{2}} (\widetilde{Y}_{i,j}\widetilde{Y}_{i,j}^T - S) \right] + \frac{\lambda_1}{2} \|S\|_1 \right\}, \quad (4.16)$$

which is the minimizer of the penalized Huber's loss function:

$$L(S) = \frac{1}{n(n-1)} \text{tr} \left[\sum_{i \neq j} \rho_{\frac{\sqrt{n(n-1)}\lambda_2}{2}} (\widetilde{Y}_{i,j}\widetilde{Y}_{i,j}^T - S) \right] + \frac{\lambda_1}{2} \|S\|_1 . \quad (4.17)$$

Recall that the estimator $\widehat{S}_\lambda$ in (4.16) is equivalent to the double-penalized least-squares estimator in (3.9). The key idea to derive an error bound for $\widehat{S}_\lambda$ is motivated by Prasad et al. (2019), which suggests that it is possible to decompose any heavy-tailed distribution as a mixture of a "well-behaved" and a contamination components. The decomposition bridges the gap between the heavy-tailed model and the outlier model (3.2), allowing us to follow an argument similar to that in Section 3. To be precise, we consider the decomposition

$$\widetilde{Y}_{i,j} = \underbrace{\widetilde{Y}_{i,j}\mathbb{1}\left\{\left\|\widetilde{Y}_{i,j}\right\|_2 \le R\right\}}_{:=\widetilde{Z}_{i,j}} + \underbrace{\widetilde{Y}_{i,j}\mathbb{1}\left\{\left\|\widetilde{Y}_{i,j}\right\|_2 > R\right\}}_{:=\widetilde{V}_{i,j}}, \tag{4.18}$$

where $R > 0$ is the truncation level that will be specified later. In the following two subsections, we will separately show that the estimator $\widehat{S}_\lambda$ in (4.16) is close to Σ in both the operator norm and the Frobenius norm.

4.1 Bounds in the operator norm

In this subsection we show that $\widehat{S}_\lambda$ is close to Σ in the operator norm with high probability. We will be interested in the effective rank of the matrix $\mathbb{E}[(H_{1,2} - \Sigma)^2]$, and denote it as

$$r_H := \mathrm{rk}(\mathbb{E}\big[(H_{1,2} - \Sigma)^2\big]) = \frac{\mathrm{tr}\left(\mathbb{E}[(H_{1,2} - \Sigma)^2]\right)}{\|\mathbb{E}[(H_{1,2} - \Sigma)^2]\|}.$$

Minsker and Wei (2020, Lemma 4.1) suggest that under the bounded kurtosis assumption (to be specified later, see (4.19)), we can upper bound r_H by the effective rank of Σ, namely, $r_H \le C \cdot \mathrm{rk}(\Sigma)$ with some absolute constant C. We first present a lemma which shows that if the tuning parameter λ_1 is too large, the estimator $\widehat{S}_\lambda$ will be a zero matrix with high probability.

Lemma 1. *Assume that $t \ge 0$, $\sigma \ge \|\mathbb{E}[(H_{1,2} - \Sigma)^2]\|^{\frac{1}{2}}$ and*

$$n \ge \max\left\{64a^2 t \cdot r_H, \frac{4b^2t^2\left\|\Sigma\right\|^2}{\sigma^2}\right\},$$

where a,b are sufficiently large constants. Then for any $\lambda_1 > \frac{\sigma}{4}\sqrt{\frac{n}{t}}$*, we have that* $\mathrm{argmin}_S\, L(S) = 0$ *with probability at least* $1 - e^{-t}$*.*

Lemma 1 immediately implies that for the choice of $\lambda_1 > \frac{\sigma}{4}\sqrt{\frac{n}{t}}$, $\left\|\widehat{S}_\lambda - \Sigma\right\| = \|\Sigma\|$ with high probability, which is bounded by the largest singular value of Σ. The following theorem provides a bound for the choice of $\lambda_1 \leq \frac{\sigma}{4}\sqrt{\frac{n}{t}}$.

Theorem 5. *Assume that* $t \geq 1$ *is such that*

$$r_H \frac{t}{n} \leq c_3$$

for some sufficiently small constant c_3*,* $\sigma \geq \|\mathbb{E}[(H_{1,2} - \Sigma)^2]\|^{\frac{1}{2}}$*, and* $n \geq \max\left\{64a^2 t \cdot r_H, \frac{4b^2t^2\|\Sigma\|^2}{\sigma^2}\right\}$ *for some sufficiently large constants a, b. Then for* $\lambda_1 \leq \frac{\sigma}{4}\sqrt{\frac{n}{t}}$ *and* $\lambda_2 \geq \sigma\sqrt{\frac{1}{(n-1)t}}$*, we have that*

$$\left\|\widehat{S}_\lambda - \Sigma\right\| \leq \frac{20}{39}\lambda_1 + \frac{80}{39}\sigma\sqrt{\frac{t}{n}} + \frac{40}{39}\lambda_2 t$$

with probability at least $1 - \left(\frac{8}{3}r_H + 1\right)e^{-t}$*.*

The proofs of Lemma 1 and Theorem 5 are presented in section A.3 of the supplementary material.

Remark 3. The bound in Theorem 5 is close to that in Minsker and Wei (2020, Corrolary 4.1), with an additional term $\frac{20}{39}\lambda_1$. This term comes from the penalization factor $\frac{\lambda_1}{2}\|S\|_1$, showing that by “shrinking” our estimator to a low rank matrix, we introduce a bias term bounded by a multiple of the tuning parameter λ_1.

Remark 4. According to Minsker and Wei (2020, Lemma 4.1), the "matrix variance" parameter σ^2 appearing in the statement of Theorem 5 can be bounded by $\|\Sigma\| \operatorname{tr}(\Sigma) = \operatorname{rk}(\Sigma) \|\Sigma\|^2$ under the bounded kurtosis assumption. More precisely, if we assume that the kurtoses of the linear forms $\langle Y, v\rangle$ are uniformly bounded by K, meaning that

$$\sup_{v: \|v\|_2 = 1} \frac{\mathbb{E}[\langle Y - \mathbb{E}[Y], v\rangle]^4}{\left[\mathbb{E}[\langle Y - \mathbb{E}[Y], v\rangle]^2\right]^2} \leq K \tag{4.19}$$

for any $v \in \mathbb{R}^d$. Then we have that

$$\left\|\mathbb{E}\left[(H_{1,2} - \Sigma)^2\right]\right\| \leq K \cdot \operatorname{rk}(\Sigma) \|\Sigma\|^2$$

and σ can be chosen as $C\sqrt{K \cdot \operatorname{rk}(\Sigma)}\, \|\Sigma\|$ with some absolute constant C. Moreover, in this case the assumptions on n and t in Lemma 1 and Theorem 5 can be reduced to a single assumption that $r_H \frac{t}{n} \leq c_3'$ for some sufficiently small constant c_3'. We will formally state condition (4.19) in the next subsection and derive additional results based on it.

4.2 Bounds in the Frobenius norm

In this subsection we show that $\widehat{S}_\lambda$ is close to the covariance matrix of Y in the Frobenius norm with high probability, under a slightly stronger assumption on the fourth moment of Y.

Definition 6. A random vector $Y \in \mathbb{R}^d$ is said to satisfy an $L_4 - L_2$ norm equivalence with constant K (also referred to as the bounded kurtosis assumption), if there exists a constant $K \geq 1$ such that

$$\left(\mathbb{E}\left[\langle Y-\mu, v\rangle^4\right]\right)^{\frac{1}{4}} \leq K\left(\mathbb{E}\left[\langle Y-\mu, v\rangle^2\right]\right)^{\frac{1}{2}} \tag{4.20}$$

for any $v \in \mathbb{R}^d$, where $\mu = \mathbb{E}[Y]$.

As previously discussed in Remark 4, condition (4.20) allows us to connect the matrix variance parameter σ^2 with $\mathrm{rk}(\Sigma_Y)$, the effective rank of the covariance matrix Σ_Y. We will assume that Y satisfies (4.20) with a constant K throughout this subsection.

Recall the decomposition

$$\widetilde{Y}_{i,j} = \underbrace{\widetilde{Y}_{i,j}\mathbb{1}\left\{\left\|\widetilde{Y}_{i,j}\right\|_2 \leq R\right\}}_{:=\widetilde{Z}_{i,j}} + \underbrace{\widetilde{Y}_{i,j}\mathbb{1}\left\{\left\|\widetilde{Y}_{i,j}\right\|_2 > R\right\}}_{:=\widetilde{V}_{i,j}}, \tag{4.21}$$

where $R > 0$ is the truncation level that will be specified later. Denote $\Sigma_Y := \mathbb{E}\left[\widetilde{Y}_{1,2}\widetilde{Y}_{1,2}^T\right]$, $\Sigma_Z := \mathbb{E}\left[\widetilde{Z}_{1,2}\widetilde{Z}_{1,2}^T\right]$ and recall that our goal is to estimate Σ_Y. Note that $\left\|\widetilde{Z}_{i,j}\right\|_2 \leq R$ almost surely, so equation (4.21) represents $\widetilde{Y}_{i,j}$ as a sum of a bounded vector $\widetilde{Z}_{i,j}$ and a "contamination" component $\widetilde{V}_{i,j}$, which is similar to (3.2). On the other hand, we note that the truncation level R should be chosen to be neither too large (to get a better truncated distribtuion) nor too small (to reduce the bias introduced by the truncation). Mendelson and Zhivotovskiy (2020) suggest that a reasonable choice is given as follows:

$$R = \left(\frac{\operatorname{tr}(\Sigma_Y) \|\Sigma_Y\| n}{\log\big(\operatorname{rk}(\Sigma_Y)\big) + \log(n)} \right)^{\frac{1}{4}}. \tag{4.22}$$

Denote $\widetilde{J} = \left\{ (i,j) \in I_n^2 : \left\| \widetilde{Y}_{i,j} \right\|_2 > R \right\}$ to be the set of indices corresponding to the nonzero outliers (i.e. $\widetilde{V}_{i,j} \neq 0$), and $\varepsilon := |\widetilde{J}|/\big(n(n-1)\big)$ to be the proportion of outliers. Under the above setup, we can derive the following lemma which provides an upper bound on ε with high probability:

Lemma 2. *Assume that Y satisfies the $L_4 - L_2$ norm equivalence with constant K, and R is chosen as in* (4.22) *. Then*

$$\varepsilon \leq c(K) \frac{\operatorname{rk}(\Sigma_Y) \left[\log\big(\operatorname{rk}(\Sigma_Y)\big) + \log(n) \right]}{n} \tag{4.23}$$

with probability at least $1 - 1/n$, where $c(K)$ is a constant only depending on K.

The proof of Lemma 2 is presented in section A.4.1 of the supplementary material. It is worth noting that the proportion of "outliers" (in a sense of the definition above) in the heavy-tailed model can be pretty small when the sample size n is large. Consequently, we can derive the following bound:

Theorem 6. *Given $A \geq 1$, assume that $Y \in \mathbb{R}^d$ is a random vector with mean $\mathbb{E}[Y] = \mu$, covariance matrix $\Sigma_Y = \mathbb{E}\big[(Y-\mu)(Y-\mu)^T\big]$, and* *satisfying an $L_4 - L_2$ norm equivalence with constant K. Let $Y_1, \ldots, Y_n$ be i.i.d samples of Y, and let $\widetilde{Z}_{i,j}$ be defined as in* (4.21). *Assume that* $n \geq c_4(K)\mathrm{rk}(\Sigma_Y)\log(n \cdot \mathrm{rk}(\Sigma_Y))$ *for some constant $c_4(K)$ depending only on K, and* $\mathrm{rank}(\Sigma_Y) \leq c_2(K) \cdot n$ *for some constant $c_2(K)$ depending only on K. Then for* $\lambda_1 = c(K)\left\|\Sigma_Y\right\|\sqrt{\frac{\mathrm{rk}(\Sigma_Y)\left[\log\left(n\cdot\mathrm{rk}(\Sigma_Y)\right)\right]}{n}}$ *and* $\lambda_2 = c(K)\left\|\Sigma_Y\right\|\sqrt{\frac{\mathrm{rk}(\Sigma_Y)\log(n)}{An}}$, *we have that*

$$\left\|\widehat{S}_\lambda - \Sigma_Y\right\|_F^2 \leq c(K)\left\|\Sigma_Y\right\|^2\left[\frac{\mathrm{rk}(\Sigma_Y)\log\left(n\cdot\mathrm{rk}(\Sigma_Y)\right)}{n}\mathrm{rank}(\Sigma_Y) + \frac{A\cdot\mathrm{rk}(\Sigma_Y)^2\log(n)^3}{n}\right]$$

with probability at least $1 - \frac{(\frac{8}{3}r_H+1)}{n^A} - \frac{4}{n}$, *where $c(K)$ is a constant depending only on K.*

The proof of Theorem 6 is given in section A.4.2 of the supplementary material.

Remark 5. Let us compare the result in Theorem 6 to the bound of Corollary 2.

1. The first part of the bound,

$$c(K)\left\|\Sigma_Y\right\|^2\frac{\mathrm{rk}(\Sigma_Y)\log\left(n\cdot\mathrm{rk}(\Sigma_Y)\right)}{n}\mathrm{rank}(\Sigma_Y),$$

   has the same order as in Corollary 2 (up to a logarithmic factor), under the assumption that Σ_Y has low rank. This part of the bound is theoretically optimal according to Remark 2.

2. The second part of the bound,

$$c(K)\left\|\Sigma_Y\right\|^2 \frac{\mathrm{rk}(\Sigma_Y)^2 \log(n)^3}{n}, \tag{4.24}$$

controls the error introduced by the outliers. It is much smaller than the corresponding quantity in Corollary 2 when ε, the proportion of the outliers, is only assumed to be a constant. The improvement is mainly due to the special structure of the heavy-tailed data, namely, the "outliers" $\widetilde{V}_{i,j}$ are mutually independent as long as the subscripts do not overlap, and hence there are many cancellations among them. Without this special structure, one can only apply Theorem 2 directly and derive a sub-optimal bound of order

$$c(K)\left\|\Sigma_Y\right\|^2 \frac{\mathrm{rk}(\Sigma_Y)^3 \log(n)^3}{n}.$$

5. Numerical experiments

In this section we present the results of our numerical experiments. Recall that our loss function is

$$\widetilde{L}(S, \boldsymbol{U_{I_n^2}}) = \lambda_1 \left\|S\right\|_1 + \lambda_2 \sum_{i\neq j} \left\|U_{i,j}\right\|_1 + \frac{1}{n(n-1)} \sum_{i\neq j} \left\| \widetilde{Y}_{i,j}\widetilde{Y}_{i,j}^T - S - \sqrt{n(n-1)}U_{i,j} \right\|_F^2. \tag{5.25}$$

We are aiming to find $(\widehat{S}_\lambda, \widehat{\boldsymbol{U}}_{\boldsymbol{I_n^2}})$, the minimizer of (5.25), numerically. Since we are only interested in $\widehat{S}_\lambda$, equation (3.11) suggests that it suffices to minimize the following function:

$$L(S) := \frac{1}{n(n-1)} \mathrm{tr} \sum_{i \neq j} \rho_{\frac{\sqrt{n(n-1)}\lambda_2}{2}} (\widetilde{Y}_{i,j} \widetilde{Y}_{i,j}^T - S) + \frac{\lambda_1}{2} \|S\|_1 \,, \tag{5.26}$$

where

$$\rho_\lambda(u) := \begin{cases} \frac{u^2}{2}, & |u| \leq \lambda \\ \lambda |u| - \frac{\lambda^2}{2}, & |u| > \lambda \end{cases} \qquad \forall u \in \mathbb{R}, \lambda \in \mathbb{R}^+$$

is the Huber's loss function. We will introduce the proximal gradient descent method in the next subsection, which helps us find the minimizer of (5.26) numerically.

5.1 Numerical algorithm

In this section, we will state our algorithm for minimizing $L(S)$. We start with an introduction to the proximal gradient method (see for example, Combettes and Wajs (2005)). Suppose that we want to minimize the function $f(x) = g(x) + h(x)$, where

- g is convex, differentiable
- h is convex, not necessarily differentiable

We define the proximal mapping and the proximal gradient descent method as follows:

Definition 7. The proximal mapping of a convex function h at the point x is defined as:

$$prox_h(x) = \underset{u}{\operatorname{argmin}} \left(h(u) + \frac{1}{2} \|u - x\|_2^2 \right).$$

Definition 8 (Proximal gradient descent (PGD) method)**.** The proximal gradient descent method for solving the problem $\operatorname{argmin}_x f(x) = \operatorname{argmin}_x g(x) + h(x)$ starts from an initial point $x^{(0)}$, and updates as:

$$x^{(k)} = \operatorname{prox}_{\alpha_k h} \left(x^{(k-1)} - \alpha_k \nabla g(x^{(k-1)}) \right),$$

where $\alpha_k > 0$ is the step size.

We have the following convergence result:

Theorem 7. *Assume that ∇g is Lipschitz continuous with constant $L > 0$:*

$$\|\nabla g(x) - \nabla g(y)\| \leq L \|x - y\|$$

and the optimal value f^* is finite and achieved at the point x^*. Then the proximal gradient algorithm with constant step size $\alpha_k = \alpha \leq L$ will yield an $\mathcal{O}(\frac{1}{k})$ convergence rate, i.e.

$$f(x^{(k)}) - f^* \leq \frac{C}{k}, \quad \forall k \in \{1, 2, \ldots\}.$$

Theorem 7 is well known (see for example, Beck (2017, Chapter 10)), but a detailed proof in our case is given in section A.5.1 of the supplementary material for the convenience of the reader. Moreover, when $g(x) =$

$\frac{1}{n}\sum_{i=1}^{n} g_i(x)$, where $g_1, \ldots, g_n$ are convex functions and $\nabla g_1, \ldots, \nabla g_n$ are Lipschitz continuous with a common constant $L > 0$, the update step of PGD will require the evaluation of n gradients, which is expensive when n is large. A natural improvement is to consider the stochastic proximal gradient descent method (SPGD), where at each iteration $k = 1, 2, \ldots$, we pick an index i_k randomly from $\{1, 2, \ldots, n\}$, and take the following update:

$$x^{(k)} = \text{prox}_{\alpha_k h}\left(x^{(k-1)} - \alpha_k \nabla g_{i_k}(x^{(k-1)})\right).$$

The advantage of SPGD over PGD is that the computational cost of SPGD per iteration is $1/n$ that of the PGD. On the other hand, since the random sampling in SPGD introduces additional variance, we need to choose a diminishing step size $\alpha_k = \mathcal{O}(\frac{1}{k})$. As a result, the SPGD only converges at a sub-linear rate (see Nitanda (2014)). To this end, we will consider the "mini-batch" PGD, which has been previously explored and widely used in large-scale learning problems (see, e.g., Shalev-Shwartz et al. (2011); Gimpel et al. (2010); Dekel et al. (2012); Khirirat et al. (2017)). This method picks a small batch of indices rather than one at each iteration to calculate the gradient, and in such a way we are able balance the computational cost of PGD and the additional variance of SPGD. The algorithm is summarized in Algorithm 1.

Algorithm 1 Stochastic proximal gradient descent (SPGD)

Input: number of iterations T, step size η_t, batch size b, tuning parameters λ_1 and λ_2, initial estimation S^0, sample size n, dimension d.

for $t = 1, 2, \ldots, T$ **do**

(1) Randomly pick $i_t, j_t \in \{1, 2, \ldots, n\}$ one by one without replacement.

(2) Compute $G_t = -\nabla g_{i,j}(S^t) = -\rho'_{\frac{\sqrt{n(n-1)}\lambda_2}{2}}(\widetilde{Y}_{i,j}\widetilde{Y}_{i,j}^T - S^t)$.

(3) If $b > 1$, then repeat (1)(2) for b times and save the average gradient in G_t.

(4) (**gradient update**)

$$T^{t+1} = S^t - G_t.$$

(5) (**proximal update**)

$$S^{t+1} = \underset{S}{\operatorname{argmin}} \left\{ \frac{1}{2} \left\| S - T^{t+1} \right\|_F^2 + \frac{\lambda_1}{2} \left\| S \right\|_1 \right\} = \gamma_{\frac{\lambda_1}{2}}(T^{t+1}),$$

where $\gamma_\lambda(u) = \mathrm{sign}(u)(|u| - \lambda)_+$.

end for

Output: S^{T+1}

5.2 Rank-one update of the spectral decomposition

Note that at each iteration of our algorithm, we need to compute the spectral decomposition of $(\widetilde{Y}_{i,j}\widetilde{Y}_{i,j}^T - S^t)$, which is computationally expensive. However, since $\widetilde{Y}_{i,j}\widetilde{Y}_{i,j}^T$ is a rank-one matrix and S^t was already saved in the spectral decomposition form after previous iteration, the problem of computing the spectral decomposition of $(\widetilde{Y}_{i,j}\widetilde{Y}_{i,j}^T - S^t)$ can be viewed as a rank-one update of the spectral decomposition, which has been extensively studied (see for example, Bunch et al. (1978), and Stange (2008)). In this subsection we will show how to use this idea to improve our algorithm.

Consider $\widetilde{B} = B + \rho bb^T$, where the spectral decomposition $B = QDQ^T$ is known, $\rho \in \mathbb{R}$ and $b \in \mathbb{R}^d$. Our target is to compute the spectral decomposition of $\widetilde{B}$. Note that

$$\widetilde{B} = B + \rho bb^T = Q(D + \rho zz^T)Q^T, \tag{5.27}$$

where $b = Qz$, so it suffices to compute the spectral decomposition of $D + \rho zz^T$. We denote $z = (\zeta_1, \ldots, \zeta_d)^T$, and without loss of generality, we can assume that $\|z\|_2 = 1$. The following theorem is fundamental for our algorithm.

Theorem 8. *(Bunch et al. (1978, Theorem 1)) Let $C = D + \rho zz^T$, where D is diagonal, $\|z\|_2 = 1$. Let $d_1 \leq d_2 \leq \ldots \leq d_d$ be the eigenvalues of*

D, and let $\widetilde{d}_1 \le \widetilde{d}_2 \le \ldots \widetilde{d}_d$ be the eigenvalues of C. Then $\widetilde{d}_i = d_i + \rho\mu_i$, $1 \le i \le d$ where $\sum_{i=1}^{n} \mu_i = 1$ and $0 \le \mu_i \le 1$. Moreover, $d_1 \le \widetilde{d}_1 \le d_2 \le \widetilde{d}_2 \le \ldots \le d_d \le \widetilde{d}_d$ if $\rho > 0$ and $\widetilde{d}_1 \le d_1 \le \widetilde{d}_2 \le d_2 \le \ldots \le \widetilde{d}_d \le d_d$ if $\rho < 0$. Finally, if d_i's are distinct and all the elements of z are nonzero, then the eigenvalues of C strictly separate those of D.

There are several cases where we can deflate the problem (i.e. reduce the size of the problem):

1. If $\zeta_i = 0$ for some i, then $\widetilde{d}_i = d_i$ and the corresponding eigenvector remains unchanged. This is because $(D + \rho z z^T)e_i = d_i e_i$ as $\zeta_i = 0$.

2. If $|\zeta_i| = 1$ for some i, then $\widetilde{d}_i = d_i + \rho$ and the corresponding eigenvector remains unchanged. Moreover, in this case $\zeta_j = 0$ for all $j \ne i$, so $\widetilde{d}_j = d_j$ and their eigenvectors are the same, so the problem is done.

3. If d_i has a multiplicity $r \ge 2$, we can reduce the size of the problem via the following steps:

   (a) Let $Q_1 = [q_{i_1}, \ldots, q_{i_r}] \in \mathbb{R}^{d \times r}$, where $\{q_{i_1}, \ldots, q_{i_r}\}$ are the eigenvectors corresponding to d_i. Also, set $z_1 = Q_1^T z$, i.e. z_1 contains rows corresponding to d_i.

   (b) Construct an Householder transformation $H \in \mathbb{R}^{r \times r}$ such that $H z_1 = - \|z_1\|_2 \, e_1$, and define $\bar{Q}_1 = Q_1 H^T$.

(c) Replace $q_{i_1}, \ldots, q_{i_r}$ by the columns of $\bar{Q}_1$, and z_1 by $\bar{Q}_1^T z_1 = -\|z_1\|_2 e_1$. This introduces $(r-1)$ more zero entries to z and possibly an entry with absolute value equals one. An application of (1)(2) gives us $(r-1)$ (or r) more eigen-pairs of $D + \rho z z^T$.

After the deflation step, it remains to work with a $k \times k$ problem ($k \leq d$), in which the eigenvalues d_i are distinct and $\zeta_i \neq 0$ for all i. We will compute the eigenvalues and eigenvectors separately.

First, Golub (1973) showed that the eigenvalues of $C = D + \rho z z^T$ are the zeros of $\omega(\lambda)$, where

$$\omega(\lambda) = 1 + \rho \sum_{j=1}^{k} \frac{\zeta_j^2}{d_j - \lambda}.$$

Alternatively, since $\widetilde{d}_1 < \ldots < \widetilde{d}_k$ and $\widetilde{d}_i = d_i + \rho\mu_i$, for each $i = 1, \ldots, k$ we can compute μ_i by solving $\omega_i(\mu_i) = 0$, where

$$\omega_i(\mu) = 1 + \sum_{j=1}^{k} \frac{\zeta_j^2}{\delta_j - \mu}. \tag{5.28}$$

and $\delta_j = \frac{d_j - d_i}{\rho}$. Bunch et al. (1978) proved that we can solve $\omega_i(\mu) = 0$ with a numerical method that converges quadratically. Details of the numerical method are presented in section A.5.2 of the supplementary material.

Second, after computing the eigenvalues $\widetilde{d}_1, \ldots, \widetilde{d}_k$, we can calculate the corresponding eigenvectors of $C = D + \rho z z^T$ by solving $C\widetilde{q}_i = \widetilde{d}_i \widetilde{q}_i$,

$i = 1, \ldots, k$. Theorem 5 in Bunch et al. (1978) shows that $\widetilde{q}_i$ can be computed via

$$\widetilde{q}_i = \frac{D_i^{-1} z}{\left\| D_i^{-1} z \right\|_2}, \tag{5.29}$$

where $D_i := D - \widetilde{d}_i I$. Finally, once we obtained the spectral decomposition of $D + \rho z z^T = \bar{Q} \widetilde{D} \bar{Q}^T$, we can easily get the decomposition of $B + \rho b b^T = (Q\bar{Q}) \widetilde{D} (Q\bar{Q})^T$. Note that computing k eigenvectors via (5.29) costs $\mathcal{O}(k^3)$ and the matrix multiplication $Q\bar{Q}$ in the last step costs $\mathcal{O}(d^3)$, so the overall complexity of the algorithm is still $\mathcal{O}(d^3)$. This can be further improved by exploiting the special structure of $\bar{Q}$, which is given by the product (see for example, Stange (2008) and Gandhi and Rajgor (2017)):

$$\bar{Q} = \underbrace{\begin{bmatrix} \zeta_1 & & \\ & \ddots & \\ & & \zeta_d \end{bmatrix}}_{:=A} \underbrace{\begin{bmatrix} \frac{1}{d_1 - \widetilde{d}_1} & \cdots & \frac{1}{d_1 - \widetilde{d}_d} \\ \vdots & & \vdots \\ \frac{1}{d_d - \widetilde{d}_1} & \cdots & \frac{1}{d_d - \widetilde{d}_d} \end{bmatrix}}_{:=C} \begin{bmatrix} \left\| \bar{c}_{\cdot 1} \right\|_2 & & \\ & \ddots & \\ & & \left\| \bar{c}_{\cdot d} \right\|_2 \end{bmatrix}^{-1} \tag{5.30}$$

where $\bar{C} := AC = [\bar{c}_{\cdot 1}, \ldots, \bar{c}_{\cdot d}]$, $\bar{c}_{\cdot i}$ represents the i^{th} column of $\bar{C}$, and $\left\| \bar{c}_{\cdot i} \right\|_2$ is the Euclidean norm of $\bar{c}_{\cdot i}$. Using (5.30), we can evaluate the matrix multiplication $Q\bar{Q}$ through the following steps:

1. Compute $QA := U = [u_{\cdot 1}, \ldots, u_{\cdot d}]$, where $u_{\cdot i} = \zeta_i q_{\cdot i}$ and $q_{\cdot i}$ is the i^{th} column of Q. This step is straightforward and requires $\mathcal{O}(d^2)$ computational time.

2. Let $u_{i\cdot}$ be the i^{th} row of U. Define

$$\widetilde{U} = UC = \begin{bmatrix} u_{1\cdot}C \\ \vdots \\ u_{d\cdot}C \end{bmatrix},$$

which requires to evaluate the product of a vector $u_{i\cdot}$ and a Cauchy matrix C d times. The problem of multiplying a Cauchy matrix with a vector is called Trummer's problem, and Gandhi and Rajgor (2017) provide an algorithm which efficiently computes such matrix-vector product in $\mathcal{O}(d \log^2 d)$ time. Consequently, the complexity of this step is $\mathcal{O}(d^2 \log^2 d)$.

3. Compute the matrix product

$$\widetilde{U} \begin{bmatrix} \|\bar{c}_{\cdot 1}\|_2 & & \\ & \ddots & \\ & & \|\bar{c}_{\cdot d}\|_2 \end{bmatrix}^{-1}.$$

This step is again straightforward and can be done in $\mathcal{O}(d^2)$ time.

The overall complexity for the computation of $Q\bar{Q}$ is now reduced to $\mathcal{O}(d^2 \log^2 d)$, which is much smaller than $\mathcal{O}(d^3)$ when d is large. We summarized our rank-one update algorithm in Algorithm 2.

Algorithm 2 Rank-one update of the spectral decomposition of $B + \rho b b^T$

Input: orthogonal matrix Q and vector d such that $B = Q\text{diag}(d)Q^T$, constant ρ, vector b

1: Set $Qb = z$. If $\|z\|_2 \neq 1$, then further set $\rho = \rho \|z\|_2^2$ and $z = z/\|z\|_2$.

2: Handle deflation cases, and record indices that have not done as a vector d_{sub}.

3: Compute eigenvalues of the $d_{sub} \times d_{sub}$ sub-problem by solving (5.28) numerically.

4: Compute eigenvectors of the $d_{sub} \times d_{sub}$ sub-problem with (5.29).

5: Combine the resulting eigenvalues in $\widetilde{d}$ and eigenvectors in $\bar{Q}$.

6: Compute $\widetilde{Q} = Q\bar{Q}$.

Output: orthogonal matrix $\widetilde{Q}$ and vector $\widetilde{d}$.

5.3 Numerical results

In this section we present some numerical results with different parameter settings. First, note that if we start with $S^0 = 0_{d\times d}$, we can easily compute the gradient in the first step of proximal gradient descent via

$$G = \rho'_{\frac{\sqrt{N}\lambda_2}{2}}(\widetilde{Y}_{i,j}\widetilde{Y}_{i,j}^T) = \frac{\rho'_{\frac{\sqrt{N}\lambda_2}{2}}(\left\|\widetilde{Y}_{i,j}\right\|_2^2)}{\left\|\widetilde{Y}_{i,j}\right\|_2^2}\widetilde{Y}_{i,j}\widetilde{Y}_{i,j}^T. \tag{5.31}$$

Here, no explicit spectral decomposition was required. Minsker and Wei (2020, Remark 4.1) provide details supporting the claim that the full gradient update at the first step helps to improve the initial guess of the estimator. Therefore, we will start with $S^0 = 0_{d\times d}$, run one step of PGD with the full data set, and use the output as the initial estimate of the solution.

Now consider the following parameter settings: $d = 100$, $n = 100$, $|J| = 3$, $\mu = (0, \ldots, 0)^T$, $\Sigma = \mathrm{diag}(10, 1, \frac{1}{100}, \ldots, \frac{1}{100})$. The samples are generated as follows: generate $n = 100$ independent samples Z_j from the Gaussian distribution $\mathcal{N}(\mu, \Sigma)$, and then replace $|J|$ of them (randomly chosen) with $Z_j + V_j$, where $V_j, j \in J$ are "outliers" to be specified later. The final results after replacement, denoted as Y_j's, are the samples we observe and that will be used as inputs for the SPGD algorithm. Next, we calculate $\widetilde{Y}_{i,j} = (Y_i - Y_j)/\sqrt{2}, i \neq j$ and perform our algorithm with $K = 500$ steps and the diminishing step size $\alpha_k = 1/k$. The initial value

S^0 is determined by a one-step full gradient update (5.31). To analyze the performance of estimators, we define

$$\text{rel\_error\_frob}(S) := \frac{\|S - \Sigma\|_F}{\|\Sigma\|}$$

to be the relative error of the estimator S in the Frobenius norm, and

$$\text{rel \_ error \_ op}(S) := \frac{\|S - \Sigma\|}{\|\Sigma\|}$$

to be the relative error of the estimator S in the operator norm, where S is an arbitrary estimator. We will compare the performance of the estimator S^* produced by our algorithm with the performance of the sample covariance matrix $\widetilde{\Sigma}_s$ introduced in (3.3). Here are some results corresponding to different types of outliers:

1. **Constant outliers.**

   Consider the outliers $V_j = (100, \ldots, 100)^T, j \in J$. We performed 200 repetitions of the experiment with $\lambda_1 = 3$, $\lambda_2 = 1$, and recorded S^*, $\widetilde{\Sigma}_s$ for each run. Histograms illustrating the distributions of relative errors are shown in Figure 1 and 2.

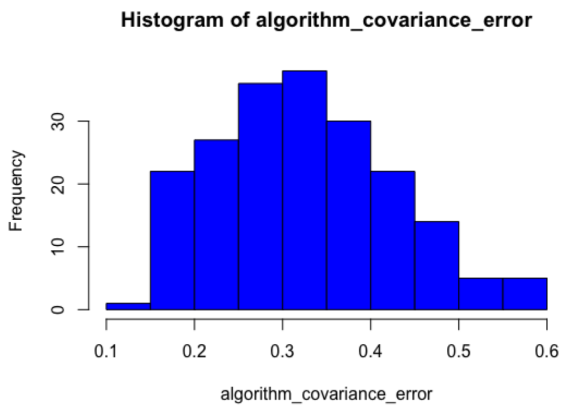

Figure 1: Distribution of the relative error of S^* in the Frobenius norm

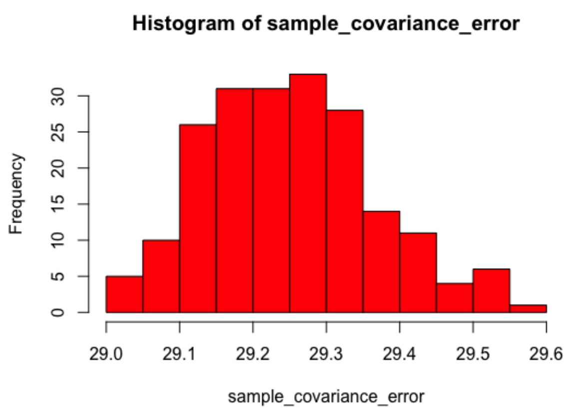

Figure 2: Distribution of the relative error of $\widetilde{\Sigma}_s$ in the Frobenius norm

The histograms show that $\widetilde{\Sigma}_s$ always produces a relative error in the Frobenius norm around 29.3, while S^* always produces a relative error in the Frobenius norm around 0.3. The average and maximum (over 200 repetitions) relative errors of S^* were 0.3250 and 0.5867 respectively, with the standard deviation of 0.0995. The corresponding values for $\widetilde{\Sigma}_s$ were 29.2512, 29.5641 and 0.1125. It is clear that in the considered scenario, estimator S^* performed noticeably better than the sample covariance $\widetilde{\Sigma}_s$.

In the meanwhile, the following histograms (Figure 3 and 4) show that S^* produces smaller relative errors in the operator norm as well. The average and maximum relative errors of S^* in the operator norm were 0.2734 and 0.5699 respectively, with the standard deviation of 0.0998. The corresponding values for $\widetilde{\Sigma}_s$ were 29.3981, 29.7125 and 0.1131.

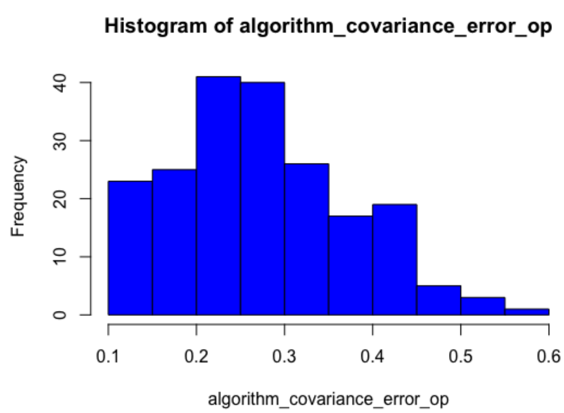

Figure 3: Distribution of the relative error of S^* in the operator norm

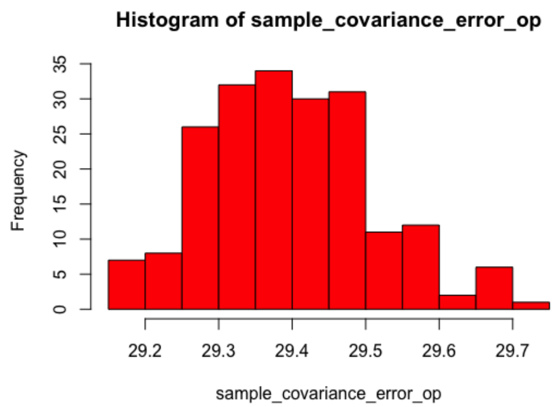

Figure 4: Distribution of the relative error of $\widetilde{\Sigma}_s$ in the operator norm

2. **Spherical Gaussian outliers.**

   Consider the case that the outliers V_j are drawn independently from a spherical Gaussian distribution $\mathcal{N}(\mu_V, \Sigma_V)$, where $\mu_V = (0, \ldots, 0)^T$, $\Sigma_V = \text{diag}(100, \ldots, 100)$. In this case, the outliers affect Z_j uniformly in all directions. We performed 200 repetitions of the experiment with $\lambda_1 = 3$, $\lambda_2 = 1$, and recorded S^*, $\widetilde{\Sigma}_s$ for each run. Histograms illustrating the distributions of relative errors are shown in Figure 5 and 6:

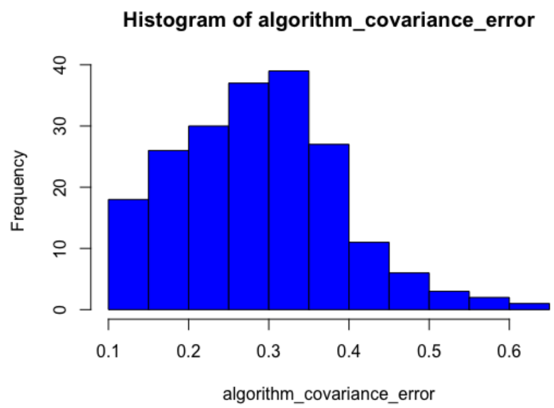

Figure 5: Distribution of the relative error of S^* in the Frobenius norm

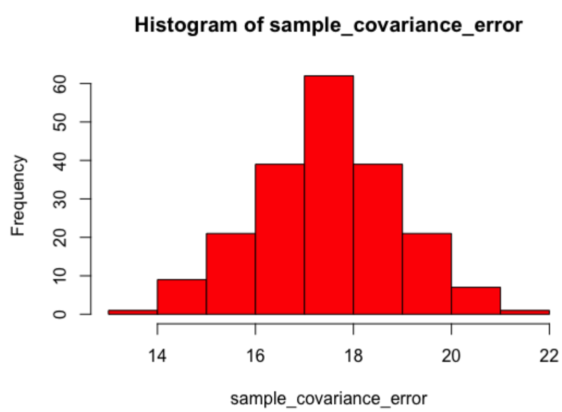

Figure 6: Distribution of the relative error of $\widetilde{\Sigma}_s$ in the Frobenius norm

The histograms show that $\widetilde{\Sigma}_s$ always produces a relative error in the Frobenius norm around 17, while S^* always produces a relative error in the Frobenius norm around 0.3. The average and maximum (over 200 repetitions) relative errors of S^* were 0.2865 and 0.6048 respectively, with the standard deviation of 0.1004. The corresponding values for $\widetilde{\Sigma}_s$ were 17.4683, 21.1723 and 1.4154. It is clear that in the considered scenario, estimator S^* performed noticeably better than the sample covariance $\widetilde{\Sigma}_s$.

In the meanwhile, the following histograms (Figure 7 and 8) show that S^* produces smaller relative errors in the operator norm as well. The average and maximum relative errors of S^* in the operator norm were 0.2690 and 0.6009 respectively, with the standard deviation of 0.1060. The corresponding values for $\widetilde{\Sigma}_s$ were 11.9925, 15.2436 and 1.1811.

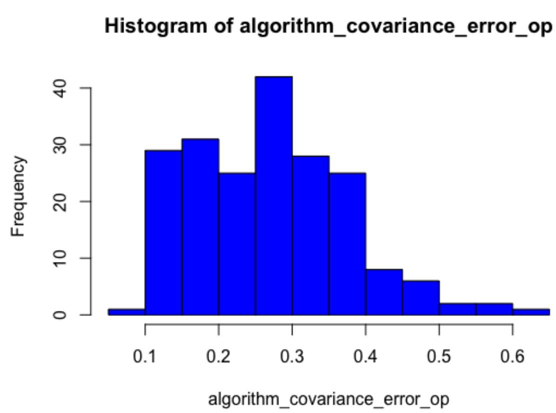

Figure 7: Distribution of the relative error of S^* in the operator norm

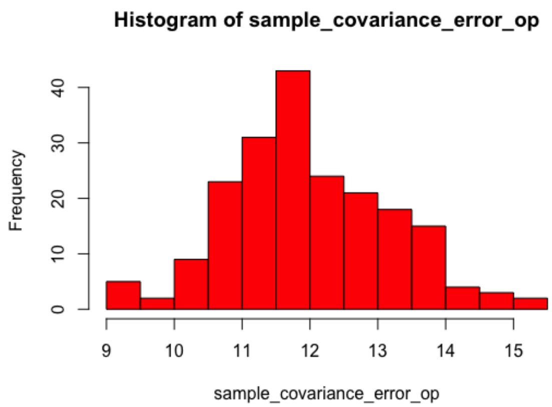

Figure 8: Distribution of the relative error of $\widetilde{\Sigma}_s$ in the operator norm

3. **Outliers that "erase" some observations.**

   Consider the case that the outliers are given as $V_j = \beta Z_j$ for $j \in J, \beta \in \mathbb{R}$. In this case, the outliers erase (when $\beta = -1$), amplify (when $\beta > 0$) or negatively amplify (when $\beta < -1$) some sample points Z_j. We performed 200 repetitions of the experiment with $\lambda_1 = \lambda_2 = 0.4$, $\beta = -50$ and recorded S^*, $\widetilde{\Sigma}_s$ for each run. Histograms illustrating the distributions of relative errors are shown in Figure 9 and 10:

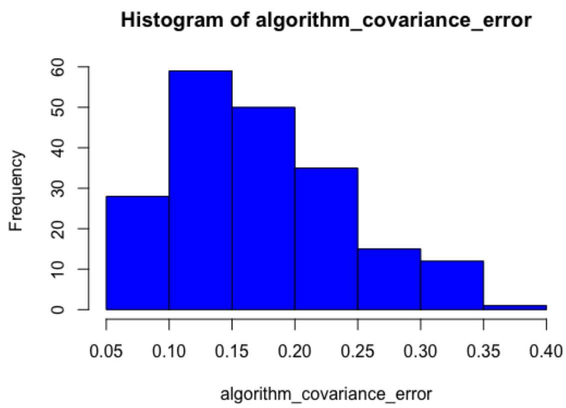

Figure 9: Distribution of the relative error of S^* in the Frobenius norm

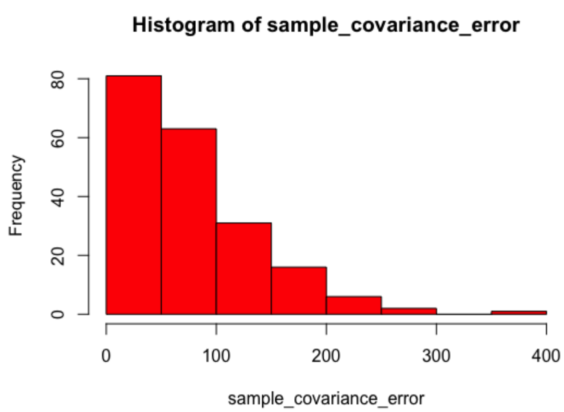

Figure 10: Distribution of the relative error of $\widetilde{\Sigma}_s$ in the Frobenius norm

The histograms show that S^* always produces a relative error in the Frobenius norm around 0.13, while $\widetilde{\Sigma}_s$ produces a relative error in the Frobenius norm around 50. Note that unlike previous examples, the performance of $\widetilde{\Sigma}_s$ is unstable in the current settings, with relative errors raising to 400 occasionally. The average and maximum (over 200 repetitions) relative errors of S^* were 0.1719 and 0.3692 respectively, with the standard deviation of 0.0680. The corresponding values for $\widetilde{\Sigma}_s$ were 80.1270, 366.5612 and 61.6144. It is clear that in the considered scenario, estimator S^* performed noticeably better than the sample covariance $\widetilde{\Sigma}_s$.

In the meanwhile, the following histograms (Figure 11 and 12) show that S^* produces smaller and more stable relative errors in the operator norm as well. The average and maximum relative errors of S^* in the operator norm were 0.1661 and 0.3689 respectively, with the standard deviation of 0.0721. The corresponding values for $\widetilde{\Sigma}_s$ were 79.6248, 368.3944 and 62.3943.

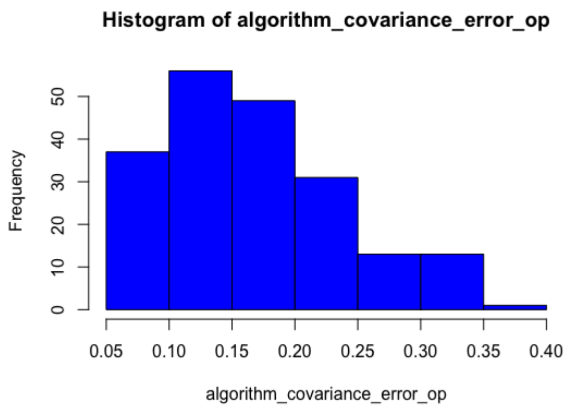

Figure 11: Distribution of the relative error of S^* in the operator norm

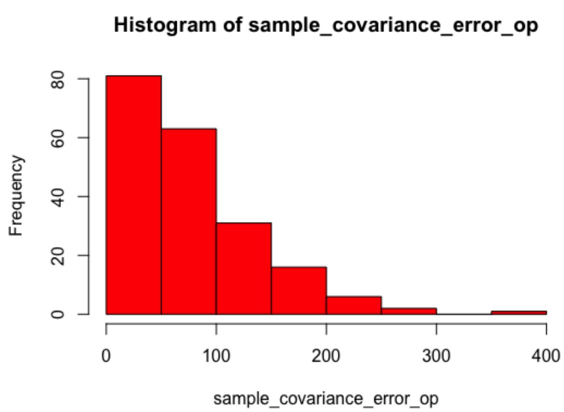

Figure 12: Distribution of the relative error of $\widetilde{\Sigma}_s$ in the operator norm

4. **Outliers in a particular direction.**

   Finally we consider the case that the outliers are all orthogonal (or parallel) to the subspace spanned by the first M principal components of Z, where Z is an $n \times d$ matrix with Z_j^T on each row. We performed 200 repetitions of the experiment with $\lambda_1 = 3$, $\lambda_2 = 1$, $M = 1$ (orthogonal case) and recorded S^*, $\widetilde{\Sigma}_s$ for each run. Histograms illustrating the distributions of relative errors are shown in Figure 13 and 14:

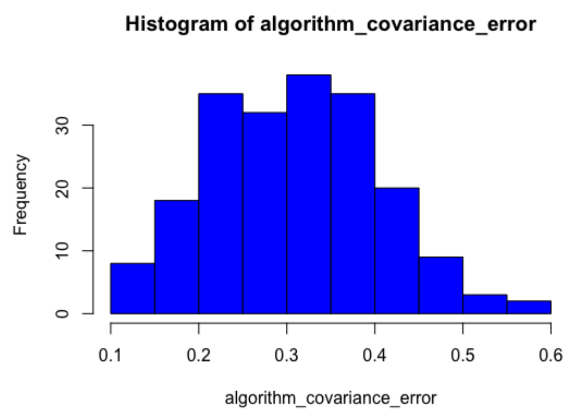

Figure 13: Distribution of the relative error of S^* in the Frobenius norm

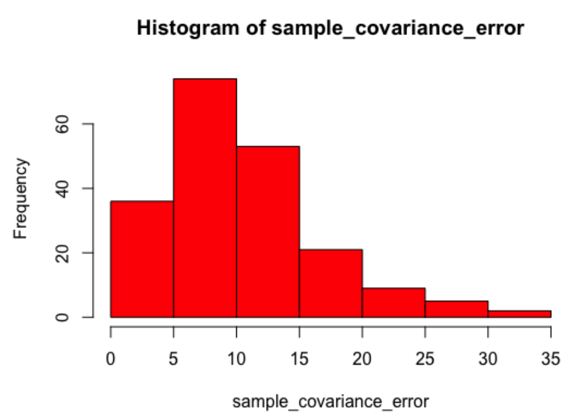

Figure 14: Distribution of the relative error of $\widetilde{\Sigma}_s$ in the Frobenius norm

The histograms show that $\widetilde{\Sigma}_s$ mainly produces a relative error in the Frobenius norm around 8, while S^* always produces a relative error in the Frobenius norm around 0.3. The average and maximum (over 200 repetitions) relative errors of S^* were 0.3083 and 0.5956 respectively, with the standard deviation of 0.0938. The corresponding values for $\widetilde{\Sigma}_s$ were 10.4196, 34.7976 and 6.0676. Note that the smallest error produced by $\widetilde{\Sigma}$ was 0.2930, which is comparable to the error produced by S^*. However, the histograms show that the small error produced by $\widetilde{\Sigma}_s$ only occurs occasionally, while S^* was producing small errors consistently. Therefore, in the considered scenario, we can still conclude that estimator S^* performed better than the sample covariance $\widetilde{\Sigma}_s$. In the meanwhile, the following histograms (Figure 15 and 16) show that S^* produces smaller relative errors in the operator norm as well. The average and maximum relative errors of S^* in the operator norm were 0.2648 and 0.5209 respectively, with the standard deviation of 0.0998. The corresponding values for $\widetilde{\Sigma}_s$ were 10.4703, 34.9707 and 6.0993.

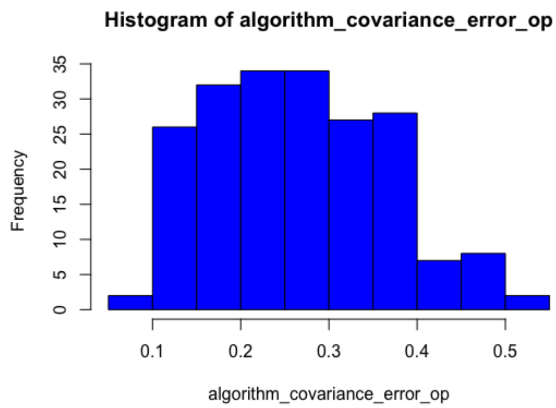

Figure 15: Distribution of the relative error of S^* in the operator norm

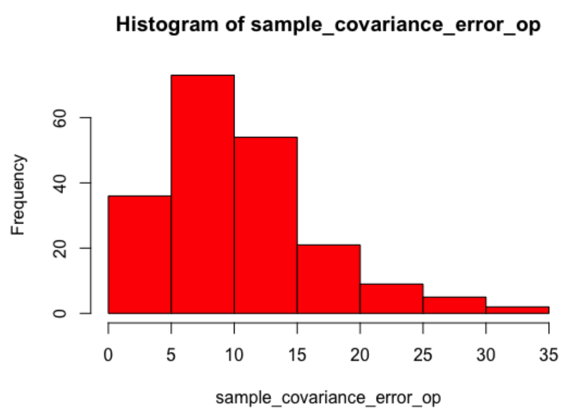

Figure 16: Distribution of the relative error of $\widetilde{\Sigma}_s$ in the operator norm

Acknowledgements

Authors acknowledge support by the National Science Foundation grant CCF-1908905.

References

Aleksandrov, A. B. and V. V. Peller (2016). Operator Lipschitz functions. *Russian Mathematical Surveys 71*(4), 605.

Beck, A. (2017). *First-order methods in optimization*. SIAM.

Belloni, A., V. Chernozhukov, and L. Wang (2011). Square-root Lasso: pivotal recovery of sparse signals via conic programming. *Biometrika 98*(4), 791–806.

Bernholt, T. (2006). Robust estimators are hard to compute. Technical report, Technical report.

Bhatia, R. (2013). *Matrix analysis*, Volume 169. Springer Science & Business Media.

Bunch, J. R., C. P. Nielsen, and D. C. Sorensen (1978). Rank-one modification of the symmetric eigenproblem. *Numerische Mathematik 31*(1), 31–48.

Cai, T. T., C.-H. Zhang, and H. H. Zhou (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics 38*(4), 2118–2144.

Chen, M., C. Gao, and Z. Ren (2018). Robust covariance and scatter matrix estimation under Huber's contamination model. *Annals of Statistics 46*(5), 1932–1960.

Cheng, Y., I. Diakonikolas, R. Ge, and D. P. Woodruff (2019). Faster algorithms for high-dimensional robust covariance estimation. In *Conference on Learning Theory*, pp. 727–757. PMLR.

Combettes, P. L. and V. R. Wajs (2005). Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation 4*(4), 1168–1200.

Dekel, O., R. Gilad-Bachrach, O. Shamir, and L. Xiao (2012). Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research 13*(1).

Diakonikolas, I., G. Kamath, D. Kane, J. Li, A. Moitra, and A. Stewart (2019). Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing 48*(2), 742–864.

Diakonikolas, I., G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart (2017). Being robust (in high dimensions) can be practical. In *International Conference on Machine Learning*, pp. 999–1008. PMLR.

Donoho, D. and A. Montanari (2016). High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields 166*(3), 935–969.

Gandhi, R. and A. Rajgor (2017). Updating singular value decomposition for rank one matrix perturbation. *arXiv preprint arXiv:1707.08369*.

Gimpel, K., D. Das, and N. A. Smith (2010). Distributed asynchronous online learning for natural language processing. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pp. 213–222.

Golub, G. H. (1973). Some modified matrix eigenvalue problems. *Siam Review 15*(2), 318–334.

Hoeffding, W. (1992). A class of statistics with asymptotically normal distribution. In *Breakthroughs in statistics*, pp. 308–334. Springer.

Huber, P. J. (1992). Robust estimation of a location parameter. In *Breakthroughs in statistics*, pp. 492–518. Springer.

Huber, P. J. (2011). Robust statistics (pp. 1248-1251). *Springer Berlin Heidelberg. Rehabilitation Psychology 46*, 382–399.

Khirirat, S., H. R. Feyzmahdavian, and M. Johansson (2017). Mini-batch gradient descent: Faster convergence under data sparsity. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pp. 2880–2887. IEEE.

Kishore Kumar, N. and J. Schneider (2017). Literature survey on low rank approximation of matrices. *Linear and Multilinear Algebra 65*(11), 2212–2244.

Koltchinskii, V. and K. Lounici (2017). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli 23*(1), 110–133.

Lai, K. A., A. B. Rao, and S. Vempala (2016). Agnostic estimation of mean and covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 665–674. IEEE.

Lounici, K. (2014). High-dimensional covariance matrix estimation with missing observations. *Bernoulli 20*(3), 1029–1058.

Maronna, R. A., R. D. Martin, V. J. Yohai, and M. Salibián-Barrera (2019). *Robust statistics: theory and methods (with R)*. John Wiley & Sons.

Mendelson, S. and N. Zhivotovskiy (2020). Robust covariance estimation under L_4-L_2 norm equivalence. *Annals of Statistics 48*(3), 1648–1664.

Minsker, S. (2017). On some extensions of Bernstein's inequality for self-adjoint operators. *Statistics & Probability Letters 127*, 111–119.

Minsker, S. and X. Wei (2020). Robust modifications of U-statistics and applications to covariance estimation problems. *Bernoulli 26*(1), 694–727.

Nesterov, Y. (2003). *Introductory lectures on convex optimization: A basic course*, Volume 87. Springer Science & Business Media.

Nesterov, Y. E. (1983). A method for solving the convex programming problem with convergence rate O (1/kˆ 2). In *Dokl. akad. nauk Sssr*, Volume 269, pp. 543–547.

Nitanda, A. (2014). Stochastic proximal gradient descent with acceleration techniques. *Advances in Neural Information Processing Systems 27*, 1574–1582.

Prasad, A., S. Balakrishnan, and P. Ravikumar (2019). A unified approach to robust mean estimation. *arXiv preprint arXiv:1907.00927*.

Shalev-Shwartz, S., Y. Singer, N. Srebro, and A. Cotter (2011). Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical programming 127*(1), 3–30.

Stange, P. (2008). On the efficient update of the singular value decomposition. In *PAMM: Proceedings in Applied Mathematics and Mechanics*, Volume 8, pp. 10827–10828. Wiley Online Library.

Tseng, P. (2008). On accelerated proximal gradient methods for convex-concave optimization.

submitted to SIAM Journal on Optimization 1.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, 448–485.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, Volume 47. Cambridge university press.

Watson, G. A. (1992). Characterization of the subdifferential of some matrix norms. *Linear algebra and its applications 170*, 33–45.

Supplementary material

A. Proofs ommited in Section 3.1

In this section, we present the proofs that were omitted from the main exposition in Section 3.1. We start by introducing some technical tools that will be useful for the proof.

A.1 Technical tools

First, we have the following useful trace duality inequalities for the Frobenius inner product.

Proposition 2. *For any $A, B \in \mathbb{R}^{d_1 \times d_2}$,*

$$\begin{aligned} |\langle A, B \rangle| &\leq \|A\|_F \|B\|_F, \\ |\langle A, B \rangle| &\leq \|A\|_1 \|B\|. \end{aligned}$$

Next, let L be a linear subspace of $\mathbb{R}^d$ and $L^\perp$ be its orthogonal complement, namely, $L^\perp = \{v \in \mathbb{R}^d : \langle v, u\rangle = 0, \forall u \in L\}$. In what follows, P_L will stand for the orthogonal projection onto L, meaning that $P_L \in \mathbb{R}^{d\times d}$ is such that $P_L^2 = P_L = P_L^T$ and $\mathrm{Im}(P_L) \subseteq L$, where $\mathrm{Im}(P_L)$ represents the image of P_L. Given the spectral decomposition of a real symmetric matrix, we have the following proposition:

Proposition 3. *Let $S \in S^d(\mathbb{R})$ be a real symmetric matrix with spectral decomposition $S = \sum_{j=1}^d \lambda_j u_j u_j^T$, where the eigenvalues satisfy $|\lambda_1| \geq \cdots \geq |\lambda_d| \geq 0$. Denote $L = \mathrm{Im}(S) = \mathrm{span}\{u_j : \lambda_j \neq 0\}$. Then $P_L = \sum_{j:\lambda_j\neq 0} u_j u_j^T$ and $P_{L^\perp} = \sum_{j:\lambda_j=0} u_j u_j^T$.*

Moreover, we will be interested in a linear operator $\mathcal{P}_L : \mathbb{R}^{d\times d} \mapsto \mathbb{R}^{d\times d}$ defined as

$$\mathcal{P}_L(A) := A - P_{L^\perp} A P_{L^\perp}. \tag{A.32}$$

The following lemma provides some results on $\mathcal{P}_L(\cdot)$ that will be useful in our proof.

Lemma 3. *Let L be a linear subspace of $\mathbb{R}^d$ and $A \in S^d(\mathbb{R})$ be an arbitrary real symmetric matrix, then*

1. $\|\mathcal{P}_L(A)\| \leq \|A\|$.

2. $\mathrm{rank}(\mathcal{P}_L(A)) \leq 2\dim(L)$.

Proof. The proof of this lemma follows straightforward from the definition of $\mathcal{P}_L$ and hence omitted here. $\square$

The following proposition characterizes the subdifferential of the convex function $A \mapsto \|A\|_1$.

Proposition 4 (Watson (1992)). *Let $A \in S^d(\mathbb{R})$ be a symmetric matrix and $A = \sum_{j=1}^{\mathrm{rank}(A)} \sigma_j u_j v_j^T$ be the singular value decomposition defined in Proposition ??. Denote $L = \mathrm{span}\{u_1, \ldots, u_r\}$, then*

$$\partial \|A\|_1 = \left\{ \sum_{j:\sigma_j>0} u_j v_j^T + P_{L^\perp} W P_{L^\perp} : \|W\| \leq 1 \right\},$$

where $P_{L^\perp}$ represents the orthogonal projection onto $L^\perp$.

Next, we state some results for the best rank-k approximation. We say that the function $|||\cdot||| : \mathbb{R}^{d_1 \times d_2} \mapsto \mathbb{R}$ is a matrix norm if for any scalar $\alpha \in \mathbb{R}$ and any matrices $A, B \in \mathbb{R}^{d_1 \times d_2}$, the following properties are satisfied:

- $|||\alpha A||| = |\alpha| |||A|||$;
- $|||A + B||| \leq |||A||| + |||B|||$;
- $|||A||| \geq 0$, and $|||A||| = 0$ if and only if $A = 0_{d_1 \times d_2}$.

The operator norm $\|\cdot\|$, the Frobenius norm $\|\cdot\|_F$ and the nuclear norm$\|\cdot\|_1$ introduced in Definition 1 are concrete examples of matrix norms. Given a nonnegative definite matrix Σ, we say that $\Sigma(k)$ is the best rank-k approximation of Σ with respect to the matrix norm $|||\cdot|||$, if

$$\Sigma(k) = \underset{S:\,\mathrm{rank}(S) \leq k}{\mathrm{argmin}} |||S - \Sigma|||.$$

The following theorem characterized the best rank-k approximation.

Theorem 9 (Kishore Kumar and Schneider (2017))**.** *Let $\Sigma \in \mathbb{R}^{d\times d}$ be a nonnegative definite matrix with spectral decomposition $\Sigma = \sum_{j=1}^{d} \lambda_j u_j u_j^T$, where the eigenvalues satisfy $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$. Then the matrix $A := \sum_{j=1}^{k} \lambda_j u_j u_j^T$ is the best rank-k approximation of Σ in both Frobenius norm and operator norm. Consequently, we have that*

$$\min_{S:\,\mathrm{rank}(S)\leq k} \|S-\Sigma\| = \lambda_{k+1}$$

and

$$\min_{S:\,\mathrm{rank}(S)\leq k} \|S-\Sigma\|_F = \sqrt{\sum_{j=k+1}^{d} \lambda_j^2}.$$

The following two corollaries will be used in our proof.

Corollary 3. *Let $\Sigma(k)$ be the best rank-k approximation of Σ in the operator norm. Then $\|\Sigma(k)-\Sigma\| \leq \|\Sigma\| \left(\frac{\mathrm{rk}(\Sigma)}{k} \wedge \sqrt{\frac{\mathrm{rk}(\Sigma)}{k}}\right)$ and $\|\Sigma(k)-\Sigma\|_F \leq \frac{tr\,(\Sigma)^2}{k}$.*

Proof. Let $\lambda_j(A)$ be the j-th largest eigenvalue of a nonnegative definite matrix, then by Theorem 9,$\|\Sigma(k)-\Sigma\| = \lambda_{k+1}(\Sigma)$. Moreover, we have that

$$\lambda_{k+1}(\Sigma) \leq \frac{\sum_{j=1}^{k+1} \lambda_j(\Sigma)}{k+1} \leq \frac{\mathrm{tr}\,(\Sigma)}{k+1} \leq \frac{\mathrm{tr}\,(\Sigma)}{k} = \|\Sigma\| \frac{\mathrm{rk}(\Sigma)}{k}. \tag{A.33}$$

Note that $\lambda_{k+1}(\Sigma) \leq \sqrt{\|\Sigma\|}\sqrt{\lambda_{k+1}(\Sigma)}$. Combining this with the previous display, we get another inequality

$$\lambda_{k+1}(\Sigma) \leq \sqrt{\frac{\|\Sigma\|\, tr(\Sigma)}{k}} = \|\Sigma\| \sqrt{\frac{\mathrm{rk}(\Sigma)}{k}}.$$

So we have that $\|\Sigma(k)-\Sigma\| \leq \|\Sigma\| \left(\frac{\mathrm{rk}(\Sigma)}{k} \wedge \sqrt{\frac{\mathrm{rk}(\Sigma)}{k}}\right)$. To obtain the bound in Frobenius norm, we note that

$$\|\Sigma(k) - \Sigma\|_F^2 = \sum_{j \geq k+1} \lambda_j(\Sigma)^2 \leq (\mathrm{tr}\,(\Sigma))^2 \sum_{j \geq k+1} j^{-2} \leq \frac{\mathrm{tr}\,(\Sigma)^2}{k},$$

where the first inequality follows from (A.33) and the second inequality follows from $\sum_{j \geq k+1} j^{-2} = \sum_{j \geq k+1} \frac{1}{j(j+1)} = \frac{1}{k+1}$. ☐

Remark 6. It is easy to see that $\frac{\mathrm{rk}(\Sigma)}{k} \leq \sqrt{\frac{\mathrm{rk}(\Sigma)}{k}}$ if and only if $\mathrm{rk}(\Sigma) \leq k$, so when Σ has low effective rank, i.e. $\mathrm{rk}(\Sigma) \leq k$, the upper bound becomes

$$\|\Sigma(k) - \Sigma\| \leq \frac{\|\Sigma\|\,\mathrm{rk}(\Sigma)}{k}.$$

Corollary 4. *Let $\Sigma(k)$ be the best rank-k approximation of Σ in the operator norm defined in Theorem 9, and $L(k) := \mathrm{Im}(\Sigma(k))$. Then $\mathcal{P}_{L(k)}(\Sigma) = \Sigma(k)$.*

Proof. By Theorem 9, $\Sigma(k)$ has spectral decomposition $\Sigma(k) = \sum_{j=1}^d \lambda_j u_j u_j^T$ with $\lambda_{k+1} = \cdots = \lambda_d = 0$. Then Proposition 3 implies that $P_{L(k)^\perp} = \sum_{j=k+1}^d u_j u_j^T$. Therefore,

$$\begin{aligned}\mathcal{P}_{L(k)}(\Sigma) &= \Sigma - P_{L(k)^\perp} \Sigma P_{L(k)^\perp} \\ &= \sum_{j=1}^d \lambda_j u_j u_j^T - \sum_{j=k+1}^d \lambda_j u_j u_j^T = \sum_{j=1}^k \lambda_j u_j u_j^T = \Sigma(k).\end{aligned}$$

☐

A.2 Proof of Theorem 2

To simplify the notations, we denote $N := n(n-1)$. Let

$$F(S,\mathbf{U}_{\mathbf{I_n^2}}) = \lambda_1 \left\|S\right\|_1 + \lambda_2 \sum_{i\neq j} \left\|U_{i,j}\right\|_1 + \frac{1}{N}\sum_{i\neq j}\left\|\widetilde{Y}_{i,j}\widetilde{Y}_{i,j}^T - S - \sqrt{N}U_{i,j}\right\|_{\mathrm{F}}^2. \quad (A.34)$$

The function F is convex, and we have that

$$\partial F(S,\mathbf{U}_{\mathbf{I_n^2}}) = \begin{pmatrix} \lambda_1\partial\left\|S\right\|_1 \\ \lambda_2\partial\left\|U_{1,2}\right\|_1 \\ \vdots \\ \lambda_2\partial\left\|U_{n,n-1}\right\|_1 \end{pmatrix} \quad (A.35)$$
$$- \quad \frac{2}{N}\begin{pmatrix} \sum_{i\neq j}(\widetilde{Y}_{i,j}\widetilde{Y}_{i,j}^T - S - \sqrt{N}U_{i,j}) \\ \sqrt{N}(\widetilde{Y}_{1,2}\widetilde{Y}_{1,2}^T - S - \sqrt{N}U_{1,2}) \\ \vdots \\ \sqrt{N}(\widetilde{Y}_{n,n-1}\widetilde{Y}_{n,n-1}^T - S - \sqrt{N}U_{n,n-1}) \end{pmatrix}$$

where $\partial\left\|A\right\|_1$ represents the subdifferential of $\left\|\cdot\right\|_1$ at A. Note that for any symmetric matrices $S, U_{1,2},\ldots,U_{n,n-1}$, the directional derivative of F at the point $\left(\widehat{S}_\lambda,\widehat{U}_{1,2},\ldots,\widehat{U}_{n,n-1}\right)$ in the direction $\left(S-\widehat{S}_\lambda, U_{1,2}-\widehat{U}_{1,2},\ldots,U_{n,n-1}-\widehat{U}_{n,n-1}\right)$ is nonnegative. In particular, we consider an arbitrary S and $U_{1,2} := \widetilde{U}^*_{1,2},\ldots,U_{n,n-1} := \widetilde{U}^*_{n,n-1}$. By the necessary condition of the minima, there exist $\widehat{V}\in\partial\|\widehat{S}_\lambda\|_1,\ \widehat{W}_{1,2}\in\partial\|\widehat{U}_{1,2}\|_1,\ldots,\widehat{W}_{n,n-1}\in\partial\|\widehat{U}_{n,n-1}\|_1$ such that

$$\begin{aligned} &\left\langle \partial F(\widehat{S},\widehat{\mathbf{U}}_{\mathbf{I_n^2}}),(S-\widehat{S}_\lambda;\widetilde{\mathbf{U}}^*_{\mathbf{I_n^2}}-\widehat{\mathbf{U}}_{\mathbf{I_n^2}})\right\rangle \\ = \quad & -\frac{2}{N}\sum_{i\neq j}\left\langle \widetilde{Y}_{i,j}\widetilde{Y}_{i,j}^T-\widehat{S}_\lambda-\sqrt{N}\widehat{U}_{i,j}, S-\widehat{S}_\lambda+\sqrt{N}(\widetilde{U}^*_{i,j}-\widehat{U}_{i,j})\right\rangle \\ + \quad & \lambda_1\left\langle \widehat{V}, S-\widehat{S}_\lambda\right\rangle + \lambda_2\sum_{i\neq j}\left\langle \widehat{W}_{i,j},\widetilde{U}^*_{i,j}-\widehat{U}_{i,j}\right\rangle \geq 0. \end{aligned}$$

For any choice of $V \in \partial \|S\|_1$, $W_{1,2} \in \partial \left\|\widetilde{U}^*_{1,2}\right\|_1, \ldots, W_{n,n-1} \in \partial \left\|\widetilde{U}^*_{n,n-1}\right\|_1$, by the monotonicity of subgradients we deduce that

$$\left\langle V - \widehat{V}, S - \widehat{S}_\lambda \right\rangle \geq 0,$$

$$\left\langle W_{i,j} - \widehat{W}_{i,j}, \widetilde{U}^*_{i,j} - \widehat{U}_{i,j} \right\rangle \geq 0, \quad (i,j) \in I_n^2.$$

Hence the previous display implies that

$$\begin{aligned} &\tfrac{2}{N} \sum_{i \neq j} \left\langle \widetilde{Y}_{i,j}\widetilde{Y}_{i,j}^T - \widehat{S}_\lambda - \sqrt{N}\widehat{U}_{i,j}, S - \widehat{S}_\lambda + \sqrt{N}(\widetilde{U}^*_{i,j} - \widehat{U}_{i,j}) \right\rangle \\ \leq\ &\lambda_1 \left\langle V, S - \widehat{S}_\lambda \right\rangle + \lambda_2 \sum_{i \neq j} \left\langle W_{i,j}, \widetilde{U}^*_{i,j} - \widehat{U}_{i,j} \right\rangle, \end{aligned}$$

which is equivalent to

$$\begin{aligned} \frac{2}{N} \sum_{i \neq j} &\left\langle \Sigma - \widehat{S}_\lambda + \sqrt{N}(\widetilde{U}^*_{i,j} - \widehat{U}_{i,j}), S - \widehat{S}_\lambda + \sqrt{N}(\widetilde{U}^*_{i,j} - \widehat{U}_{i,j}) \right\rangle \\ &\leq -\lambda_1 \left\langle V, \widehat{S}_\lambda - S \right\rangle - \lambda_2 \sum_{i \neq j} \left\langle W_{i,j}, \widehat{U}_{i,j} - \widetilde{U}^*_{i,j} \right\rangle \\ &- 2 \left\langle \frac{1}{N} \sum_{i \neq j} \widetilde{X}_{i,j} - \Sigma, S - \widehat{S}_\lambda \right\rangle - \frac{2}{\sqrt{N}} \sum_{i \neq j} \left\langle \widetilde{X}_{i,j} - \Sigma, \widetilde{U}^*_{i,j} - \widehat{U}_{i,j} \right\rangle, \quad \text{(A.36)} \end{aligned}$$

where $\widetilde{X}_{i,j} = \widetilde{Z}_{i,j}\widehat{Z}_{i,j}^T$. We will bound (A.36) in two cases.

Case 1: Assume that

$$\frac{2}{N} \sum_{i \neq j} \Big\langle \Sigma - \widehat{S}_\lambda + \sqrt{N}(\widetilde{U}^*_{i,j} - \widehat{U}_{i,j}),$$

$$S - \widehat{S}_\lambda + \sqrt{N}(\widetilde{U}^*_{i,j} - \widehat{U}_{i,j}) \geq 0.$$

Applying the law of cosines, $2\langle A, B\rangle = \|A\|_F^2 + \|B\|_F^2 - \|A-B\|_F^2$, $\forall A, B \in \mathbb{R}^{d\times d}$, to the left hand side of (A.36), we get that

$$\begin{aligned} & \frac{1}{N}\sum_{i\neq j}\left\|\Sigma - \widehat{S}_\lambda + \sqrt{N}(\widetilde{U}^*_{i,j} - \widehat{U}_{i,j})\right\|_F^2 \\ & +\frac{1}{N}\sum_{i\neq j}\left\|S - \widehat{S}_\lambda + \sqrt{N}(\widetilde{U}^*_{i,j} - \widehat{U}_{i,j})\right\|_F^2 \\ \leq\ & \|\Sigma - S\|_F^2 + 2\left\langle \frac{1}{N}\sum_{i\neq j}\widetilde{X}_{i,j} - \Sigma, \widehat{S}_\lambda - S\right\rangle \\ & +\frac{2}{\sqrt{N}}\sum_{i\neq j}\left\langle \widetilde{X}_{i,j} - \Sigma, \widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\rangle \\ & -\lambda_1\left\langle V, \widehat{S}_\lambda - S\right\rangle - \lambda_2\sum_{i\neq j}\left\langle W_{i,j}, \widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\rangle. \end{aligned} \tag{A.37}$$

We will now analyze the terms on the right-hand side of equation (A.37) one by one. First, let $S = \sum_{j=1}^{\text{rank}(S)} \sigma_j(S) u_j v_j^T$ be the singular value decomposition of S, where $\sigma_j(S)$ is the j-th largest singular value of S. Then we can represent any $V \in \partial\|S\|_1$ by $V = \sum_{j=1}^{\text{rank}(S)} u_j v_j^T + P_{L^\perp} W P_{L^\perp}$ for some $\|W\| \leq 1$, where $L = \text{span}\{u_1, \ldots, u_{\text{rank}(S)}\}$. From this representation, we have that $\mathcal{P}_L(V) = V - P_{L^\perp} V P_{L^\perp} = \sum_{j=1}^{\text{rank}(S)} u_j v_j^T$, and

$$\begin{aligned} & -\left\langle V, \widehat{S}_\lambda - S\right\rangle \\ =\ & -\left\langle \mathcal{P}_L(V), \widehat{S}_\lambda - S\right\rangle - \left\langle P_{L^\perp} V P_{L^\perp}, \widehat{S}_\lambda - S\right\rangle \\ =\ & -\left\langle \mathcal{P}_L(V), \widehat{S}_\lambda - S\right\rangle - \left\langle W, P_{L^\perp}\widehat{S}_\lambda P_{L^\perp}\right\rangle \\ \leq\ & \left|\left\langle \mathcal{P}_L(V), \widehat{S}_\lambda - S\right\rangle\right| - \left\|P_{L^\perp}\widehat{S}_\lambda P_{L^\perp}\right\|_1 \\ \leq\ & \|\mathcal{P}_L(V)\|_F \left\|\widehat{S}_\lambda - S\right\|_F - \left\|P_{L^\perp}\widehat{S}_\lambda P_{L^\perp}\right\|_1 \\ =\ & \sqrt{\text{rank}(S)}\left\|\widehat{S}_\lambda - S\right\|_F - \left\|P_{L^\perp}\widehat{S}_\lambda P_{L^\perp}\right\|_1, \end{aligned} \tag{A.38}$$

where we chose W such that $\left\langle W, P_{L^\perp}\widehat{S}_\lambda P_{L^\perp}\right\rangle = \left\|P_{L^\perp}\widehat{S}_\lambda P_{L^\perp}\right\|_1$. Similarly, let $L_{i,j}$ be the image of $\widetilde{U}^*_{i,j}, (i,j) \in I_n^2$, then for properly chosen $W_{1,2} \in \partial\left\|\widetilde{U}^*_{1,2}\right\|_1, \ldots, W_{n,n-1} \in \partial\left\|\widetilde{U}^*_{n,n-1}\right\|_1$, we have that

$$\begin{aligned}
-\sum_{i\neq j}&\left\langle W_{i,j}, \widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\rangle \\
&\leq -\sum_{i\neq j}\left\|P_{L^\perp_{i,j}}\widehat{U}_{i,j}P_{L^\perp_{i,j}}\right\|_1 + \sum_{i\neq j}\left|\left\langle \mathcal{P}_{L_{i,j}}(W_{i,j}), \widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\rangle\right| \\
&\leq \sum_{i\neq j}\sqrt{\operatorname{rank}(\widetilde{U}^*_{i,j})}\left\|\widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\|_F - \sum_{i\neq j}\left\|P_{L^\perp_{i,j}}\widehat{U}_{i,j}P_{L^\perp_{i,j}}\right\|_1 \\
&\leq \sum_{(i,j)\in\widetilde{J}}\sqrt{2}\left\|\widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\|_F - \sum_{i\neq j}\left\|P_{L^\perp_{i,j}}\widehat{U}_{i,j}P_{L^\perp_{i,j}}\right\|_1, \qquad \text{(A.39)}
\end{aligned}$$

where we used the fact that $\operatorname{rank}(\widetilde{U}^*_{i,j}) \leq 2$, and $\widetilde{U}^*_{i,j} = 0$, $L_{i,j} = \{0\}$ for $(i,j) \notin \widetilde{J}$. Next, we denote $\Delta := \frac{1}{N}\sum_{i\neq j}\widetilde{X}_{i,j} - \Sigma$ and recall the linear operator defined in (A.32):

$$\mathcal{P}(A) = A - P_{L^\perp}AP_{L^\perp}.$$

It is easy to check that $\mathcal{P}_L(\Delta) = P_{L^\perp}\Delta P_L + P_L\Delta$, which implies $\operatorname{rank}(\mathcal{P}_L(\Delta)) \leq 2\operatorname{rank}(S)$. Therefore,

$$\begin{aligned}
\left\langle \Delta, \widehat{S}_\lambda - S\right\rangle &= \left\langle \mathcal{P}_L(\Delta), \widehat{S}_\lambda - S\right\rangle + \left\langle P_{L^\perp}\Delta P_{L^\perp}, \widehat{S}_\lambda - S\right\rangle \\
&= \left\langle \mathcal{P}_L(\Delta), \widehat{S}_\lambda - S\right\rangle + \left\langle \Delta, P_{L^\perp}(\widehat{S}_\lambda - S)P_{L^\perp}\right\rangle \\
&\leq \|\mathcal{P}_L(\Delta)\|_F\left\|\widehat{S}_\lambda - S\right\|_F + \|\Delta\|\left\|P_{L^\perp}\widehat{S}_\lambda P_{L^\perp}\right\|_1 \\
&\leq \sqrt{\operatorname{rank}(\mathcal{P}_L(\Delta))}\,\|\mathcal{P}_L(\Delta)\|\left\|\widehat{S}_\lambda - S\right\|_F + \|\Delta\|\left\|P_{L^\perp}\widehat{S}_\lambda P_{L^\perp}\right\|_1 \\
&\leq \sqrt{2\operatorname{rank}(S)}\,\|\Delta\|\left\|\widehat{S}_\lambda - S\right\|_F + \|\Delta\|\left\|P_{L^\perp}\widehat{S}_\lambda P_{L^\perp}\right\|_1, \qquad \text{(A.40)}
\end{aligned}$$

where the last inequality follows from the bound $\|\mathcal{P}_L(\Delta)\| \leq \|\Delta\|$. Finally, it is easy to see that

$$
\begin{aligned}
\sum_{i\neq j}\left\langle \widetilde{X}_{i,j}-\Sigma,\widehat{U}_{i,j}-\widetilde{U}^*_{i,j}\right\rangle &= \sum_{i\neq j}\left\langle \mathcal{P}_{L_{i,j}}(\widetilde{X}_{i,j}-\Sigma),\widehat{U}_{i,j}-\widetilde{U}^*_{i,j}\right\rangle \\
&\quad+\sum_{i\neq j}\left\langle P_{L^{\perp}_{i,j}}(\widetilde{X}_{i,j}-\Sigma)P_{L^{\perp}_{i,j}},\widehat{U}_{i,j}-\widetilde{U}^*_{i,j}\right\rangle \\
&\le \sum_{i\neq j}\left\|\mathcal{P}_{L_{i,j}}(\widetilde{X}_{i,j}-\Sigma)\right\|_F\left\|\widehat{U}_{i,j}-\widetilde{U}^*_{i,j}\right\|_F \\
&\quad+\sum_{i\neq j}\left\|\widetilde{X}_{i,j}-\Sigma\right\|\left\|P_{L^{\perp}_{i,j}}\widehat{U}_{i,j}P_{L^{\perp}_{i,j}}\right\|_1 \\
&\le \sum_{(i,j)\in\widetilde{J}}\sqrt{2\,\mathrm{rank}(\widetilde{U}^*_{i,j})}\left\|\widetilde{X}_{i,j}-\Sigma\right\|\left\|\widehat{U}_{i,j}-\widetilde{U}^*_{i,j}\right\|_F \\
&\quad+\sum_{i\neq j}\left\|\widetilde{X}_{i,j}-\Sigma\right\|\left\|P_{L^{\perp}_{i,j}}\widehat{U}_{i,j}P_{L^{\perp}_{i,j}}\right\|_1 \\
&\le \sum_{(i,j)\in\widetilde{J}}\sqrt{4}\left\|\widetilde{X}_{i,j}-\Sigma\right\|\left\|\widehat{U}_{i,j}-\widetilde{U}^*_{i,j}\right\|_F \\
&\quad+\sum_{i\neq j}\left\|\widetilde{X}_{i,j}-\Sigma\right\|\left\|P_{L^{\perp}_{i,j}}\widehat{U}_{i,j}P_{L^{\perp}_{i,j}}\right\|_1. \qquad \text{(A.41)}
\end{aligned}
$$

Combining inequalities (A.38, A.39, A.40, A.41) with (A.37), we deduce that

$$
\begin{aligned}
&\frac{1}{N}\sum_{i\neq j}\left\|\Sigma-\widehat{S}_\lambda+\sqrt{N}(\widetilde{U}^*_{i,j}-\widehat{U}_{i,j})\right\|_F^2 \\
&\quad+\frac{1}{N}\sum_{i\neq j}\left\|S-\widehat{S}_\lambda+\sqrt{N}(\widetilde{U}^*_{i,j}-\widehat{U}_{i,j})\right\|_F^2 \\
&\le \|\Sigma-S\|_F^2+2\left(\sqrt{2\,\mathrm{rank}(S)}\,\|\Delta\|\left\|\widehat{S}_\lambda-S\right\|_F+\|\Delta\|\left\|P_{L^\perp}\widehat{S}_\lambda P_{L^\perp}\right\|_1\right) \\
&\quad+\frac{2}{\sqrt{N}}\Bigg(\sum_{(i,j)\in\widetilde{J}}\sqrt{4}\left\|\widetilde{X}_{i,j}-\Sigma\right\|\left\|\widehat{U}_{i,j}-\widetilde{U}^*_{i,j}\right\|_F \\
&\quad+\sum_{i\neq j}\left\|\widetilde{X}_{i,j}-\Sigma\right\|\left\|P_{L^{\perp}_{i,j}}\widehat{U}_{i,j}P_{L^{\perp}_{i,j}}\right\|_1\Bigg)+\lambda_1\Big(\sqrt{\mathrm{rank}(S)}\left\|\widehat{S}_\lambda-S\right\|_F \\
&\quad-\left\|P_{L^\perp}\widehat{S}_\lambda P_{L^\perp}\right\|_1\Big) \\
&\quad+\lambda_2\left(\sum_{(i,j)\in\widetilde{J}}\sqrt{2}\left\|\widehat{U}_{i,j}-\widetilde{U}^*_{i,j}\right\|_F-\sum_{i\neq j}\left\|P_{L^{\perp}_{i,j}}\widehat{U}_{i,j}P_{L^{\perp}_{i,j}}\right\|_1\right),
\end{aligned}
$$

which is equivalent to

$$\frac{1}{N}\sum_{i\neq j}\left\|\Sigma-\widehat{S}_\lambda+\sqrt{N}(\widetilde{U}^*_{i,j}-\widehat{U}_{i,j})\right\|_F^2+(\lambda_1-2\left\|\Delta\right\|)\left\|P_{L^\perp}\widehat{S}_\lambda P_{L^\perp}\right\|_1$$
$$+\frac{1}{N}\sum_{i\neq j}\left\|S-\widehat{S}_\lambda+\sqrt{N}(\widetilde{U}^*_{i,j}-\widehat{U}_{i,j})\right\|_F^2$$
$$+\left(\lambda_2-\frac{2}{\sqrt{N}}\max_{i\neq j}\left\|\widetilde{X}_{i,j}-\Sigma\right\|\right)\left\|P_{L^\perp_{i,j}}\widehat{U}_{i,j}P_{L^\perp_{i,j}}\right\|_1$$
$$\leq\left\|\Sigma-S\right\|_F^2+\left(2\sqrt{2\operatorname{rank}(S)}\left\|\Delta\right\|+\lambda_1\sqrt{\operatorname{rank}(S)}\right)\left\|\widehat{S}_\lambda-S\right\|_F$$
$$+\sum_{(i,j)\in\widetilde{J}}\left(\frac{4}{\sqrt{N}}\left\|\widetilde{X}_{i,j}-\Sigma\right\|+\lambda_2\sqrt{2}\right)\left\|\widehat{U}_{i,j}-\widetilde{U}^*_{i,j}\right\|_F. \quad \text{(A.42)}$$

Now consider the event

$$\mathcal{E}_1:=\left\{\lambda_1\geq 2\left\|\Delta\right\|,\lambda_2\geq\frac{3}{\sqrt{N}}\max_{i\neq j}\left\|\widetilde{X}_{i,j}-\Sigma\right\|\right\}.$$

We will derive a bound for $\left\|\widehat{S}_\lambda-\Sigma\right\|_F^2+\sum_{i\neq j}\left\|\widehat{U}_{i,j}-\widetilde{U}^*_{i,j}\right\|_F^2$ on $\mathcal{E}_1$. Applying the identity $\left\|A+B\right\|_F^2=\left\|A\right\|_F^2+\left\|B\right\|_F^2+2\left\langle A,B\right\rangle$ to the the left-hand side of (A.42), we get that on the event $\mathcal{E}_1$,

$$\left\|\Sigma-\widehat{S}_\lambda\right\|_F^2+\left\|S-\widehat{S}_\lambda\right\|_F^2+2\sum_{i\neq j}\left\|\widetilde{U}^*_{i,j}-\widehat{U}_{i,j}\right\|_F^2$$
$$\leq\left\|\Sigma-S\right\|_F^2+\left(2\sqrt{2\operatorname{rank}(S)}\left\|\Delta\right\|+\lambda_1\sqrt{\operatorname{rank}(S)}\right)\left\|\widehat{S}_\lambda-S\right\|_F$$
$$+\sum_{(i,j)\in\widetilde{J}}\left(\frac{4}{\sqrt{N}}\left\|\widetilde{X}_{ij}-\Sigma\right\|+\lambda_2\sqrt{2}\right)\left\|\widehat{U}_{i,j}-\widetilde{U}^*_{i,j}\right\|_F$$
$$+\frac{2}{\sqrt{N}}\sum_{i\neq j}\left\langle\Sigma-\widehat{S}_\lambda,\widetilde{U}^*_{i,j}-\widehat{U}_{i,j}\right\rangle+\frac{2}{\sqrt{N}}\sum_{i\neq j}\left\langle S-\widehat{S}_\lambda,\widetilde{U}^*_{i,j}-\widehat{U}_{i,j}\right\rangle. \quad \text{(A.43)}$$

We now bound the inner product terms on the right-hand side. First, combining inequalities (A.38, A.39, A.40, A.41) with (A.36), we deduce the following bound:

$$\frac{2}{N}\sum_{i\neq j}\left\langle\Sigma-\widehat{S}_\lambda+\sqrt{N}(\widetilde{U}^*_{i,j}-\widehat{U}_{i,j}),S-\widehat{S}_\lambda+\sqrt{N}(\widetilde{U}^*_{i,j}-\widehat{U}_{i,j})\right\rangle$$

$$+\left(\lambda_2-\frac{2}{\sqrt{N}}\max_{i\neq j}\left\|\widetilde{X}_{i,j}-\Sigma\right\|\right)\left\|P_{L_{i,j}^{\perp}}\widehat{U}_{i,j}P_{L_{i,j}^{\perp}}\right\|_1$$

$$+\left(\lambda_1-2\left\|\Delta\right\|\right)\left\|P_{L^{\perp}}\widehat{S}_{\lambda}P_{L^{\perp}}\right\|_1$$

$$\leq\left(2\sqrt{2\,\mathrm{rank}(S)}\left\|\Delta\right\|+\lambda_1\sqrt{\mathrm{rank}(S)}\right)\left\|\widehat{S}_{\lambda}-S\right\|_F$$

$$+\sum_{(i,j)\in\widetilde{J}}\left(\frac{4}{\sqrt{N}}\left\|\widetilde{X}_{i,j}-\Sigma\right\|+\lambda_2\sqrt{2}\right)\left\|\widehat{U}_{i,j}-\widetilde{U}^*_{i,j}\right\|_F. \quad \text{(A.44)}$$

On the event $\mathcal{E}_1$ along with the assumption that

$$\frac{2}{N}\sum_{i\neq j}\Big\langle\Sigma-\widehat{S}_{\lambda}+\sqrt{N}(\widetilde{U}^*_{i,j}-\widehat{U}_{i,j}),$$

$$S-\widehat{S}_{\lambda}+\sqrt{N}(\widetilde{U}^*_{i,j}-\widehat{U}_{i,j})\geq 0.$$

(A.44) implies that

$$\frac{1}{3}\lambda_2\left\|P_{L_{i,j}^{\perp}}\widehat{U}_{i,j}P_{L_{i,j}^{\perp}}\right\|_1$$

$$\leq\quad\left(2\sqrt{2\,\mathrm{rank}(S)}\left\|\Delta\right\|+\lambda_1\sqrt{\mathrm{rank}(S)}\right)\left\|\widehat{S}_{\lambda}-S\right\|_F$$

$$+\quad\sum_{(i,j)\in\widetilde{J}}\left(\frac{4}{\sqrt{N}}\left\|\widetilde{X}_{i,j}-\Sigma\right\|+\lambda_2\sqrt{2}\right)\left\|\widehat{U}_{i,j}-\widetilde{U}^*_{i,j}\right\|_F.$$

Recall that $L_{i,j}=\{0\}$ for any $(i,j)\notin\widetilde{J}$, hence

$$\lambda_2\sum_{(i,j)\notin\widetilde{J}}\left\|P_{L_{i,j}^{\perp}}\widehat{U}_{i,j}P_{L_{i,j}^{\perp}}\right\|_1 \quad \text{(A.45)}$$

$$=\quad\lambda_2\sum_{(i,j)\notin\widetilde{J}}\left\|P_{L_{i,j}^{\perp}}\left(\widehat{U}_{i,j}-\widetilde{U}^*_{i,j}\right)P_{L_{i,j}^{\perp}}\right\|_1$$

$$=\quad\lambda_2\sum_{(i,j)\notin\widetilde{J}}\left\|\widehat{U}_{i,j}-\widetilde{U}^*_{i,j}\right\|_1$$

$$\leq\quad 3\left(2\sqrt{2\,\mathrm{rank}(S)}\left\|\Delta\right\|+\lambda_1\sqrt{\mathrm{rank}(S)}\right)\left\|\widehat{S}_{\lambda}-S\right\|_F$$

$$+\quad 3\sum_{(i,j)\in\widetilde{J}}\left(\frac{4}{\sqrt{N}}\left\|\widetilde{X}_{i,j}-\Sigma\right\|+\lambda_2\sqrt{2}\right)\left\|\widehat{U}_{i,j}-\widetilde{U}^*_{i,j}\right\|_F.$$

Next, we can estimate $\sum_{(i,j)\notin \widetilde{J}} \left|\left\langle \widehat{S}_\lambda - S, \widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\rangle\right|$ as follows:

$$\begin{aligned}
&\sum_{(i,j)\notin \widetilde{J}} \left|\left\langle \widehat{S}_\lambda - S, \widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\rangle\right| \leq \left\|\widehat{S}_\lambda - S\right\| \sum_{(i,j)\notin \widetilde{J}} \left\|\widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\|_1 \\
&\leq \frac{3\left\|\widehat{S}_\lambda - S\right\|}{\lambda_2}\Bigg[\left(2\sqrt{2\,\mathrm{rank}(S)}\,\|\Delta\| + \lambda_1\sqrt{\mathrm{rank}(S)}\right)\left\|\widehat{S}_\lambda - S\right\|_F \\
&\qquad + \sum_{(i,j)\in \widetilde{J}} \left(\frac{4}{\sqrt{N}}\left\|\widetilde{X}_{i,j} - \Sigma\right\| + \lambda_2\sqrt{2}\right)\left\|\widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\|_F\Bigg] \\
&\leq 3\left\|\widehat{S}_\lambda - S\right\|\left[(\sqrt{2}+1)\sqrt{\mathrm{rank}(S)}\frac{\lambda_1}{\lambda_2}\left\|\widehat{S}_\lambda - S\right\|_F + \left(\frac{4}{3}+\sqrt{2}\right)\sum_{(i,j)\in \widetilde{J}}\left\|\widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\|_F\right],
\end{aligned} \tag{A.46}$$

where the last inequality holds on event $\mathcal{E}_1$. This implies that

$$\begin{aligned}
&\frac{2}{\sqrt{N}}\sum_{i\neq j}\left\langle S - \widehat{S}_\lambda, \widetilde{U}^*_{i,j} - \widehat{U}_{i,j}\right\rangle \leq \frac{2}{\sqrt{N}}\sum_{(i,j)\in \widetilde{J}}\left\langle S - \widehat{S}_\lambda, \widetilde{U}^*_{i,j} - \widehat{U}_{i,j}\right\rangle + \frac{6}{\sqrt{N}}\left\|\widehat{S}_\lambda - S\right\| \\
&\qquad \times\left[(\sqrt{2}+1)\sqrt{\mathrm{rank}(S)}\frac{\lambda_1}{\lambda_2}\left\|\widehat{S}_\lambda - S\right\|_F + \left(\frac{4}{3}+\sqrt{2}\right)\sum_{(i,j)\in \widetilde{J}}\left\|\widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\|_F\right] \\
&\leq 2\cdot\frac{\left\|\widehat{S}_\lambda - S\right\|_F}{2}\cdot 2\frac{1}{\sqrt{N}}\sqrt{\left|\widetilde{J}\right|}\sqrt{\sum_{(i,j)\in \widetilde{J}}\left\|\widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\|_F^2} + (6\sqrt{2}+6)\frac{\lambda_1}{\lambda_2}\sqrt{\frac{\mathrm{rank}(S)}{N}}\left\|\widehat{S}_\lambda - S\right\|_F^2 \\
&\qquad + 2\cdot\frac{\left\|\widehat{S}_\lambda - S\right\|_F}{2}\cdot(8+6\sqrt{2})\frac{1}{\sqrt{N}}\sqrt{\left|\widetilde{J}\right|}\sqrt{\sum_{(i,j)\in \widetilde{J}}\left\|\widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\|_F^2} \\
&\leq \left[(6\sqrt{2}+6)\sqrt{\frac{\mathrm{rank}(S)}{N}}\frac{\lambda_1}{\lambda_2} + \frac{1}{2}\right]\left\|\widehat{S}_\lambda - S\right\|_F^2 + \left[4 + (6\sqrt{2}+8)^2\right]\frac{|\widetilde{J}|}{N}\sum_{(i,j)\in \widetilde{J}}\left\|\widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\|_F^2,
\end{aligned} \tag{A.47}$$

where the second inequality follows from the fact that $\|A\| \leq \|A\|_F$ for any symmetric matrix A, and the last inequality follows from the fact that $2ab \leq a^2 + b^2$ for any real numbers a, b.

Similarly, we deduce that

$$\frac{2}{\sqrt{N}}\sum_{i\neq j}\left\langle \Sigma-\widehat{S}_\lambda, \widetilde{U}^*_{i,j}-\widehat{U}_{i,j}\right\rangle \leq (6\sqrt{2}+6)\sqrt{\frac{\mathrm{rank}(S)}{N}}\frac{\lambda_1}{\lambda_2}\left\|\widehat{S}_\lambda - S\right\|_F^2$$
$$+\left[(6\sqrt{2}+6)\sqrt{\frac{\mathrm{rank}(S)}{N}}\frac{\lambda_1}{\lambda_2}+\frac{1}{2}\right]\left\|\widehat{S}_\lambda-\Sigma\right\|_F^2+\left[4+(6\sqrt{2}+8)^2\right]\frac{|\widetilde{J}|}{N}\sum_{(i,j)\in\widetilde{J}}\left\|\widehat{U}_{i,j}-\widetilde{U}^*_{i,j}\right\|_F^2. \tag{A.48}$$

Combining (A.47, A.48) with (A.43), one sees that on event $\mathcal{E}_1$,

$$\left\|\Sigma-\widehat{S}_\lambda\right\|_F^2+\left\|S-\widehat{S}_\lambda\right\|_F^2+2\sum_{i\neq j}\left\|\widetilde{U}^*_{i,j}-\widehat{U}_{i,j}\right\|_F^2$$
$$\leq \|\Sigma - S\|_F^2 + \left(2\sqrt{2\,\mathrm{rank}(S)}\,\|\Delta\|+\lambda_1\sqrt{\mathrm{rank}(S)}\right)\left\|\widehat{S}_\lambda - S\right\|_F$$
$$+\sum_{(i,j)\in\widetilde{J}}\left(\frac{4}{\sqrt{N}}\left\|\widetilde{X}_{i,j}-\Sigma\right\|+\lambda_2\sqrt{2}\right)\left\|\widehat{U}_{i,j}-\widetilde{U}^*_{i,j}\right\|_F+2\left[4+(6\sqrt{2}+8)^2\right]\frac{|\widetilde{J}|}{N}\sum_{(i,j)\in\widetilde{J}}\left\|\widehat{U}_{i,j}-\widetilde{U}^*_{i,j}\right\|_F^2$$
$$+\left[2(6\sqrt{2}+6)\sqrt{\frac{\mathrm{rank}(S)}{N}}\frac{\lambda_1}{\lambda_2}+\frac{1}{2}\right]\left\|\widehat{S}_\lambda-S\right\|_F^2+\left[(6\sqrt{2}+6)\sqrt{\frac{\mathrm{rank}(S)}{N}}\frac{\lambda_1}{\lambda_2}+\frac{1}{2}\right]\left\|\widehat{S}_\lambda-\Sigma\right\|_F^2$$
$$\leq \|\Sigma - S\|_F^2 + 2\lambda_1\left(\sqrt{2}+1\right)\sqrt{\mathrm{rank}(S)}\cdot\frac{1}{2}\left\|\widehat{S}_\lambda - S\right\|_F$$
$$+2\frac{1}{\sqrt{2}}\lambda_2(\frac{4}{3}+\sqrt{2})\sqrt{|\widetilde{J}|}\cdot\frac{1}{\sqrt{2}}\sqrt{\sum_{(i,j)\in\widetilde{J}}\left\|\widehat{U}_{i,j}-\widetilde{U}^*_{i,j}\right\|_F^2}$$
$$+\left[2(6\sqrt{2}+6)\sqrt{\frac{\mathrm{rank}(S)}{N}}\frac{\lambda_1}{\lambda_2}+\frac{1}{2}\right]\left\|\widehat{S}_\lambda-S\right\|_F^2+\left[(6\sqrt{2}+6)\sqrt{\frac{\mathrm{rank}(S)}{N}}\frac{\lambda_1}{\lambda_2}+\frac{1}{2}\right]\left\|\widehat{S}_\lambda-\Sigma\right\|_F^2$$
$$+2\left[4+(6\sqrt{2}+8)^2\right]\frac{|\widetilde{J}|}{N}\sum_{(i,j)\in\widetilde{J}}\left\|\widehat{U}_{i,j}-\widetilde{U}^*_{i,j}\right\|_F^2$$
$$\leq \|\Sigma - S\|_F^2+\frac{1}{4}\left\|\widehat{S}_\lambda - S\right\|_F^2+\frac{1}{2}\sum_{i\neq j}\left\|\widetilde{U}^*_{i,j}-\widehat{U}_{i,j}\right\|_F^2+\lambda_1^2\left(\sqrt{2}+1\right)^2\mathrm{rank}(S)+\frac{1}{2}\lambda_2^2(\frac{4}{3}+\sqrt{2})^2|\widetilde{J}|$$
$$+\left[2(6\sqrt{2}+6)\sqrt{\frac{\mathrm{rank}(S)}{N}}\frac{\lambda_1}{\lambda_2}+\frac{1}{2}\right]\left\|\widehat{S}_\lambda-S\right\|_F^2+\left[(6\sqrt{2}+6)\sqrt{\frac{\mathrm{rank}(S)}{N}}\frac{\lambda_1}{\lambda_2}+\frac{1}{2}\right]\left\|\widehat{S}_\lambda-\Sigma\right\|_F^2$$
$$+2\left[4+(6\sqrt{2}+8)^2\right]\frac{|\widetilde{J}|}{N}\sum_{(i,j)\in\widetilde{J}}\left\|\widehat{U}_{i,j}-\widetilde{U}^*_{i,j}\right\|_F^2. \tag{A.49}$$

Assuming that $2(6+6\sqrt{2})\sqrt{\frac{\mathrm{rank}(S)}{N}}\frac{\lambda_1}{\lambda_2}\leq\frac{1}{8}$ and $2\left[4+(6\sqrt{2}+8)^2\right]\frac{|\widetilde{J}|}{N}\leq\frac{11}{8}$, we conclude that

$$\frac{1}{8}\left(\left\|\Sigma-\widehat{S}_\lambda\right\|_F^2+\sum_{i\neq j}\left\|\widetilde{U}^*_{i,j}-\widehat{U}_{i,j}\right\|_F^2\right)$$

$$\leq\|\Sigma-S\|_F^2+\operatorname{rank}(S)\lambda_1^2\left(\sqrt{2}+1\right)^2+\lambda_2^2\frac{(4/3+\sqrt{2})^2}{2}|\widetilde{J}|. \quad \text{(A.50)}$$

The assumptions above are valid provided $\operatorname{rank}(S)\leq\frac{1}{56000}\cdot\frac{n^2\lambda_2^2}{\lambda_1^2}$ and $|\widetilde{J}|\leq\frac{N}{402}$. Note that if we apply the inequality $2ab\leq a^2+b^2$ in the derivation above with different choices of constants, we can reduce the conditions on $\operatorname{rank}(S)$ and $|\widetilde{J}|$ to

$$\operatorname{rank}(S)\leq c_1\frac{n^2\lambda_2^2}{\lambda_1^2},\qquad\forall c_1\leq\frac{1}{5980} \quad \text{(A.51)}$$

and

$$|\widetilde{J}|\leq c_2N,\qquad\forall c_2\leq\frac{1}{295}.$$

Case 2: Assume that

$$\frac{2}{N}\sum_{i\neq j}\left\langle\Sigma-\widehat{S}_\lambda+\sqrt{N}(\widetilde{U}^*_{i,j}-\widehat{U}_{i,j}),S-\widehat{S}_\lambda+\sqrt{N}(\widetilde{U}^*_{i,j}-\widehat{U}_{i,j})\right\rangle<0.$$

We start with several lemmas.

Lemma 4. *On the event*

$$\mathcal{E}_2:=\left\{\lambda_1\geq4\left\|\frac{1}{N}\sum_{i\neq j}\widetilde{X}_{i,j}-\Sigma(k)\right\|,\lambda_2\geq\frac{4}{\sqrt{N}}\max_{i\neq j}\left\|\widetilde{X}_{i,j}-\Sigma(k)\right\|\right\},$$

the following inequality holds

$$\lambda_1\left\|P_{L(k)^\perp}\widehat{S}_\lambda P_{L(k)^\perp}\right\|_1+\lambda_2\sum_{i\neq j}\left\|P_{L^\perp_{i,j}}\widehat{U}_{i,j}P_{L^\perp_{i,j}}\right\|_1$$

$$\leq 3\left(\lambda_1\left\|\mathcal{P}_{L(k)}(\widehat{S}_\lambda-\Sigma(k))\right\|_1+\lambda_2\sum_{(i,j)\in\widetilde{J}}\left\|\mathcal{P}_{L_{i,j}}(\widehat{U}_{i,j}-\widetilde{U}^*_{i,j})\right\|_1\right),$$

where $L(k)=\operatorname{Im}(\Sigma(k))$, $L_{i,j}=\operatorname{Im}\left(\widetilde{U}^*_{i,j}\right)$, *and* $P_{L(k)}$, $P_{L_{i,j}}$ *are the orthogonal projections onto the corresponding subspaces.*

Proof of Lemma 4. Denote

$$Q(S,U_{1,2},\ldots,U_{n,n-1}):=\frac{1}{N}\sum_{i\neq j}\left\|\widetilde{Y}_{i,j}\widetilde{Y}^T_{i,j}-S-\sqrt{N}U_{i,j}\right\|_F^2.$$

By definition of $\widehat{S}_\lambda$,

$$Q(\widehat{S}_\lambda,\widehat{U}_{1,2},\ldots,\widehat{U}_{n,n-1})-Q(\Sigma(k),\widetilde{U}^*_{1,2},\ldots,\widetilde{U}^*_{n,n-1})$$
$$\leq\lambda_1\left(\|\Sigma(k)\|_1-\|\widehat{S}_\lambda\|_1\right)+\lambda_2\sum_{i\neq j}(\|\widetilde{U}^*_{i,j}\|_1-\|\widehat{U}_{i,j}\|_1). \tag{A.52}$$

By convexity of the $\|\cdot\|_1$ norm, for any $V\in\partial\|\Sigma(k)\|_1$, $\|\Sigma(k)\|_1-\|\widehat{S}_\lambda\|_1\leq\left\langle V,\Sigma(k)-\widehat{S}_\lambda\right\rangle$. Let $r=\operatorname{rank}(\Sigma(k))\leq k$, we have the representation $V=\sum_{j=1}^r v_jv_j^T+P_{L(k)^\perp}WP_{L(k)^\perp}$, where $\|W\|\leq 1$. By duality between the spectral and nuclear norm (Proposition 2), we deduce that with an appropriate choice of W,

$$\|\Sigma(k)\|_1-\|\widehat{S}_\lambda\|_1\leq\left\langle V,\Sigma(k)-\widehat{S}_\lambda\right\rangle$$
$$=\left\langle\mathcal{P}_{L(k)}(V),\Sigma(k)-\widehat{S}_\lambda\right\rangle+\left\langle P_{L(k)^\perp}WP_{L(k)^\perp},\Sigma(k)-\widehat{S}_\lambda\right\rangle$$
$$\leq\left\|\mathcal{P}_{L(k)}(\Sigma(k)-\widehat{S}_\lambda)\right\|_1-\left\|P_{L(k)^\perp}\widehat{S}_\lambda P_{L(k)^\perp}\right\|_1. \tag{A.53}$$

Similalry,

$$\sum_{i\neq j}(\|\widetilde{U}^*_{i,j}\|_1-\|\widehat{U}_{i,j}\|_1)\leq\sum_{i\neq j}\left(\left\|\mathcal{P}_{L_{i,j}}(\widetilde{U}^*_{i,j}-\widehat{U}_{i,j})\right\|_1-\left\|P_{L^\perp_{i,j}}\widehat{U}_{i,j}P_{L^\perp_{i,j}}\right\|_1\right), \tag{A.54}$$

where $L_{i,j}$ is the image of $\widetilde{U}^*_{i,j}$, $\forall (i,j) \in I_n^2$.

On the other hand, recall that $\widetilde{Y}_{i,j}\widetilde{Y}_{i,j}^T = \widetilde{X}_{i,j} + \sqrt{N}\widetilde{U}^*_{i,j}$ and ∇Q is given by the first term in equation (A.35). Convexity of Q implies that

$$\begin{aligned}
&Q(\widehat{S}_\lambda, \widehat{U}_{1,2}, \ldots, \widehat{U}_{n,n-1}) - Q(\Sigma(k), \widetilde{U}^*_{1,2}, \ldots, \widetilde{U}^*_{n,n-1}) \\
&\geq \left\langle \nabla Q\left(\Sigma(k), \widetilde{U}^*_{1,2}, \ldots, \widetilde{U}^*_{n,n-1}\right), (\widehat{S}_\lambda - \Sigma(k), \widehat{U}_{1,2} - \widetilde{U}^*_{1,2}, \ldots, \widehat{U}_{n,n-1} - \widetilde{U}^*_{n,n-1})\right\rangle \\
&= -\frac{2}{N}\sum_{i\neq j}\left\langle \widetilde{X}_{i,j} - \Sigma(k), \widehat{S}_\lambda - \Sigma(k)\right\rangle - \frac{2}{\sqrt{N}}\sum_{i\neq j}\left\langle \widetilde{X}_{i,j} - \Sigma(k), \widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\rangle \\
&= 2\left\langle \Sigma(k) - \frac{1}{N}\sum_{i\neq j}\widetilde{X}_{i,j}, \widehat{S}_\lambda - \Sigma(k)\right\rangle + \frac{2}{\sqrt{N}}\sum_{i\neq j}\left\langle \Sigma(k) - \widetilde{X}_{i,j}, \widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\rangle \\
&\geq -2\left\|\Sigma(k) - \frac{1}{N}\sum_{i\neq j}\widetilde{X}_{i,j}\right\|\left\|\widehat{S}_\lambda - \Sigma(k)\right\|_1 - \frac{2}{\sqrt{N}}\max_{i\neq j}\left\|\Sigma(k) - \widetilde{X}_{i,j}\right\|\sum_{i\neq j}\left\|\widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\|_1.
\end{aligned} \tag{A.55}$$

On the event

$$\mathcal{E}_2 := \left\{\lambda_1 \geq 4\left\|\frac{1}{N}\sum_{i\neq j}\widetilde{X}_{i,j} - \Sigma(k)\right\|, \lambda_2 \geq \frac{4}{\sqrt{N}}\max_{i\neq j}\left\|\widetilde{X}_{i,j} - \Sigma(k)\right\|\right\},$$

the inequality (A.55) implies that

$$Q(\widehat{S}_\lambda, \widehat{U}_{1,2}, \ldots, \widehat{U}_{n,n-1}) - Q(\Sigma(k), \widetilde{U}^*_{1,2}, \ldots, \widetilde{U}^*_{n,n-1}) \tag{A.56}$$

$$\geq -\frac{1}{2}\left(\lambda_1\left\|\widehat{S}_\lambda - \Sigma(k)\right\|_1 + \lambda_2\sum_{i\neq j}\left\|\widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\|_1\right). \tag{A.57}$$

Moreover, note that

$$\left\|\widehat{S}_\lambda - \Sigma(k)\right\|_1 \leq \left\|\mathcal{P}_{L(k)}(\widehat{S}_\lambda - \Sigma(k))\right\|_1 + \left\|P_{L(k)^\perp}\widehat{S}_\lambda P_{L(k)^\perp}\right\|_1$$

and

$$\left\|\widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\|_1 \leq \left\|\mathcal{P}_{L_{i,j}}(\widehat{U}_{i,j} - \widetilde{U}^*_{i,j})\right\|_1 + \left\|P_{L_{i,j}^\perp}\widehat{U}_{i,j}P_{L_{i,j}^\perp}\right\|_1.$$

Combining these inequalities with (A.56), we get the lower bound

$$Q(\widehat{S}_\lambda, \widehat{U}_{1,2}, \ldots, \widehat{U}_{n,n-1}) - Q(\Sigma(k), \widetilde{U}^*_{1,2}, \ldots, \widetilde{U}^*_{n,n-1})$$
$$\geq -\frac{1}{2}\Bigg[\lambda_1 \left(\left\|\mathcal{P}_{L(k)}(\widehat{S}_\lambda - \Sigma(k))\right\|_1 + \left\|P_{L(k)^\perp}\widehat{S}_\lambda P_{L(k)^\perp}\right\|_1\right)$$
$$+ \lambda_2 \left(\left\|\mathcal{P}_{L_{i,j}}(\widehat{U}_{i,j} - \widetilde{U}^*_{i,j})\right\|_1 + \left\|P_{L^\perp_{i,j}}\widehat{U}_{i,j}P_{L^\perp_{i,j}}\right\|_1\right)\Bigg].$$

Combining (A.52, A.53, A.54) with the lower bound (A.58), we deduce the "sparsity inequality"

$$\lambda_1 \left\|P_{L(k)^\perp}\widehat{S}_\lambda P_{L(k)^\perp}\right\|_1 + \lambda_2 \sum_{i\neq j}\left\|P_{L^\perp_{i,j}}\widehat{U}_{i,j}P_{L^\perp_{i,j}}\right\|_1$$
$$\leq 3\left(\lambda_1 \left\|\mathcal{P}_{L(k)}(\widehat{S}_\lambda - \Sigma(k))\right\|_1 + \lambda_2 \sum_{(i,j)\in\widetilde{J}}\left\|\mathcal{P}_{L_{i,j}}(\widehat{U}_{i,j} - \widetilde{U}^*_{i,j})\right\|_1\right).$$

□

Lemma 5. *Assume that* $\max\left\{6\sqrt{2}\cdot\frac{\lambda_1}{\lambda_2}\sqrt{\frac{k}{N}}, 7\sqrt{\frac{|\widetilde{J}|}{N}}\right\} \leq \frac{1}{4}$. *Then on the event* $\mathcal{E}_2$ *of Lemma 4, the following inequality holds:*

$$\left\|\widehat{S}_\lambda - \Sigma\right\|_F^2 + \sum_{i\neq j}\left\|\widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\|_F^2 \leq \frac{2}{N}\sum_{i\neq j}\left\|\widehat{S}_\lambda - \Sigma + \sqrt{N}(\widehat{U}_{i,j} - \widetilde{U}^*_{i,j})\right\|_F^2.$$

Proof of Lemma 5. First, we consider the decomposition

$$\frac{2}{\sqrt{N}}\sum_{i\neq j}\left\langle \widehat{S}_\lambda - \Sigma, \widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\rangle$$
$$= \underbrace{\frac{2}{\sqrt{N}}\sum_{(i,j)\in\widetilde{J}}\left\langle \widehat{S}_\lambda - \Sigma, \widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\rangle}_{\text{I}} + \underbrace{\frac{2}{\sqrt{N}}\sum_{(i,j)\notin\widetilde{J}}\left\langle \widehat{S}_\lambda - \Sigma, \widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\rangle}_{\text{II}}. \quad \text{(A.58)}$$

For the term I, we have that

$$\text{I} \leq \frac{2\left\|\widehat{S}_\lambda - \Sigma\right\|_F}{\sqrt{N}}\sum_{(i,j)\in\widetilde{J}}\left\|\widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\|_F \leq 2\left\|\widehat{S}_\lambda - \Sigma\right\|_F\sqrt{\frac{|\widetilde{J}|}{N}}\sqrt{\sum_{(i,j)\in\widetilde{J}}\left\|\widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\|_F^2}. \quad \text{(A.59)}$$

To estimate the term II, note that $\mathcal{P}_{L(k)}(\Sigma) = \Sigma(k) = \mathcal{P}_{L(k)}(\Sigma(k))$ as $L(k) = \operatorname{Im}(\Sigma(k))$. Moreover, $\sum_{(i,j)\notin \widetilde{J}} \left\| P_{L_{i,j}^{\perp}} \widehat{U}_{i,j} P_{L_{i,j}^{\perp}} \right\|_1 = \sum_{(i,j)\notin \widetilde{J}} \left\| \widehat{U}_{i,j} - \widetilde{U}^*_{i,j} \right\|_1$, hence Lemma 4 yields that on the event $\mathcal{E}_2$,

$$\begin{aligned}
\mathrm{II} &\le \frac{2\left\|\widehat{S}_\lambda - \Sigma\right\|}{\sqrt{N}} \sum_{(i,j)\notin \widetilde{J}} \left\|\widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\|_1 \\
&\le \frac{2\left\|\widehat{S}_\lambda - \Sigma\right\|}{\sqrt{N}} \cdot 3 \left(\frac{\lambda_1}{\lambda_2} \left\| \mathcal{P}_{L(k)}(\widehat{S}_\lambda - \Sigma(k)) \right\|_1 + \sum_{(i,j)\in \widetilde{J}} \left\| \mathcal{P}_{L_{i,j}}(\widehat{U}_{i,j} - \widetilde{U}^*_{i,j}) \right\|_1 \right) \\
&\le \frac{6\left\|\widehat{S}_\lambda - \Sigma\right\|}{\sqrt{N}} \Bigg(\frac{\lambda_1}{\lambda_2} \sqrt{2\,\mathrm{rank}(\Sigma(k))} \left\| \mathcal{P}_{L(k)}(\widehat{S}_\lambda - \Sigma) \right\|_F \\
&\qquad\qquad + \sum_{(i,j)\in \widetilde{J}} \sqrt{2\,\mathrm{rank}(\widetilde{U}^*_{i,j})} \left\| \mathcal{P}_{L_{i,j}}(\widehat{U}_{i,j} - \widetilde{U}^*_{i,j}) \right\|_F \Bigg) \\
&\le \frac{6\left\|\widehat{S}_\lambda - \Sigma\right\|}{\sqrt{N}} \left(\frac{\lambda_1}{\lambda_2}\sqrt{2k} \left\|\widehat{S}_\lambda - \Sigma\right\|_F + \sum_{(i,j)\in \widetilde{J}} \sqrt{4} \left\| \widehat{U}_{i,j} - \widetilde{U}^*_{i,j} \right\|_F \right) \\
&\le 6 \left\|\widehat{S}_\lambda - \Sigma\right\|_F \left(\frac{\lambda_1}{\lambda_2} \sqrt{\frac{2k}{N}} \left\|\widehat{S}_\lambda - \Sigma\right\|_F + 2\sqrt{\frac{|\widetilde{J}|}{N}} \sqrt{\sum_{(i,j)\in \widetilde{J}} \left\| \widehat{U}_{i,j} - \widetilde{U}^*_{i,j} \right\|_F^2} \right). \qquad \text{(A.60)}
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\frac{2}{\sqrt{N}} \sum_{i\neq j} \left\langle \widehat{S}_\lambda - \Sigma, \widehat{U}_{i,j} - \widetilde{U}^*_{i,j} \right\rangle \\
&\ge -6 \left\|\widehat{S}_\lambda - \Sigma\right\|_F^2 \frac{\lambda_1}{\lambda_2}\sqrt{\frac{2k}{N}} - \sqrt{\frac{|\widetilde{J}|}{N}} \cdot 2 \cdot \left(\sqrt{7} \left\|\widehat{S}_\lambda - \Sigma\right\|_F \right) \cdot \left(\sqrt{7} \sqrt{\sum_{(i,j)\in \widetilde{J}} \left\| \widehat{U}_{i,j} - \widetilde{U}^*_{i,j} \right\|_F^2} \right) \\
&\ge -6 \left\|\widehat{S}_\lambda - \Sigma\right\|_F^2 \frac{\lambda_1}{\lambda_2}\sqrt{\frac{2k}{N}} - \sqrt{\frac{|\widetilde{J}|}{N}} \left(7 \left\|\widehat{S}_\lambda - \Sigma\right\|_F^2 + 7 \sum_{(i,j)\in \widetilde{J}} \left\| \widehat{U}_{i,j} - \widetilde{U}^*_{i,j} \right\|_F^2 \right) \\
&\ge -\left(6\sqrt{2}\frac{\lambda_1}{\lambda_2}\sqrt{\frac{k}{N}} + 7\sqrt{\frac{|\widetilde{J}|}{N}} \right) \left\|\widehat{S}_\lambda - \Sigma\right\|_F^2 - 7\sqrt{\frac{|\widetilde{J}|}{N}} \sum_{(i,j)\in \widetilde{J}} \left\| \widehat{U}_{i,j} - \widetilde{U}^*_{i,j} \right\|_F^2, \qquad \text{(A.61)}
\end{aligned}$$

where we used $2ab \le a^2 + b^2$ in the second inequality. Finally, given the assumption that

$$\max\left\{ 6\sqrt{2} \cdot \frac{\lambda_1}{\lambda_2}\sqrt{\frac{k}{N}}, 7\sqrt{\frac{\widetilde{J}}{N}} \right\} \le \frac{1}{4},$$

on the event $\mathcal{E}_2$ we have that

$$\begin{aligned}
&\frac{2}{N}\sum_{i\neq j}\left\|\widehat{S}_\lambda - \Sigma + \sqrt{N}(\widehat{U}_{i,j} - \widetilde{U}^*_{i,j})\right\|_F^2 \\
&= 2\left(\left\|\widehat{S}_\lambda - \Sigma\right\|_F^2 + \sum_{i\neq j}\left\|\widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\|_F^2 + \frac{2}{\sqrt{N}}\sum_{i\neq j}\left\langle \widehat{S}_\lambda - \Sigma, \widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\rangle\right) \\
&\geq 2\left(\left\|\widehat{S}_\lambda - \Sigma\right\|_F^2 + \sum_{i\neq j}\left\|\widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\|_F^2 - \frac{1}{2}\left\|\widehat{S}_\lambda - \Sigma\right\|_F^2 - \frac{1}{4}\sum_{i\neq j}\left\|\widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\|_F^2\right) \\
&\geq \left\|\widehat{S}_\lambda - \Sigma\right\|_F^2 + \sum_{i\neq j}\left\|\widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\|_F^2. \qquad \text{(A.62)}
\end{aligned}$$

□

Remark 7. We now consider the intersection of events $\mathcal{E}_1$ and $\mathcal{E}_2$. Consider $k = \lfloor \frac{N\lambda_2^2}{1200\lambda_1^2} \rfloor$, $\left|\widetilde{J}\right| \leq \frac{N}{6400}$ (implying that $\left|\widetilde{J}\right| \leq \frac{N}{800}$). Corollary 3 guarantees that $\|\Sigma(k) - \Sigma\| \leq \|\Sigma\| \sqrt{\frac{\mathrm{rk}(\Sigma)}{k}}$, so

$$\begin{aligned}
4\left\|\frac{1}{N}\sum_{i\neq j}\widetilde{X}_{i,j} - \Sigma(k)\right\| &\leq 4\left(\left\|\frac{1}{N}\sum_{i\neq j}\widetilde{X}_{ij} - \Sigma\right\| + \|\Sigma - \Sigma(k)\|\right) \\
&\leq 4\left\|\frac{1}{N}\sum_{i\neq j}\widetilde{X}_{i,j} - \Sigma\right\| + 4\|\Sigma\|\sqrt{\frac{\mathrm{rk}(\Sigma)}{k}} \leq 4\left\|\frac{1}{N}\sum_{i\neq j}\widetilde{X}_{i,j} - \Sigma\right\| + 140\|\Sigma\|\sqrt{\frac{\mathrm{rk}(\Sigma)}{N}}.
\end{aligned}$$

Similarly, for the second term we have that

$$\begin{aligned}
\frac{4}{\sqrt{N}}\max_{i\neq j}\left\|\widetilde{X}_{i,j} - \Sigma(k)\right\| &\leq \frac{4}{\sqrt{N}}\max_{i\neq j}\left(\left\|\widetilde{X}_{i,j} - \Sigma\right\| + \|\Sigma(k) - \Sigma\|\right) \\
&\leq \frac{4}{\sqrt{N}}\max_{i\neq j}\left\|\widetilde{X}_{i,j} - \Sigma\right\| + \frac{4}{\sqrt{N}}\|\Sigma(k) - \Sigma\| \\
&\leq 4\frac{1}{\sqrt{N}}\max_{i\neq j}\left\|\widetilde{X}_{i,j} - \Sigma\right\| + \frac{4}{\sqrt{N}}\|\Sigma\|\sqrt{\frac{\mathrm{rk}(\Sigma)}{k}} \\
&\leq 4\frac{1}{\sqrt{N}}\max_{i\neq j}\left\|\widetilde{X}_{i,j} - \Sigma\right\| + 140\frac{\|\Sigma\|\sqrt{\mathrm{rk}(\Sigma)}}{N}.
\end{aligned}$$

Therefore, the event

$$\mathcal{E} := \Bigg\{ \lambda_1 \geq \frac{140\,\|\Sigma\|}{\sqrt{N}} \sqrt{\mathrm{rk}(\Sigma)} + 4 \left\| \frac{1}{N} \sum_{i \neq j} \widetilde{X}_{i,j} - \Sigma \right\|,$$

$$\lambda_2 \geq \frac{140\,\|\Sigma\|}{N} \sqrt{\mathrm{rk}(\Sigma)} + 4 \frac{1}{\sqrt{N}} \max_{i \neq j} \left\| \widetilde{X}_{i,j} - \Sigma \right\| \Bigg\}$$

is a subset of both event $\mathcal{E}_1$ and event $\mathcal{E}_2$, and all previous results hold on the event $\mathcal{E}$ naturally.

Now applying law of cosines to the left-hand side of

$$\frac{2}{N} \sum_{i \neq j} \left\langle \Sigma - \widehat{S}_\lambda + \sqrt{N}(\widetilde{U}^*_{i,j} - \widehat{U}_{i,j}), S - \widehat{S}_\lambda + \sqrt{N}(\widetilde{U}^*_{i,j} - \widehat{U}_{i,j}) \right\rangle < 0,$$

we get that

$$\frac{1}{N} \sum_{i \neq j} \left\| \Sigma - \widehat{S}_\lambda + \sqrt{N}(\widetilde{U}^*_{i,j} - \widehat{U}_{i,j}) \right\|_F^2 + \frac{1}{N} \sum_{i \neq j} \left\| S - \widehat{S}_\lambda + \sqrt{N}(\widetilde{U}^*_{i,j} - \widehat{U}_{i,j}) \right\|_F^2$$

$$< \|\Sigma - S\|_F^2 .$$

This implies the inequality

$$\frac{1}{N} \sum_{i \neq j} \left\| \Sigma - \widehat{S}_\lambda + \sqrt{N}(\widetilde{U}^*_{i,j} - \widehat{U}_{i,j}) \right\|_F^2 < \|\Sigma - S\|_F^2 . \tag{A.63}$$

On the event $\mathcal{E}$ with $k = \lfloor \frac{N\lambda_2^2}{1200\lambda_1^2} \rfloor$, $\left|\widetilde{J}\right| \leq \frac{N}{800}$, we can combine the result of Lemma 5 with the equation (A.63) to get that

$$\frac{1}{2} \left(\left\| \widehat{S}_\lambda - \Sigma \right\|_F^2 + \sum_{i \neq j} \left\| \widehat{U}_{i,j} - \widetilde{U}^*_{i,j} \right\|_F^2 \right) \leq \|\Sigma - S\|_F^2 . \tag{A.64}$$

This bound is consistent with (A.50), which provides upper bounds for both the estimation of Σ and $\widetilde{U}^*_{i,j}, (i,j) \in I_n^2$. To complete the proof, we repeat part of the previous argument to derive an improved bound for the estimation of Σ only, while treating $\widetilde{U}^*_{i,j}$ as "nuisance parameters". Let

$$G(S) := \frac{1}{N}\sum_{i\neq j}\left\|\widetilde{Y}_{i,j}\widetilde{Y}_{i,j}^T - S - \sqrt{N}\widehat{U}_{i,j}\right\|_F^2 + \lambda_1\|S\|_1 \tag{A.65}$$

as before, and we note that the directional derivative of G at the point $\widehat{S}_\lambda$ in the direction $S-\widehat{S}_\lambda$ is nonnegative for any symmetric matrix S, implying that there exists $\widehat{V}\in\partial\left\|\widehat{S}_\lambda\right\|_1$ such that

$$-\frac{2}{N}\sum_{i\neq j}\left\langle \widetilde{Y}_{i,j}\widetilde{Y}_{i,j}^T - \widehat{S}_\lambda - \sqrt{N}\widehat{U}_{i,j}, S-\widehat{S}_\lambda\right\rangle + \lambda_1\left\langle \widehat{V}, S-\widehat{S}_\lambda\right\rangle \geq 0.$$

Proceeding as before, we see that there exists $V\in\partial\|S\|_1$ such that

$$\frac{2}{N}\sum_{i\neq j}\left\langle \Sigma-\widehat{S}_\lambda, S-\widehat{S}_\lambda\right\rangle \leq \lambda_1\left\langle V, S-\widehat{S}_\lambda\right\rangle + \frac{2}{N}\sum_{i\neq j}\left\langle \widetilde{X}_{i,j}-\Sigma, \widehat{S}_\lambda - S\right\rangle + \frac{2}{\sqrt{N}}\sum_{i\neq j}\left\langle \widetilde{U}^*_{i,j}-\widehat{U}_{i,j}, \widehat{S}_\lambda - S\right\rangle.$$

Combining (A.38, A.40) with the inequality above and applying the law of cosines, we deduce that

$$\begin{aligned}\left\|\Sigma-\widehat{S}_\lambda\right\|_F^2 &+ \left\|S-\widehat{S}_\lambda\right\|_F^2 + (\lambda_1 - 2\|\Delta\|)\left\|P_{L^\perp}\widehat{S}_\lambda P_{L^\perp}\right\|_1 \\ &\leq \frac{2}{\sqrt{N}}\sum_{i\neq j}\left\langle \widetilde{U}^*_{i,j}-\widehat{U}_{i,j}, \widehat{S}_\lambda - S\right\rangle + \|\Sigma-S\|_F^2 \\ &\quad + \left(2\sqrt{2\,\mathrm{rank}(S)}\,\|\Delta\| + \lambda_1\sqrt{\mathrm{rank}(S)}\right)\left\|\widehat{S}_\lambda - S\right\|_F,\end{aligned} \tag{A.66}$$

where, as before, $\Delta = \frac{1}{N}\sum_{i\neq j}\widetilde{X}_{i,j}-\Sigma$. On the event $\mathcal{E}$, we have that $\lambda_1 \geq 2\|\Delta\|$, so using the inequality $2ab \leq a^2 + b^2$, we get that

$$\begin{aligned}
&\left(2\sqrt{2\,\mathrm{rank}(S)}\,\|\Delta\| + \lambda_1\sqrt{\mathrm{rank}(S)}\right)\left\|\widehat{S}_\lambda - S\right\|_F \\
&= 2\sqrt{2}\left(2\sqrt{2\,\mathrm{rank}(S)}\,\|\Delta\| + \lambda_1\sqrt{\mathrm{rank}(S)}\right)\left(\frac{\left\|\widehat{S}_\lambda - S\right\|_F}{2\sqrt{2}}\right) \\
&\leq \frac{1}{8}\left\|\widehat{S}_\lambda - S\right\|_F^2 + 2\lambda_1^2\,\mathrm{rank}(S)(\sqrt{2}+1)^2. \qquad \text{(A.67)}
\end{aligned}$$

On the other hand, we can repeat the reasoning in (A.60) and apply Lemma 4 to deduce that

$$\begin{aligned}
&\sum_{(i,j)\notin\widetilde{J}}\left\|\widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\|_1 = \sum_{\widetilde{J}}\left\|P_{L^\perp_{i,j}}\widehat{U}_{i,j}P_{L^\perp_{i,j}}\right\|_1 \\
&\leq 3\left(\frac{\lambda_1}{\lambda_2}\left\|\mathcal{P}_{L(k)}\left(\widehat{S}_\lambda - \Sigma(k)\right)\right\|_1 + \sum_{\widetilde{J}}\left\|\mathcal{P}_{L_{i,j}}\left(\widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right)\right\|_1\right) \\
&\leq 3\left(\frac{\lambda_1}{\lambda_2}\cdot\sqrt{2k}\left\|\widehat{S}_\lambda - \Sigma\right\|_F + 2\sum_{\widetilde{J}}\left\|\widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\|_F\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\frac{2}{\sqrt{N}}\sum_{i\neq j}\left\langle\widetilde{U}^*_{i,j} - \widehat{U}_{i,j}, \widehat{S}_\lambda - S\right\rangle \\
&\leq \frac{2\left\|S - \widehat{S}_\lambda\right\|_F}{\sqrt{N}}\cdot\sum_{(i,j)\in\widetilde{J}}\left\|\widetilde{U}^*_{i,j} - \widehat{U}_{i,j}\right\|_F + \frac{2\left\|S - \widehat{S}_\lambda\right\|}{\sqrt{N}}\cdot\sum_{(i,j)\notin\widetilde{J}}\left\|\widetilde{U}^*_{i,j} - \widehat{U}_{i,j}\right\|_1 \\
&\leq \frac{2\left\|S - \widehat{S}_\lambda\right\|_F}{\sqrt{N}}\left(7\sum_{\widetilde{J}}\left\|\widetilde{U}^*_{i,j} - \widehat{U}_{i,j}\right\|_F + \frac{\lambda_1}{\lambda_2}\cdot 3\sqrt{2k}\left\|\widehat{S}_\lambda - \Sigma\right\|_F\right) \\
&\leq \frac{\lambda_1}{\lambda_2}\cdot 3\sqrt{2}\sqrt{\frac{k}{N}}\left(\left\|S - \widehat{S}_\lambda\right\|_F^2 + \left\|\widehat{S}_\lambda - \Sigma\right\|_F^2\right) + 14\left\|S - \widehat{S}_\lambda\right\|_F\sqrt{\frac{|\widetilde{J}|}{N}}\sqrt{\sum_{\widetilde{J}}\left\|\widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\|_F^2}.
\end{aligned}$$

(A.68)

To estimate $\sqrt{\sum_{\widetilde{J}}\left\|\widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\|_F^2}$, we apply the inequality (A.50) which entails that

$$\sqrt{\sum_{\widetilde{J}} \left\|\widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\|_F^2} \leq 2\sqrt{2}\left(\|\Sigma - S\|_F + \sqrt{\operatorname{rank}(S)}\lambda_1(\sqrt{2}+1) + \lambda_2 \frac{(4/3+\sqrt{2})}{\sqrt{2}}\sqrt{|\widetilde{J}|}\right),$$

given that $k = \lfloor \frac{N\lambda_2^2}{1200\lambda_1^2} \rfloor$, $\operatorname{rank}(S) \leq \frac{1}{56000} \cdot \frac{n^2\lambda_2^2}{\lambda_1^2}$, $|\widetilde{J}| \leq \frac{N}{6400}$. Therefore, by applying the bound $2ab \leq a^2 + b^2$ several times, we deduce that

$$\begin{aligned}
&14 \left\|S - \widehat{S}_\lambda\right\|_F \sqrt{\frac{|\widetilde{J}|}{N}} \sqrt{\sum_{\widetilde{J}} \left\|\widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\|_F^2} \\
&\leq 2 \cdot 14\sqrt{2}\sqrt{\frac{|\widetilde{J}|}{N}} \left\|S - \widehat{S}_\lambda\right\|_F \cdot \|\Sigma - S\|_F + 2 \cdot \frac{\left\|S - \widehat{S}_\lambda\right\|_F}{2} \cdot 28(4/3+\sqrt{2})\lambda_2 \sqrt{\frac{|\widetilde{J}|^2}{N}} \\
&\qquad + 2 \cdot 2\sqrt{2}(\sqrt{2}+1)\sqrt{\frac{|\widetilde{J}|}{N}} \left\|S - \widehat{S}_\lambda\right\|_F \cdot 7\lambda_1 \sqrt{\operatorname{rank}(S)} \\
&\leq 14\sqrt{2}\sqrt{\frac{|\widetilde{J}|}{N}} \left(\left\|S - \widehat{S}_\lambda\right\|_F^2 + \|\Sigma - S\|_F^2 \right) + \left(8(\sqrt{2}+1)^2 \frac{|\widetilde{J}|}{N} + \frac{1}{4} \right) \left\|S - \widehat{S}_\lambda\right\|_F^2 \\
&\qquad + 49\lambda_1^2 \operatorname{rank}(S) + \left(28(4/3+\sqrt{2})\right)^2 \frac{|\widetilde{J}|^2}{N}\lambda_2^2.
\end{aligned}$$

Combining this with (A.66,A.67,A.68), we obtain that

$$\begin{aligned}
\left\|\Sigma - \widehat{S}_\lambda\right\|_F^2 \leq \frac{11}{5} \|\Sigma - S\|_F^2 + \frac{8}{5}\left(2(\sqrt{2}+1)^2 + 49\right)\lambda_1^2 \operatorname{rank}(S) \\
+ \frac{8}{5}\left(28(4/3+\sqrt{2})\right)^2 \frac{|\widetilde{J}|^2}{N}\lambda_2^2 \quad \text{(A.69)}
\end{aligned}$$

under the assumptions that $\operatorname{rank}(S) \leq \frac{1}{56000} \cdot \frac{n^2\lambda_2^2}{\lambda_1^2}$, $3\sqrt{2}\frac{\lambda_1}{\lambda_2}\sqrt{\frac{k}{N}} + 14\sqrt{2}\sqrt{\frac{|\widetilde{J}|}{N}} \leq \frac{3}{8}$ and $8(\sqrt{2}+1)^2\frac{|\widetilde{J}|}{N} + \frac{1}{4} \leq \frac{1}{2}$. The assumptions hold for $\operatorname{rank}(S) \leq \frac{1}{56000} \cdot \frac{n^2\lambda_2^2}{\lambda_1^2}$, $k = \lfloor \frac{N\lambda_2^2}{1200\lambda_1^2} \rfloor$ and $|\widetilde{J}| \leq \frac{N}{6400}$. Note that the coefficient $\frac{11}{5}$ can be made smaller. Given $\delta \in (0, \frac{3}{8})$, we assume that $3\sqrt{2}\frac{\lambda_1}{\lambda_2}\sqrt{\frac{k}{N}} + 14\sqrt{2}\sqrt{\frac{|\widetilde{J}|}{N}} \leq \delta$, which holds with the choices of $k \leq \frac{N\lambda_2^2}{72\lambda_1^2} \cdot \delta^2$ and $|\widetilde{J}| \leq \frac{N}{1568} \cdot \delta^2$ respectively. Also, we assume that $\operatorname{rank}(S) \leq c_1\frac{n^2\lambda_2^2}{\lambda_1^2}$ for some constant $c_1 \leq \frac{1}{5980}$ according to (A.51). Then (A.69) becomes

$$\left\|\Sigma - \widehat{S}_\lambda\right\|_F^2 \le \frac{1+\delta}{1-\delta}\left\|\Sigma - S\right\|_F^2 + \frac{1}{1-\delta}\left(2(\sqrt{2}+1)^2 + 49\right)\lambda_1^2 \operatorname{rank}(S) + \frac{1}{1-\delta}\left(28(4/3+\sqrt{2})\right)^2 \frac{|\widetilde{J}|^2}{N}\lambda_2^2, \quad \text{(A.70)}$$

where $\frac{1+\delta}{1-\delta} \in (1, \frac{11}{5}]$ is a number close to 1. Finally, by (3.8), we see that $\frac{|\widetilde{J}|}{N} \le 2\frac{|J|}{n}$, so we can write the last term of the inequality (A.70) as

$$\left(28(4/3+\sqrt{2})\right)^2 \frac{|\widetilde{J}|^2}{N}\lambda_2^2 = \left(28(4/3+\sqrt{2})\right)^2 \lambda_2^2 |J|^2 \frac{(2n-|J|-1)^2}{n(n-1)} \le 4\left(28(4/3+\sqrt{2})\right)^2 \lambda_2^2 |J|^2$$

under the assumption that $|J| \le \frac{n\delta^2}{3136}$. This completes the proof.

A.2.1 Proof of Theorem 3

In this section we prove Theorem 3, which provides the lower bound for the choice of λ_1. We start with a well-known theorem on the concentration of sample covariance matrix.

Theorem 10 (Koltchinskii and Lounici (2017, Theorem 9))**.** *Assume that Z is L-sub-Gaussian with mean zero and sample covariance matrix Σ. Let $Z_1, \dots, Z_n$ be independent samples of Z, then there exists $c(L) > 0$ depending only on L, such that*

$$\left\|\frac{1}{n}\sum_{j=1}^{n} Z_j Z_j^T - \Sigma\right\| \le c(L)\left\|\Sigma\right\|\left(\sqrt{\frac{\operatorname{rk}(\Sigma)}{n}} \vee \frac{\operatorname{rk}(\Sigma)}{n} \vee \sqrt{\frac{t}{n}} \vee \frac{t}{n}\right)$$

with probability at least $1 - e^{-t}$.

Remark 8. Assuming that $\operatorname{rk}(\Sigma) \le n$ and $t \le n$, the bound can be reduced to

$$c(L)\left\|\Sigma\right\|\left(\sqrt{\frac{\operatorname{rk}(\Sigma)}{n}} \vee \sqrt{\frac{t}{n}}\right).$$

Now we prove Theorem 3.

Proof of Theorem 3. First, it is well-known that

$$\frac{1}{n(n-1)}\sum_{i\neq j}\frac{(Z_i-Z_j)(Z_i-Z_j)^T}{2}=\frac{1}{n-1}\sum_{i=1}^{n}(Z_i-\bar{Z})(Z_i-\bar{Z})^T, \tag{A.71}$$

where $\bar{Z}:=\frac{1}{n}\sum_{i=1}^{n}Z_i$. Therefore,

$$\Delta:=\frac{1}{n(n-1)}\sum_{i\neq j}\widetilde{Z}_{ij}\widetilde{Z}_{ij}^T-\Sigma=\frac{1}{n(n-1)}\sum_{i\neq j}\frac{(Z_i-Z_j)(Z_i-Z_j)^T}{2}-\Sigma=\widetilde{\Sigma}_s-\Sigma.$$

Recall $\mathbb{E}[Z_j]=\mu, j=1,\ldots,n$, and note that

$$\widetilde{\Sigma}_s=\frac{1}{n-1}\left(\sum_{i=1}^{n}(Z_i-\mu)(Z_i-\mu)^T-n(\bar{Z}-\mu)(\bar{Z}-\mu)^T\right),$$

hence we have the decomposition

$$\begin{aligned}(n-1)\left\|\Delta\right\|&=\left\|(n-1)\widetilde{\Sigma}_s-n\Sigma+\Sigma\right\|\\&\leq\left\|\sum_{i=1}^{n}(Z_i-\mu)(Z_i-\mu)^T-n\Sigma\right\|+\left\|\Sigma-n(\bar{Z}-\mu)(\bar{Z}-\mu)^T\right\|.\end{aligned} \tag{A.72}$$

We will bound the two terms on the right-hand side of (A.72) one by one. First, note that $Z_j-\mu, j=1,\ldots,n$ are i.i.d L-sub-Gaussian random vectors with mean zero and covariance matrix Σ, hence Theorem 10 immediately gives that

$$\left\|\sum_{i=1}^{n}(Z_i-\mu)(Z_i-\mu)^T-n\Sigma\right\|\leq nc(L)\left\|\Sigma\right\|\left(\sqrt{\frac{\mathrm{rk}(\Sigma)}{n}}\vee\frac{\mathrm{rk}(\Sigma)}{n}\vee\sqrt{\frac{t}{n}}\vee\frac{t}{n}\right) \tag{A.73}$$

with probability at least $1-e^{-t}$. To bound the second term, consider the random variable $Y:=\sqrt{n}\left(\bar{Z}-\mu\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Z_i-\mu)$. Clearly, $\mathbb{E}[Y]=0$ and

$$\mathbb{E}\left[YY^T\right]=n\mathbb{E}\left[(\bar{Z}-\mu)(\bar{Z}-\mu)^T\right]=\frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}(Z_i-\mu)(Z_j-\mu)^T\right]$$

$$= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}(Z_i-\mu)(Z_i-\mu)^T\right] = \Sigma,$$

where we used the independence of $Z_i, i = 1, \ldots, n$ in the third equality. Moreover, Corollary 1 guarantees that Y is L-sub-Gaussian. Therefore, Y satisfies the conditions in Theorem 10, and a direct application of the theorem implies that

$$\left\|\Sigma - n(\bar{Z}-\mu)(\bar{Z}-\mu)^T\right\| = \left\|YY^T - \Sigma\right\| \leq c(L)\left\|\Sigma\right\|(\mathrm{rk}(\Sigma)\vee t) \tag{A.74}$$

with probability at least $1-e^{-t}$, given that $t \geq 1$. Combining (A.72, A.73, A.74), we deduce that for any $t \geq 1$,

$$\begin{aligned}\|\Delta\| &\leq c(L)\left[\frac{n}{n-1}\|\Sigma\|\left(\sqrt{\frac{\mathrm{rk}(\Sigma)}{n}}\vee\frac{\mathrm{rk}(\Sigma)}{n}\vee\sqrt{\frac{t}{n}}\vee\frac{t}{n}\right)+\frac{1}{n-1}\|\Sigma\|(\mathrm{rk}(\Sigma)\vee t)\right]\\ &\leq c(L)\|\Sigma\|\left[\sqrt{\frac{\mathrm{rk}(\Sigma)+t}{n}}+\frac{\mathrm{rk}(\Sigma)+t}{n}\right]\end{aligned}$$

with probability at least $1-2e^{-t}$, where $c(L)$ is an absolute constant that only depends on L but could vary from step to step.

□

A.2.2 Proof of Theorem 4

In this section we prove Theorem 4, which provides the lower bound for the choice of λ_2.

Proof of Theorem 4. Fix $i \in \{1, \ldots, n\}$, we apply Theorem 10 to Z_i and deduce that for any $u \geq 1$,

$$\left\|Z_iZ_i^T - \Sigma\right\| \leq c(L)\left\|\Sigma\right\|(\mathrm{rk}(\Sigma)+u) = c(L)\left(\mathrm{tr}\,(\Sigma)+\left\|\Sigma\right\|u\right)$$

with probability at least $1-e^{-u}$.

Therefore, by union bound we have that for $t \geq 1$ and $n \geq 1$,

$$\begin{aligned}
&P\left(\max_i \left\|Z_iZ_i^T-\Sigma\right\| \geq c(L)\left[tr(\Sigma)+\log(n)\left\|\Sigma\right\|+\left\|\Sigma\right\|t\right]\right)\\
&\leq \sum_{i=1}^n P\left(\left\|Z_iZ_i^T-\Sigma\right\| \geq c(L)\left[tr(\Sigma)+\log(n)\left\|\Sigma\right\|+\left\|\Sigma\right\|t\right]\right)\\
&= nP\left(\left\|Z_iZ_i^T-\Sigma\right\| \geq c(L)\left[tr(\Sigma)+\left\|\Sigma\right\|\left(\log(n)+t\right)\right]\right)\\
&\leq ne^{-log(n)-t}=e^{-t}.
\end{aligned}$$

In other words, for $t \geq 1$, we have that with probability at least $1-e^{-t}$,

$$\begin{aligned}
\max_i \left\|Z_iZ_i^T-\Sigma\right\| &\leq c(L)\left[\mathrm{tr}\left(\Sigma\right)+\left\|\Sigma\right\|\left(\log(n)+t\right)\right]\\
&= c(L)\left\|\Sigma\right\|\left(\mathrm{rk}(\Sigma)+\log(n)+t\right),
\end{aligned}$$

as desired. □

A.3 Proof of Lemma 1 and Theorem 5

In this subsection we present the proof of Lemma 1 and Theorem 5, which provide error bounds of the estimator in (4.16) in the operator norm. To simplify the expressions, we introduce the following notations, which are valid in this subsection only:

- Denote
$$\rho(u) := \rho_1(u) = \begin{cases} \frac{u^2}{2}, & |u| \leq 1 \\ |u|-\frac{1}{2}, & |u|>1 \end{cases} \qquad \forall u \in \mathbb{R}.$$
- Denote $k_0 = \lfloor n/2 \rfloor$ and $N = n(n-1)$.

- Denote $\theta_2 := \frac{2}{\lambda_2\sqrt{n(n-1)}}$ and
$$\theta_\sigma := \frac{1}{\sigma}\sqrt{\frac{2t}{k_0}}$$
where $\sigma > 0$, $t > 0$ are constants to be specified later.

It is easy to check that $\rho_\lambda(u) = \lambda^2\rho(\frac{u}{\lambda})$ and $\rho'_\lambda(u) = \lambda\rho(\frac{u}{\lambda})$, so with the above notations, we can rewrite the loss function in (4.17) as

$$L(S) = \frac{1}{\theta_2^2}\frac{1}{N}\sum_{i\neq j}\rho\left(\theta_2(H_{i,j} - S)\right) + \frac{\lambda_1}{2}\left\|S\right\|_1. \tag{A.75}$$

The gradient of the loss function is

$$\nabla L(S) = -\frac{1}{N\theta_2}\sum_{i\neq j}\rho'\left(\theta_2(H_{i,j} - S)\right) + \frac{\lambda_1}{2}\partial\left\|S\right\|_1. \tag{A.76}$$

Given $\alpha \in (0,1)$, one can easily verify that $\rho'(\cdot)$ is Hölder continuous on $\mathbb{R}$, namely, $|\rho'(x) - \rho'(y)| \leq 2|x-y|^\alpha, \forall x, y \in \mathbb{R}$. The following theorem shows that $\rho'(\cdot)$ is Hölder continuous in the operator norm, which is crucial for the next part of the proof:

Theorem 11. *(Aleksandrov and Peller (2016, Theorem 1.7.2)) Assume that $f(x)$ is Hölder continuous on $\mathbb{R}$ with $\alpha \in (0,1)$, i.e. $|f(x) - f(y)| \leq C_0|x-y|^\alpha$, $\forall x, y \in \mathbb{R}$. Then there exists an absolute constant c such that*

$$\|f(A) - f(B)\| \leq c(1-\alpha)^{-1}C_0\left\|A - B\right\|^\alpha$$

for any symmetric matrices A and B.

We now present the proofs of Lemma 1 and Theorem 5.

A.3.1 Proof of Lemma 1

It is worth noting that the proof follows the argument in Minsker and Wei (2020, Section 5).

Recall the loss function and its gradient:

$$L(S) = \frac{1}{\theta_2^2}\frac{1}{N}\sum_{i\neq j}\rho(\theta_2(H_{i,j}-S)) + \frac{\lambda_1}{2}\left\|S\right\|_1,$$

$$\nabla L(S) = -\frac{1}{N\theta_2}\sum_{i\neq j}\rho'(\theta_2(H_{i,j}-S)) + \frac{\lambda_1}{2}\partial\left\|S\right\|_1,$$

where $H_{i,j} = \widetilde{Y}_{i,j}\widetilde{Y}_{i,j}^T$. Consider the choice $\lambda_1 > \frac{1}{2\theta_\sigma}$, where $\theta_\sigma := \frac{1}{\sigma}\sqrt{\frac{2t}{k_0}}$. We assume that the minimizer $\widetilde{S} = V \neq 0$. Since $L(S)$ is convex, we have

$$L(V) - L(0) \geq \langle\nabla L(0), V - 0\rangle.$$

Plugging in the explicit form of $\nabla L(0)$, we get that for any $W \in \partial\|S\|_{1|_{S=0}}$,

$$L(V) - L(0) \geq \left\langle -\frac{1}{N\theta_2}\sum_{i\neq j}\rho'(\theta_2 H_{i,j}) + \frac{\lambda_1}{2}W, V\right\rangle,$$

hence

$$L(V) - L(0) \geq \sup_{W\in\partial\|S\|_1|_{S=0}}\left\langle -\frac{1}{N\theta_2}\sum_{i\neq j}\rho'(\theta_2 H_{i,j}) + \frac{\lambda_1}{2}W, V\right\rangle. \tag{A.77}$$

Consider the random variable $\mathcal{X}_{i,j} := \mathbb{1}\{\|H_{i,j} - \Sigma\| \leq \frac{1}{a\theta_\sigma}\}$, $a \geq 2$, and set $\theta_2 = \theta_\sigma$ in what follows. By Chebyshev's inequality,

$$P(\mathcal{X}_{i,j} = 0) \leq a^2\theta_\sigma^2 \operatorname{tr}\mathbb{E}\left[(H_{i,j} - \Sigma)^2\right] \leq a^2\frac{2t}{k_0}r_H,$$

where $r_H = \operatorname{rk}(\mathbb{E}\left[(H_{i,j} - \Sigma)^2\right])$. Define the event

$$\mathcal{E} := \left\{\frac{1}{N}\sum_{i\neq j}(1 - \mathcal{X}_{i,j}) \leq r_H\frac{2a^2t}{k_0}(1 + \sqrt{\frac{3}{2a^2 r_H}})\right\}$$

By the finite difference inequality (see for example, Minsker and Wei (2020, Fact 4-6)),

$$P\Big(\frac{1}{N}\sum_{i\neq j}(1-\mathcal{X}_{i,j}) \geq r_H\frac{2a^2t}{k_0}(1+\tau)\Big) \leq e^{-\tau^2 2a^2 t r_H/3}, \quad 0<\tau<1.$$

Setting $\tau = \sqrt{\frac{3}{2a^2 r_H}}$ we get $P(\mathcal{E}) \geq 1-e^{-t}$. Therefore, for $k_0 \geq 32a^2 t r_H$, we have that

$$\frac{1}{N}\sum_{i\neq j}(1-\mathcal{X}_{i,j}) \leq \frac{1}{8}$$

with probability $\geq 1-e^{-t}$. Note that on the event $\{\mathcal{X}_{i,j}=0\}$,

$$\left\|\rho'(\theta_\sigma H_{i,j})\right\| \leq 1$$

since $|\rho'(x)| \leq 1$ for any $x\in\mathbb{R}$. On the other hand, on the event $\{\mathcal{X}_{i,j}=1\}$, we have that $\|H_{i,j}-\Sigma\| \leq \frac{1}{a\theta_\sigma}$, hence

$$\|H_{i,j}\| \leq \frac{1}{a\theta_\sigma} + \|\Sigma\| \leq \frac{1}{a\theta_\sigma} + \frac{1}{b\theta_\sigma}$$

given that $k_0 \geq \frac{2b^2t^2\|\Sigma\|^2}{\sigma^2}$. Therefore, by Theorem 11 we have that

$$\left\|\rho'(\theta_\sigma H_{i,j})\right\| \leq 2c(1-\alpha)^{-1}\left\|\theta_\sigma H_{i,j}\right\|^\alpha \leq 2c(1-\alpha)^{-1}\left(\frac{1}{a}+\frac{1}{b}\right)^\alpha.$$

Setting a, b large enough such that $2c(1-\alpha)^{-1}\left(\frac{1}{a}+\frac{1}{b}\right)^\alpha + \frac{1}{8} \leq \frac{1}{4}$, we have that

$$\begin{aligned}\left\|\frac{1}{\theta_\sigma N}\sum_{i\neq j}\rho'(\theta_\sigma H_{i,j})\right\| &\leq \left\|\frac{1}{\theta_\sigma N}\sum_{i\neq j}\rho'(\theta_\sigma H_{i,j})\mathcal{X}_{i,j}\right\| + \left\|\frac{1}{\theta_\sigma N}\sum_{i\neq j}\rho'(\theta_\sigma H_{i,j})(1-\mathcal{X}_{i,j})\right\| \\ &\leq \frac{1}{\theta_\sigma}2c(1-\alpha)^{-1}(\frac{1}{a}+\frac{1}{b})^\alpha\frac{1}{N}\sum_{i\neq j}\mathcal{X}_{i,j} + \frac{1}{8\theta_\sigma} \leq \frac{1}{4}\frac{1}{\theta_\sigma}.\end{aligned}$$

Therefore,

$$L(V) - L(0) \geq \sup_{W \in \partial \|S\|_1|_{S=0}} \left\langle -\frac{1}{N\theta_2} \sum_{i \neq j} \rho'(\theta_2 H_{i,j}) + \frac{\lambda_1}{2} W, V \right\rangle$$
$$= \left\langle -\frac{1}{N\theta_2} \sum_{i \neq j} \rho'(\theta_2 H_{i,j}), V \right\rangle + \frac{\lambda_1}{2} \|V\|_1 \geq -\frac{1}{4\theta_\sigma} \|V\|_1 + \frac{\lambda_1}{2} \|V\|_1 > 0,$$

where we used the fact that $\partial \|S\| |_{S=0} = \{W : \|W\| \leq 1\}$ and $\sup_{W:\|W\|\leq 1} \langle W, V\rangle = \|V\|_1$. This is a contradiction to the fact that V is a minimizer of the loss function $L(S)$, and hence we conclude that $\operatorname{argmin}_S L(S) = 0$ with probability at least $1 - e^{-t}$.

A.3.2 Proof of Theorem 5

Recall the loss function

$$L(S) = \frac{1}{\theta_2^2} \frac{1}{N} \sum_{i \neq j} \rho\big(\theta_2 (H_{i,j} - S)\big) + \frac{\lambda_1}{2} \|S\|_1$$

and its gradient

$$\nabla L(S) = -\frac{1}{N\theta_2} \sum_{i \neq j} \rho'\big(\theta_2 (H_{i,j} - S)\big) + \frac{\lambda_1}{2} \partial \|S\|_1 \,,$$

where $H_{i,j} = \widetilde{Y}_{i,j} \widetilde{Y}_{i,j}^T$, $\rho(\cdot) = \rho_1(\cdot)$ and $\theta_2 = \frac{2}{\sqrt{N}\lambda_2}$. Consider the proximal gradient descent iteration:

1. $S^0 := \mathbb{E}[H] = \Sigma$.

2. For $t = 1, 2, \ldots$, do:

   - $T^{t+1} := S^t + \frac{1}{N\theta_2} \sum_{i \neq j} \rho' \left(\theta_2 (H_{i,j} - S^t) \right)$.

   - $S^{t+1} := \operatorname{argmin}_S \{ \frac{1}{2} \left\| S - T^{t+1} \right\|_F^2 + \frac{\lambda_1}{2} \|S\|_1 \}$.

We will show that with an appropriate choice of θ_2, S^{t+1} does not escape a small neighborhood of Σ with high probability, and the result will easily follow. First, the following lemma bounds $\left\|S^{t+1} - T^{t+1}\right\|$:

Lemma 6.

$$\left\|S^{t+1} - T^{t+1}\right\| \leq \frac{\lambda_1}{2}.$$

Proof. Repeating the reasoning for equation (A.116) in section A.6.1 of the supplementary material, we can solve for S^{t+1} explicitly:

$$S^{t+1} = \underset{S}{\operatorname{argmin}} \left\{ \frac{1}{2} \left\|S - T^{t+1}\right\|_F^2 + \frac{\lambda_1}{2} \|S\|_1 \right\} = \gamma_{\frac{\lambda_1}{2}}(T^{t+1}),$$

where $\gamma_\lambda(u) = \text{sign}(u)(|u| - \lambda)_+$ is the function that shrinks eigenvalues to 0. A direct calculation gives that

$$\left\|S^{t+1} - T^{t+1}\right\| = \left\|S^{t+1} - T^{t+1}\right\| = \left\|\gamma_{\frac{\lambda_1}{2}}(T^{t+1}) - T^{t+1}\right\| \leq \frac{\lambda_1}{2}.$$

□

Applying Lemma 6 , we see that

$$\left\|S^{t+1} - \Sigma\right\| \leq \left\|S^{t+1} - T^{t+1}\right\| + \left\|T^{t+1} - \Sigma\right\| \leq \frac{\lambda_1}{2} + \left\|T^{t+1} - \Sigma\right\|.$$

It remains to bound $\left\|T^{t+1} - \Sigma\right\|$. Note that

$$\left\|T^{t+1} - \Sigma\right\| = \left\|S^t - \Sigma + \frac{1}{N\theta_2} \sum_{i \neq j} \rho' \left(\theta_2(H_{i,j} - S^t)\right)\right\|$$

$$\leq \underbrace{\left\| \frac{1}{N\theta_2} \sum_{i \neq j} \left[\rho' \left(\theta_2 (H_{i,j} - S^t) \right) - \rho'(\theta_2(H_{i,j} - \Sigma)) \right] + S^t - \Sigma \right\|}_{:=\mathrm{I}}$$

$$+ \underbrace{\left\| \frac{1}{N\theta_2} \sum_{i \neq j} \rho' \left(\theta_2 (H_{i,j} - \Sigma) \right) \right\|}_{:=\mathrm{II}}. \quad \text{(A.78)}$$

We will bound terms I and II separately. Set $k_0 = \lfloor n/2 \rfloor$ and define

$$Y_{i,j}(S;\theta) := \rho'(\theta(H_{i,j} - S)),$$

$$W_{i_1,\dots,i_n}(S;\theta) := \frac{1}{k_0} \left[Y_{i_1,i_2}(S;\theta) + \cdots + Y_{i_{2k_0-1},i_{2k_0}}(S;\theta) \right],$$

where $(i_1, \dots, i_n) \in \pi_n$ is a permutation. Fact 6 in Minsker and Wei (2020) implies that

$$\mathrm{I} = \left\| \frac{1}{\theta_2 n!} \sum_{\pi_n} \left(W_{i_1,\dots,i_n}(S^t;\theta_2) - W_{i_1,\dots,i_n}(\Sigma;\theta_2) \right) + S^t - \Sigma \right\|, \quad \text{(A.79)}$$

$$\mathrm{II} = \left\| \frac{1}{\theta_2 n!} \sum_{\pi_n} W_{i_1,\dots,i_n}(\Sigma;\theta_2) \right\|. \quad \text{(A.80)}$$

For a given $\sigma^2 \geq \left\| \mathbb{E}\left[(H_{i,j} - \Sigma)^2\right] \right\|$ and $\theta_\sigma := \frac{1}{\sigma}\sqrt{\frac{2t}{k_0}}$, the following lemma provides a bound for the term II:

Lemma 7. *Recall that* $r_H = \mathrm{rk}(\mathbb{E}\left[(H_{i,j} - \Sigma)^2\right])$. *Given* $t \geq 1$, *we have that*

$$\left\| \frac{1}{\theta_2 n!} \sum_{\pi_n} W_{i_1,\dots,i_n}(\Sigma;\theta_2) \right\| \leq \theta_2 \sigma^2 + \frac{t}{\theta_2 k_0}$$

with probability at least $1 - \frac{8}{3} r_H e^{-t}$. *When* $\theta_2 = \theta_\sigma$, *the upper bound takes the form* $\frac{3}{\sqrt{2}}\sigma\sqrt{\frac{t}{k_0}}$.

Proof. It is easy to verify that for any $x \in \mathbb{R}$,

$$-\log(1 - x + x^2) \leq \rho'(x) \leq \log(1 + x + x^2),$$

and the rest of the proof follows from the argument in Minsker and Wei (2020, Section 5.5). □

To estimate the term I, consider the random variable

$$L_n(\delta) := \sup_{\|S-\Sigma\|\le\delta} \left\| \frac{1}{\theta_\sigma n!} \sum_{\pi_n} \left(W_{i_1,\ldots,i_n}(S;\theta_\sigma) - W_{i_1,\ldots,i_n}(\Sigma;\theta_\sigma)\right) + S - \Sigma \right\|.$$

Lemma 8. *Given $\alpha \in (0,1)$, we have that for all $\delta \le \frac{1}{2}\frac{1}{\theta_\sigma}$,*

$$L_n(\delta) \le r_H \frac{13t}{k_0} \delta \left(1 + 4c(1-\alpha)^{-1}\right)$$

with probability at least $1 - e^{-t}$, where $c > 0$ is an absolute constant specified in Theorem 11.

Proof. Define $\mathcal{X}_{i,j} := \mathbb{1}\left\{ \|H_{i,j} - \Sigma\| \le \frac{1}{2\theta_\sigma} \right\}$ and consider the event $\mathcal{E} := \left\{ \sum_{i\ne j}(1-\mathcal{X}_{i,j}) \le N\frac{8t}{k_0} r_H (1+\sqrt{\frac{3}{8r_H}}) \right\}$. Minsker and Wei (2020) proves that $P(\mathcal{E}) \ge 1 - e^{-t}$. For S with $\|S - \Sigma\| \le \delta \le \frac{1}{2\theta_\sigma}$, we have that

$$\begin{aligned}
&\frac{1}{\theta_\sigma n!} \sum_{\pi_n} \left(W_{i_1,\ldots,i_n}(S;\theta_\sigma) - W_{i_1,\ldots,i_n}(\Sigma;\theta_\sigma)\right) + S - \Sigma \\
&\qquad = \frac{1}{N\theta_\sigma} \sum_{i\ne j} \left[\rho'\left(\theta_\sigma(H_{i,j} - S)\right) - \rho'(\theta_\sigma(H_{i,j} - \Sigma))\right] + S - \Sigma \\
&\qquad = \left(\frac{1}{N\theta_\sigma} \sum_{i\ne j} \left[\rho'\left(\theta_\sigma(H_{i,j} - S)\right) - \rho'(\theta_\sigma(H_{i,j} - \Sigma))\right] \mathcal{X}_{i,j} + S - \Sigma \right) \\
&\qquad\qquad + \left(\frac{1}{N\theta_\sigma} \sum_{i\ne j} \left[\rho'\left(\theta_\sigma(H_{i,j} - S)\right) - \rho'(\theta_\sigma(H_{i,j} - \Sigma))\right] (1 - \mathcal{X}_{i,j}) \right).
\end{aligned}$$

We will separately control the two terms on the right-hand side of the equality above. First, note that when $\mathcal{X}_{i,j} = 1$, we have that $\|H_{i,j} - \Sigma\| \le \frac{1}{2\theta_\sigma} \le \frac{1}{\theta_\sigma}$, and $\|H_{i,j} - S\| \le \|H_{i,j} - \Sigma\| + \|\Sigma - S\| \le \frac{1}{\theta_\sigma}$. Therefore, on the event $\mathcal{E}$,

$$\left\| \frac{1}{N\theta_\sigma} \sum_{i\ne j} \left[\rho'\left(\theta_\sigma(H_{i,j} - S)\right) - \rho'(\theta_\sigma(H_{i,j} - \Sigma))\right] \mathcal{X}_{i,j} + S - \Sigma \right\|$$

$$= \left\| \frac{1}{N\theta_\sigma} \sum_{i \neq j} \left[\theta_\sigma(H_{i,j} - S) - \theta_\sigma(H_{i,j} - \Sigma) \right] \mathcal{X}_{i,j} + S - \Sigma \right\|$$

$$= \left\| \frac{1}{N} \sum_{i \neq j} (S - \Sigma)(1 - \mathcal{X}_{i,j}) \right\| \leq r_H \frac{8t}{k_0}(1 + \sqrt{\frac{3}{8r_H}})\delta \leq r_H \frac{13t}{k_0}\delta. \quad \text{(A.81)}$$

Next, recall that for any $\alpha \in (0,1)$, $|\rho'(x) - \rho'(y)| \leq 2|x-y|^\alpha$ for any $x, y \in \mathbb{R}$, so by Theorem 11, there exists a constant $c > 0$ such that

$$\left\| \rho'(A) - \rho'(B) \right\| \leq 2c(1-\alpha)^{-1} \left\| A - B \right\|^\alpha$$

for any symmetric matrices A and B. Therefore, on the event $\mathcal{E}$,

$$\left\| \frac{1}{N\theta_\sigma} \sum_{i \neq j} \left[\rho'\left(\theta_\sigma(H_{i,j} - S)\right) - \rho'(\theta_\sigma(H_{i,j} - \Sigma)) \right] (1 - \mathcal{X}_{i,j}) \right\|$$

$$\leq 2c(1-\alpha)^{-1} \left\| \theta_\sigma(\Sigma - S) \right\|^\alpha \cdot \frac{1}{N\theta_\sigma} \sum_{i \neq j} (1 - \mathcal{X}_{i,j})$$

$$\leq 2c(1-\alpha)^{-1}(\frac{1}{2})^{\alpha - 1}\delta \cdot r_H \frac{13t}{k_0} \leq 4c(1-\alpha)^{-1} \cdot r_H \frac{13t}{k_0}\delta. \quad \text{(A.82)}$$

Combining (A.81), (A.82) and $P(\mathcal{E}) \geq 1 - e^{-t}$, we have that

$$L_n(\delta) \leq r_H \frac{13t}{k_0} \delta \left(1 + 4c(1-\alpha)^{-1}\right)$$

with probability at least $1 - e^{-t}$. □

Now we can bound $\left\| S^{t+1} - \Sigma \right\|$ as follows:

For $t = 0, 1, \ldots$, define

$$\delta_0 = 0,$$

$$\delta_{t+1} = r_H \frac{13t}{k_0} \left(1 + 4c(1-\alpha)^{-1}\right) \delta_t + 5.75\sigma\sqrt{\frac{t}{k_0}} + \frac{\lambda_1}{2}.$$

Choose t, k such that $r_H \frac{13t}{k_0} \left(1 + 4c(1-\alpha)^{-1}\right) \leq \frac{1}{20}$ and $t \leq \frac{k_0}{520}$, we have that $5.75\sigma\sqrt{\frac{t}{k_0}} \leq \frac{1}{40\theta_\sigma}$, hence

$$\delta_{t+1} \leq \frac{1}{20}\delta_t + \frac{1}{40\theta_\sigma} + \frac{\lambda_1}{2} \leq \frac{1}{2\theta_\sigma}$$

given that $\delta_t \leq \frac{1}{2\theta_\sigma}$ and $\lambda_1 \leq \frac{1}{2\theta_\sigma}$. Since $\left\|S^0 - \Sigma\right\| = 0 \leq \frac{1}{2\theta_\sigma}$, we have that for $t = 0, 1, \ldots$,

$$\begin{aligned}\left\|S^{t+1} - \Sigma\right\| &\leq \frac{\lambda_1}{2} + L_n(\delta_t) + \frac{3}{\sqrt{2}}\sigma\sqrt{\frac{t}{k_0}} \\ &\leq \frac{\lambda_1}{2} + r_H \frac{13t}{k_0} \left(1 + 4c(1-\alpha)^{-1}\right) \delta_t + \frac{3}{\sqrt{2}}\sigma\frac{t}{k_0} \leq \delta_{t+1}\end{aligned}$$

with probability at least $1 - (\frac{8}{3}r_H + 1)e^{-t}$. Finally, for $\gamma := r_H \frac{13t}{k_0} \left(1 + 4c(1-\alpha)^{-1}\right) \leq \frac{1}{40}$, it is easy to check that for $t = 0, 1, \ldots$,

$$\begin{aligned}\delta_{t+1} &= \gamma^{t+1}\delta_0 + \sum_{l=0}^{t} \gamma^t (\frac{\lambda_1}{2} + \frac{3}{\sqrt{2}}\sigma\sqrt{\frac{t}{k_0}}) \\ &\leq \sum_{l\geq 0} \frac{1}{40^l} (\frac{\lambda_1}{2} + \frac{3}{\sqrt{2}}\sigma\sqrt{\frac{t}{k_0}}) = \frac{20}{39}\lambda_1 + \frac{20\sqrt{2}}{13}\sigma\sqrt{\frac{t}{k_0}}.\end{aligned}$$

By Theorem 7, $S^t \to \widehat{S}$ pointwise as $t \to \infty$, so the result follows.

To this end, we note that the proof above can be repeated with $\theta_2 < \theta_\sigma := \frac{1}{\sigma}\sqrt{\frac{2t}{k_0}}$, in which case Lemma 1 will be valid for

$$k_0 \geq \max\left\{44a^2 tr_H, \frac{2b^2t^2 \left\|\Sigma\right\|^2}{\sigma^2}\right\}.$$

Moreover, the upper bounds in Lemma 7 and Lemma 8 will become

$$\sigma\sqrt{\frac{2t}{k_0}} + \frac{t}{\theta_2 k_0}$$

and

$$8\theta_2^2\sigma^2 r_H \left(1+4c(1-\alpha)^{-1}\right)\delta$$

respectively. Consequently, we can deduce that whenever

$$8\theta_2^2\sigma^2 r_H \left(1+2c(1-\alpha)^{-1}\right) \le \frac{1}{20},$$

which is valid as long as

$$r_H \frac{t}{k_0}\left(1+2c(1-\alpha)^{-1}\right) \le \frac{1}{640},$$

the following inequality holds with probability at least $1-(\frac{8}{3}r_H+1)e^{-t}$:

$$\left\|\widehat{S}_\lambda - \Sigma\right\| \le \frac{20}{39}\lambda_1 + \frac{40}{39}\left[\sigma\sqrt{\frac{2t}{k_0}} + \lambda_2 t\right].$$

This completes the proof.

A.4 Proofs omitted in Section 4.2

A.4.1 Proof of Lemma 2

In this section we prove that the fraction of outliers is small with high probability for heavy-tailed data. Denote $k_0 = \lfloor n/2 \rfloor$, $\Sigma = \Sigma_Y$ and $\chi_{i,j} = \mathbb{1}\left\{\left\|\widetilde{Y}_{i,j}\right\|_2 \le R\right\}$, which are valid in this proof only. Then we have that $|\widetilde{J}| = \sum_{i\neq j}(1-\chi_{i,j})$, and by Markov's inequality,

$$P(\chi_{i,j}=0) \le \frac{\mathbb{E}\left[\left\|\widetilde{Y}_{i,j}\right\|_2^2\right]}{R^2} = \frac{\mathrm{tr}\,(\Sigma)}{R^2}. \tag{A.83}$$

By the finite difference inequality, we have that for $0<\tau<1$,

$$P\left(\frac{1}{N}\sum_{i\neq j}(1-\chi_{i,j}) \ge (1+\tau)\frac{\mathrm{tr}\,(\Sigma)}{R^2}\right) \le \exp\left\{\frac{-\tau^2 k_0 \mathrm{tr}\,(\Sigma)}{3R^2}\right\}.$$

Setting $\tau = R\sqrt{\frac{3t}{k_0 \mathrm{tr}\,(\Sigma)}}$ and assuming that $R\sqrt{\frac{3t}{k_0 \mathrm{tr}\,(\Sigma)}} < 1$, we see that

$$\epsilon = \frac{|\widetilde{J}|}{N} \le \frac{\mathrm{tr}\,(\Sigma)}{R^2} + \frac{\sqrt{\mathrm{tr}\,(\Sigma)}}{R}\sqrt{\frac{3t}{k_0}}$$

with probability at least $1 - e^{-t}$. Note that when R is chosen as

$$R = \left(\frac{\mathrm{tr}\,(\Sigma)\,\|\Sigma\|\,n}{\log\big(n \cdot \mathrm{rk}(\Sigma)\big)}\right)^{\frac{1}{4}},$$

the assumption $R\sqrt{\frac{3t}{k_0 \mathrm{tr}\,(\Sigma)}} < 1$ is equivalent to

$$\frac{\mathrm{tr}\,(\Sigma)\,\|\Sigma\|\,n}{\log\big(n \cdot \mathrm{rk}(\Sigma)\big)} < \left(\frac{k_0 \mathrm{tr}\,(\Sigma)}{3t}\right)^2 \le \left(\frac{n \mathrm{tr}\,(\Sigma)}{3t}\right)^2,$$

which is valid as long as $n \ge \frac{9t^2}{\log(n)}$. With this choice of R, we have that

$$\epsilon \le \frac{\mathrm{tr}\,(\Sigma)}{R^2}\left(1 + R\sqrt{\frac{3t}{k_0 \mathrm{tr}\,(\Sigma)}}\right) < \frac{2\mathrm{tr}\,(\Sigma)}{R^2} = 2\sqrt{\frac{\mathrm{rk}(\Sigma) \cdot \log\big(n \cdot \mathrm{rk}(\Sigma)\big)}{n}} \tag{A.84}$$

with probability at least $1 - e^{-t}$.

Moreover, when $\widetilde{Y}_{i,j}$ satisfies $L_4 - L_2$ norm equivalence with constant K, we can improve the bound in (A.83) to $K^4 \mathrm{tr}\,(\Sigma)^2/R^4$. By finite difference inequality again, we have that for $0 < \tau < 1$,

$$P\left(\frac{1}{N}\sum_{i \ne j}(1 - \chi_{i,j}) \ge (1+\tau)K^4\frac{\mathrm{tr}\,(\Sigma)^2}{R^4}\right) \le \exp\left\{\frac{-\tau^2 k_0 K^4 \mathrm{tr}\,(\Sigma)^2}{3R^4}\right\}.$$

Assuming that $R^2\sqrt{\frac{3t}{k_0 \mathrm{tr}\,(\Sigma)^2 K^4}} < 1$, or equivalently

$$\frac{\mathrm{tr}\,(\Sigma)\,\|\Sigma\|\,n}{\log\big(n \cdot \mathrm{rk}(\Sigma)\big)} < \frac{n \mathrm{tr}\,(\Sigma)^2 L^4}{3t},$$

we can set $\tau = R^2\sqrt{\frac{3t}{k_0 \mathrm{tr}\,(\Sigma)^2 K^4}}$ and derive the following improved bound:

$$\epsilon \leq 2K^4 \frac{\mathrm{rk}(\Sigma) \cdot \log\big(n \cdot \mathrm{rk}(\Sigma)\big)}{n}, \tag{A.85}$$

which holds with probability at least $1 - e^{-t}$. Note that the assumption above is valid when $K^4 \mathrm{rk}(\Sigma) \log\big(n \cdot \mathrm{rk}(\Sigma)\big) > 3t$, which requires the order of t to be at most $\log(n)$.

A.4.2 Proof of Theorem 6

In this section we present the proof of Theorem 6, which gives an improved error bound in the Frobenius norm for heavy-tailed data. The main idea making the improvement possible is the fact that for heavy-tailed data, the "outliers" $\widetilde{V}_{i,j}$ are nonzero if and only if the "well-behaved" term $\widetilde{Z}_{i,j}$ equals zero. We will repeat parts of the proof of Theorem 2 using this fact along with the inequality of Theorem 5 instead of the inequality $\|A\| \leq \|A\|_F$ to derive an improved bound.

We start with some notations, which are specific to this proof. Let $N = n(n-1)$, and $c(K)$ be a constant depending on K only, which can vary from step to step. Consider the events

$$\mathcal{E}_1 = \left\{\epsilon \leq c(K) \frac{\mathrm{rk}(\Sigma_Y) \log\big(n \cdot \mathrm{rk}(\Sigma_Y)\big)}{n}\right\},$$

$$\mathcal{E}_2 = \left\{\left\|\widehat{S} - \Sigma_Y\right\| \leq c(K) \left\|\Sigma_Y\right\| \sqrt{\frac{\mathrm{rk}(\Sigma_Y) \log(n)^3}{n}}\right\},$$

and

$$\mathcal{E} = \left\{\lambda_1 \geq \frac{140 \left\|\Sigma_Z\right\|}{\sqrt{n(n-1)}} \sqrt{\mathrm{rk}(\Sigma_Z)} + 4 \left\| \frac{1}{n(n-1)} \sum_{i \neq j} \widetilde{Z}_{i,j} \widetilde{Z}_{i,j}^T - \Sigma_Z \right\|,\right.$$

$$\lambda_2 \geq \frac{140\,\|\Sigma_Z\|}{n(n-1)}\sqrt{\mathrm{rk}(\Sigma_Z)} + 4\frac{1}{\sqrt{n(n-1)}}\max_{i\neq j}\left\|\widetilde{Z}_{i,j}\widetilde{Z}_{i,j}^T - \Sigma_Z\right\| \Bigg\}.$$

We will need to condition on these three events throughout the proof, so we will first estimate their probabilities.

1. In the view of Lemma 2, we have that $P(\mathcal{E}_1) \geq 1 - \frac{1}{n}$.

2. For $\mathcal{E}$, we need to choose λ_1 and λ_2 appropriately in order to guarantee that $\mathcal{E}$ happens with high probability. Since $\left\|\widetilde{Z}_{i,j}\right\|_2 \leq R$ almost surely, we can invoke the following version of matrix Bernstein inequality, which is a corollary of Theorem 3.1 from Minsker (2017),

Theorem 12. *Let $Z_1, \ldots, Z_n$ be i.i.d. random vectors with $\mathbb{E}[Z_1] = 0$ and $\|Z_1\|_2 \leq R$ almost surely. Denote $\Sigma_Z = \mathbb{E}\left[Z_1 Z_1^T\right]$ and $B = \mathbb{E}\left[(Z_1 Z_1^T)^2\right]$, then for $t \geq \frac{9n\|B\|}{16R^4}$,*

$$\left\|\frac{1}{n}\sum_i Z_i Z_i^T - \Sigma_Z\right\| \leq C\left(\sqrt{\frac{\|B\|\left(\log(\mathrm{rk}(B)) + t\right)}{n}} \vee \frac{R^2\left(\log(\mathrm{rk}(B)) + t\right)}{n}\right)$$

with probability at least $1 - e^{-t}$.

Following the same argument as in Section A.2.1, we can derive the following corollary for the transformed data:

Corollary 5. *Let $\widetilde{Z}_{i,j}$ be defined as in (4.18), namely, $\widehat{Z} = \mathbb{1}\left\{\left\|\widetilde{Y}_{i,j}\right\|_2 \leq R\right\}$. Then*

$$\left\|\frac{1}{N}\sum_{i\neq j}\widetilde{Z}_{i,j}\widetilde{Z}_{i,j}^T - \Sigma_Z\right\| \leq C\left(\sqrt{\frac{\|B\|\left(\log(\mathrm{rk}(B)) + t\right)}{n}} \vee \frac{R^2\left(\log(\mathrm{rk}(B)) + t\right)}{n}\right) \quad \text{(A.86)}$$

with probability at least $1 - 2e^{-t}$, where $B = \mathbb{E}\left[(\widetilde{Z}_{i,j}\widetilde{Z}_{i,j}^T)^2\right]$.

It remains to estimate $\|B\|$ and $\mathrm{rk}(B)$. Mendelson and Zhivotovskiy (2020) showed that if Y satisfies an $L_4 - L_2$ norm equivalence with constant K, then

$$c\,\|\Sigma_Z\|\,\mathrm{tr}\,(\Sigma_Z) \le \|B\| \le c(K)\,\|\Sigma_Y\|\,\mathrm{tr}\,(\Sigma_Y)$$

and

$$\mathrm{tr}\,(B) \le c(K)\mathrm{tr}\,(\Sigma_Y)^2.$$

Combining these two bounds, we have that

$$\mathrm{rk}(B) = \frac{\mathrm{tr}\,(B)}{\|B\|} \le c(K)\frac{\mathrm{tr}\,(\Sigma_Y)^2}{\|\Sigma_Z\|\,\mathrm{tr}\,(\Sigma_Z)}. \tag{A.87}$$

On the other hand, we have the following lemma which guarantees that Σ_Z is close to Σ_Y.

Lemma 9. *Let $Y \in \mathbb{R}^d$ be a mean zero random vector satisfying the $L_4 - L_2$ norm equivalence with constant K. Then*

$$\|\Sigma_Z - \Sigma_Y\| \le c(K)\frac{\|\Sigma_Y\|\,tr\,(\Sigma_Y)}{R^2} = c(K)\frac{\|\Sigma_Y\|^2\,\mathrm{rk}(\Sigma_Y)}{R^2}, \tag{A.88}$$

$$|tr\,(\Sigma_Z) - tr\,(\Sigma_Y)| \le c(K)\frac{tr^2(\Sigma_Y)}{R^2} = c(K)\frac{\|\Sigma_Y\|^2\,\mathrm{rk}(\Sigma_Y)^2}{R^2}, \tag{A.89}$$

$$\|\Sigma_Z - \Sigma_Y\|_F \le c(K)\frac{\|\Sigma_Y\|^{\frac{1}{2}}\,tr\,(\Sigma_Y)^{\frac{3}{2}}}{R^2} = c(K)\frac{\|\Sigma_Y\|^2\,\mathrm{rk}(\Sigma_Y)^{\frac{3}{2}}}{R^2}, \tag{A.90}$$

where $\Sigma_Y = \mathbb{E}\big[YY^T\big]$, $Z = Y\mathbb{1}\{\|Y\|_2 \le R\}$, $\Sigma_Z = \mathbb{E}\big[ZZ^T\big]$, and $c(K)$ is a constant depending only on K.

The proof of Lemma 9 is presented in section A.6.2 of the supplementary material. In particular, it implies that both $\|\Sigma_Z\|$ and $\operatorname{tr}(\Sigma_Z)$ are equivalent up to a multiplicative constant factor to $\|\Sigma_Y\|$ and $\operatorname{tr}(\Sigma_Y)$ respectively, as long as $R \geq c(K)\sqrt{\operatorname{tr}(\Sigma_Y)}$. The condition is valid given that $n \geq c(K)\operatorname{rk}(\Sigma_Y)\Big[\log\big(\operatorname{rk}(\Sigma_Y)\big) + \log(n)\Big]$, and hence by (A.87),

$$\operatorname{rk}(B) \leq c(K)\operatorname{rk}(\Sigma_Y).$$

Combining the bounds on $\|B\|$ and $\operatorname{rk}(B)$ with Corollary 5, the choice of R as

$$R = \left(\frac{\operatorname{tr}(\Sigma_Y)\|\Sigma_Y\| n}{\log\big(n \cdot \operatorname{rk}(\Sigma_Y)\big)}\right)^{\frac{1}{4}},$$

and the choice of $t = \log(n)$, we deduce that

$$\left\| \frac{1}{n(n-1)} \sum_{i \neq j} \widetilde{Z}_{i,j} \widetilde{Z}_{i,j}^T - \Sigma_Z \right\| \leq c(K)\|\Sigma_Y\| \sqrt{\frac{\operatorname{rk}(\Sigma_Y)\left[\log\big(n \cdot \operatorname{rk}(\Sigma_Y)\big)\right]}{n}} \tag{A.91}$$

with probability at least $1 - \frac{2}{n}$. Similarly, applying Theorem 12 to each single point $\widetilde{Z}_{i,j}$ and proceeding in a similar way in Section A.2.2, we deduce that

$$\max_{i \neq j} \left\| \widetilde{Z}_{i,j} \widetilde{Z}_{i,j}^T - \Sigma_Z \right\| \leq c(K)\|\Sigma_Y\| \sqrt{n \cdot \operatorname{rk}(\Sigma_Y)\left[\log\big(n \cdot \operatorname{rk}(\Sigma_Y)\big)\right]} \tag{A.92}$$

with probability at least $1 - \frac{1}{n}$. It follows that with the choices of

$$\lambda_1 \geq c(K)\|\Sigma_Y\| \sqrt{\frac{\operatorname{rk}(\Sigma_Y)\left[\log\big(n \cdot \operatorname{rk}(\Sigma_Y)\big)\right]}{n}} \tag{A.93}$$

and

$$\lambda_2 \geq c(K)\|\Sigma_Y\| \sqrt{\frac{\operatorname{rk}(\Sigma_Y)\left[\log\big(n \cdot \operatorname{rk}(\Sigma_Y)\big)\right]}{n}}, \tag{A.94}$$

we have that $P(\mathcal{E}) \geq 1 - \frac{3}{n}$.

3. To estimate the probability of the event $\mathcal{E}_2$, we first state a modified version of Theorem 5.

Remark 9. Following the same argument as in the proof of Theorem 5, we can show that for $A \geq 1$, with the choice of $\lambda_1 \leq c(K) \|\Sigma_Y\| \sqrt{n \log(n) \mathrm{rk}(\Sigma_Y)}$, $\lambda_2 = c(K) \|\Sigma_Y\| \sqrt{\frac{\mathrm{rk}(\Sigma_Y) \log(n)}{An}}$ and under the assumptions that

$$r_H \frac{\log(n)}{n} \left(1 + 2c(1-\alpha)^{-1}\right) \leq \frac{1}{1280},$$

the following inequality holds with probability at least $1 - (\frac{8}{3} r_H + 1) \frac{1}{n^A}$:

$$\left\| \widehat{S}_\lambda - \Sigma_Y \right\| \leq \frac{20}{39} \lambda_1 + \frac{40}{13} c(K) \|\Sigma_Y\| \sqrt{\frac{\mathrm{rk}(\Sigma_Y) A \log(n)^3}{n}}. \tag{A.95}$$

Applying Remark 9 with $\lambda_1 = c(K) \|\Sigma_Y\| \sqrt{\frac{\mathrm{rk}(\Sigma_Y) \left[\log\left(n \cdot \mathrm{rk}(\Sigma_Y)\right)\right]}{n}}$, $\lambda_2 = c(K) \|\Sigma_Y\| \sqrt{\frac{\mathrm{rk}(\Sigma_Y) \log(n)}{An}}$ and the assumption $n \geq c(K) \mathrm{rk}(\Sigma_Y) \left[\log\left(n \cdot \mathrm{rk}(\Sigma_Y)\right)\right]$, we get that

$$\left\| \widehat{S} - \Sigma_Y \right\| \leq c(K) \|\Sigma_Y\| \sqrt{\frac{\mathrm{rk}(\Sigma_Y) A \log(n)^3}{n}} \tag{A.96}$$

with probability at least $1 - \frac{(\frac{8}{3} r_H + 1)}{n^A}$. This confirms that $P(\mathcal{E}_2) \geq 1 - \frac{(\frac{8}{3} r_H + 1)}{n^A}$. Note that the choices of λ_1 and λ_2 coincide with (A.93) and (A.94).

For what follows, we will condition on the events $\mathcal{E}$, $\mathcal{E}_1$ and $\mathcal{E}_2$. Repeating parts of the argument in Section A.2, we can arrive at the inequality

$$\left\| \Sigma_Z - \widehat{S}_\lambda \right\|_F^2 + \left\| S - \widehat{S}_\lambda \right\|_F^2 \leq \|\Sigma_Z - S\|_F^2 + \frac{1}{8} \left\| \widehat{S}_\lambda - S \right\|_F^2 + 2\lambda_1^2 \,\mathrm{rank}(S)(\sqrt{2}+1)^2$$

$$+\frac{2}{\sqrt{N}}\sum_{i\neq j}\left\langle \widetilde{U}^*_{i,j}-\widehat{U}_{i,j},\widehat{S}_\lambda - S\right\rangle. \quad \text{(A.97)}$$

By Lemma 9 and choosing R as

$$R=\left(\frac{\operatorname{tr}\left(\Sigma_Y\right)\|\Sigma_Y\|\,n}{\log\left(n\cdot \operatorname{rk}(\Sigma_Y)\right)}\right)^{\frac{1}{4}}, \quad \text{(A.98)}$$

we have that

$$\|\Sigma_Z-\Sigma_Y\|_F\leq c(K)\,\|\Sigma_Y\|\,\frac{\operatorname{rk}(\Sigma_Y)\sqrt{\log\left(\operatorname{rk}(\Sigma_Y)\right)+\log(n)}}{\sqrt{n}}.$$

Therefore, we can deduce from (A.97) that

$$\left\|\Sigma_Y-\widehat{S}_\lambda\right\|_F^2+\left\|S-\widehat{S}_\lambda\right\|_F^2\leq \|\Sigma_Z-S\|_F^2+\frac{1}{8}\left\|\widehat{S}_\lambda-S\right\|_F^2+2\lambda_1^2\operatorname{rank}(S)(\sqrt{2}+1)^2$$
$$+\frac{2}{\sqrt{N}}\sum_{i\neq j}\left\langle \widetilde{U}^*_{i,j}-\widehat{U}_{i,j},\widehat{S}_\lambda - S\right\rangle+c(L)\,\|\Sigma_Y\|^2\,\frac{\operatorname{rk}(\Sigma_Y)^2\log\left(n\cdot\operatorname{rk}(\Sigma_Y)\right)}{n}. \quad \text{(A.99)}$$

It remains to bound the expression

$$\frac{2}{\sqrt{N}}\sum_{i\neq j}\left\langle \widetilde{U}^*_{i,j}-\widehat{U}_{i,j},\widehat{S}_\lambda - S\right\rangle.$$

First, note that

$$\sum_{(i,j)\notin\widetilde{J}}\left\|\widehat{U}_{i,j}-\widetilde{U}^*_{i,j}\right\|_1=\sum_{(i,j)\notin\widetilde{J}}\left\|P_{L^\perp_{i,j}}\widehat{U}_{i,j}P_{L^\perp_{i,j}}\right\|_1$$

and that

$$\sum_{(i,j)\in\widetilde{J}}\left\|\widehat{U}_{i,j}-\widetilde{U}^*_{i,j}\right\|_1\leq\sum_{(i,j)\in\widetilde{J}}\left\|\mathcal{P}_{L_{i,j}}(\widehat{U}_{i,j}-\widetilde{U}^*_{i,j})\right\|_1+\sum_{(i,j)\in\widetilde{J}}\left\|P_{L^\perp_{i,j}}\widehat{U}_{i,j}P_{L^\perp_{i,j}}\right\|_1.$$

By Lemma 4, we have that

$$\sum_{i\neq j}\left\|\widehat{U}_{i,j}-\widetilde{U}^*_{i,j}\right\|_1 \leq \sum_{(i,j)\in\widetilde{J}}\left\|\mathcal{P}_{L_{i,j}}(\widehat{U}_{i,j}-\widetilde{U}^*_{i,j})\right\|_1+\sum_{i\neq j}\left\|P_{L^{\perp}_{i,j}}\widehat{U}_{i,j}P_{L^{\perp}_{i,j}}\right\|_1$$
$$\leq \sum_{(i,j)\in\widetilde{J}}\left\|\mathcal{P}_{L_{i,j}}(\widehat{U}_{i,j}-\widetilde{U}^*_{i,j})\right\|_1+3\left(\frac{\lambda_1}{\lambda_2}\left\|\mathcal{P}_{L(k)}(\widehat{S}-\Sigma(k))\right\|_1+\sum_{(i,j)\in\widetilde{J}}\left\|\mathcal{P}_{L_{i,j}}(\widehat{U}_{i,j}-\widetilde{U}^*_{i,j})\right\|_1\right)$$
$$\leq 4\sum_{(i,j)\in\widetilde{J}}\left\|\mathcal{P}_{L_{i,j}}(\widehat{U}_{i,j}-\widetilde{U}^*_{i,j})\right\|_1+3\frac{\lambda_1}{\lambda_2}\left\|\mathcal{P}_{L(k)}(\widehat{S}-\Sigma(k))\right\|_1$$

Repeating the argument behind (A.60), we have that

$$\frac{2}{\sqrt{N}}\sum_{i\neq j}\left\langle\widetilde{U}^*_{i,j}-\widehat{U}_{i,j},\widehat{S}-S\right\rangle \leq \frac{2\left\|\widehat{S}-S\right\|}{\sqrt{N}}\sum_{i\neq j}\left\|\widetilde{U}^*_{i,j}-\widehat{U}_{i,j}\right\|_1$$
$$\leq \frac{2\left\|\widehat{S}-S\right\|}{\sqrt{N}}\left(4\sum_{(i,j)\in\widetilde{J}}\left\|\mathcal{P}_{L_{i,j}}(\widehat{U}_{i,j}-\widetilde{U}^*_{i,j})\right\|_1+3\frac{\lambda_1}{\lambda_2}\left\|\mathcal{P}_{L(k)}(\widehat{S}-\Sigma(k))\right\|_1\right)$$
$$\leq \frac{2\left\|\widehat{S}-S\right\|}{\sqrt{N}}\left(3\sqrt{2k}\frac{\lambda_1}{\lambda_2}\left\|\widehat{S}-\Sigma_Z\right\|_F+8\sum_{(i,j)\in\widetilde{J}}\left\|\widehat{U}_{i,j}-\widetilde{U}_{i,j}\right\|_F\right)$$
$$\leq 6\left\|\widehat{S}-S\right\|\frac{\lambda_1}{\lambda_2}\sqrt{\frac{2k}{N}}\left\|\widehat{S}-\Sigma_Z\right\|_F+16\left\|\widehat{S}-S\right\|\sqrt{\frac{|\widetilde{J}|}{N}}\sqrt{\sum_{(i,j)\in\widetilde{J}}\left\|\widehat{U}_{i,j}-\widehat{U}^*_{i,j}\right\|_F^2}. \tag{A.100}$$

We will estimate the two terms on the right-hand side of the above inequality one by one. Note that we did not apply the crude bound $\left\|\widehat{S}-S\right\|\leq\left\|\widehat{S}-S\right\|_F$ since $\left\|\widehat{S}-S\right\|$ is strictly smaller for the heavy tailed data due to the independence of the "outliers". By triangle inequality, Lemma 9 and the choice of $R=\left(\frac{\operatorname{tr}(\Sigma)\|\Sigma\|n}{\log\left(n\cdot\operatorname{rk}(\Sigma)\right)}\right)^{\frac{1}{4}}$, we have that on the event $\mathcal{E}_2$,

$$\left\|\widehat{S}_\lambda-S\right\|\leq\left\|\widehat{S}_\lambda-\Sigma_Y\right\|+\|\Sigma_Y-\Sigma_Z\|+\|\Sigma_Z-S\|$$
$$\leq c(L)\left(\|\Sigma_Y\|\sqrt{\frac{\operatorname{rk}(\Sigma_Y)A\log(n)^3}{n}}+\frac{\|\Sigma_Y\|\operatorname{tr}(\Sigma_Y)}{R^2}\right)+\|\Sigma_Z-S\|$$
$$\leq \underbrace{c(L)\|\Sigma_Y\|\sqrt{\frac{\operatorname{rk}(\Sigma_Y)A\log(n)^3}{n}}}_{:=\mathrm{I}}+\|\Sigma_Z-S\|. \tag{A.101}$$

Note that the term I is of the order $\sqrt{\frac{\mathrm{rk}(\Sigma_Y)}{n}}$, up to the logarithmic factors. For what follows, we set $S = \Sigma_Y$, and (A.101) implies that $\left\|\widehat{S}_\lambda - S\right\| = \left\|\widehat{S}_\lambda - \Sigma_Y\right\| \leq \mathrm{I}$. To estimate $\sqrt{\sum_{\widetilde{J}} \left\|\widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\|_F^2}$, we can apply inequality (A.49) which entails that

$$\sqrt{\sum_{\widetilde{J}} \left\|\widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\|_F^2} \leq 2\sqrt{2}\Bigg(\|\Sigma_Z - \Sigma_Y\|_F + \sqrt{\mathrm{rank}(S)}\lambda_1(\sqrt{2}+1) + \lambda_2 \frac{(4/3+\sqrt{2})}{\sqrt{2}}\sqrt{|\widetilde{J}|} + \sqrt{2(6\sqrt{2}+6)\frac{\lambda_1}{\lambda_2}} \left(\frac{\mathrm{rank}(\Sigma_Y)}{N}\right)^{\frac{1}{4}} \left\|\widehat{S}_\lambda - \Sigma_Y\right\|_F \Bigg),$$

given that $k = \lfloor \frac{N\lambda_2^2}{1200\lambda_1^2} \rfloor$ and $|\widetilde{J}| \leq \frac{N}{6400}$. For simplicity, we denote $B = \sqrt{2(6\sqrt{2}+6)\frac{\lambda_1}{\lambda_2}} \left(\frac{\mathrm{rank}(\Sigma_Y)}{N}\right)^{\frac{1}{4}}$.

Now we will estimate the two terms in (A.100):

- First,

$$6\left\|\widehat{S}_\lambda - \Sigma_Y\right\| \frac{\lambda_1}{\lambda_2}\sqrt{\frac{2k}{N}} \left\|\widehat{S}_\lambda - \Sigma_Z\right\|_F \leq 6 \cdot \mathrm{I} \cdot \frac{\lambda_1}{\lambda_2}\sqrt{\frac{2k}{N}} \left\|\widehat{S}_\lambda - \Sigma_Z\right\|_F. \tag{A.102}$$

  This term is independent of the outliers, and a direct application of the inequality $2ab \leq a^2 + b^2$ gives that

$$6\left\|\widehat{S}_\lambda - \Sigma_Y\right\| \frac{\lambda_1}{\lambda_2}\sqrt{\frac{2k}{N}} \left\|\widehat{S}_\lambda - \Sigma_Z\right\|_F \leq 3 \cdot \frac{\lambda_1}{\lambda_2}\sqrt{\frac{2k}{N}} \left(\mathrm{I}^2 + \left\|\widehat{S}_\lambda - \Sigma_Z\right\|_F^2\right). \tag{A.103}$$

- Second,

$$16\left\|\widehat{S}_\lambda - \Sigma_Y\right\| \sqrt{\frac{|\widetilde{J}|}{N}} \sqrt{\sum_{(i,j)\in\widetilde{J}} \left\|\widehat{U}_{i,j} - \widetilde{U}^*_{i,j}\right\|_F^2}$$
$$\leq 16\left\|\widehat{S}_\lambda - \Sigma_Y\right\| \sqrt{\frac{|\widetilde{J}|}{N}} \cdot 2\sqrt{2}\Bigg(\|\Sigma_Z - \Sigma_Y\|_F + \sqrt{\mathrm{rank}(\Sigma_Y)}\lambda_1(\sqrt{2}+1) + \lambda_2 \frac{(4/3+\sqrt{2})}{\sqrt{2}}\sqrt{|\widetilde{J}|} + B\left\|\widehat{S}_\lambda - \Sigma_Y\right\|_F \Bigg)$$

$$\leq 16\sqrt{2}\sqrt{\frac{|\widetilde{J}|}{N}}\left(\left\|\Sigma_Y - \widehat{S}_\lambda\right\|^2 + \|\Sigma_Z - \Sigma_Y\|_F^2\right) + 8(\sqrt{2}+1)^2\frac{|\widetilde{J}|}{N}\left\|\Sigma_Y - \widehat{S}_\lambda\right\|^2$$
$$+ 64\lambda_1^2 \operatorname{rank}(\Sigma_Y) + 32(4/3+\sqrt{2})\cdot \mathrm{I}\cdot\lambda_2\sqrt{\frac{|\widetilde{J}|^2}{N}}$$
$$+ 16\sqrt{2}\left(\sqrt{n}\cdot \mathrm{I}^2\cdot B^2 + \frac{|\widetilde{J}|}{N}\cdot\frac{1}{\sqrt{n}}\left\|\widehat{S}_\lambda - \Sigma_Y\right\|_F^2\right). \quad \text{(A.104)}$$

Combining (A.99, A.100, A.103,A.104), and assuming that $6\cdot\frac{\lambda_1}{\lambda_2}\sqrt{\frac{2k}{N}} + 96\sqrt{2}\sqrt{\frac{|\widetilde{J}|}{N}} \leq \delta \leq \frac{3}{8}$ and $B^2 \leq \frac{\delta\sqrt{A}}{16\sqrt{2}n}$, we deduce that

$$(1-\delta)\left\|\Sigma_Y - \widehat{S}_\lambda\right\|_F^2 \leq c(\delta)76\lambda_1^2 \operatorname{rank}(\Sigma_Y) + \delta\cdot \mathrm{I}^2 + 32(4/3+\sqrt{2})\cdot \mathrm{I}\cdot\lambda_2\epsilon n$$
$$+ c(L,\delta)\|\Sigma_Y\|^2\frac{\mathrm{rk}(\Sigma_Y)^2 \log\left(n\cdot \mathrm{rk}(\Sigma_Y)\right)}{n}, \quad \text{(A.105)}$$

where $\epsilon = \frac{|\widetilde{J}|}{N}$ is the proportion of outliers. Finally, we recall that the choices of λ_1 and λ_2 are

$$\lambda_1 = c(K)\|\Sigma_Y\|\sqrt{\frac{\mathrm{rk}(\Sigma_Y)\left[\log\left(n\cdot \mathrm{rk}(\Sigma_Y)\right)\right]}{n}} \quad \text{(A.106)}$$

and

$$\lambda_2 = c(K)\|\Sigma_Y\|\sqrt{\frac{\mathrm{rk}(\Sigma_Y)\log(n)}{An}}. \quad \text{(A.107)}$$

Also, recall the definition of I in (A.101):

$$\mathrm{I} = c(K)\|\Sigma_Y\|\sqrt{\frac{\mathrm{rk}(\Sigma_Y)A\log(n)^3}{n}}. \quad \text{(A.108)}$$

Combining the equations (A.106, A.106, A.108) with (A.105), we derive that

$$\left\|\Sigma_Y - \widehat{S}_\lambda\right\|_F^2$$
$$\leq c(K,\delta)\|\Sigma_Y\|^2\frac{\mathrm{rk}(\Sigma_Y)\log\left(n\cdot \mathrm{rk}(\Sigma_Y)\right)}{n}\operatorname{rank}(\Sigma_Y) + c(K,\delta)\|\Sigma_Y\|^2\frac{\mathrm{rk}(\Sigma_Y)A\log(n)^3}{n}$$

$$
\begin{aligned}
&+ c(K,\delta)\epsilon \cdot n \cdot \|\Sigma_Y\|^2 \frac{\mathrm{rk}(\Sigma_Y)\log(n)^2}{n} + c(K,\delta)\|\Sigma_Y\|^2 \frac{\mathrm{rk}(\Sigma_Y)^2 \log\big(n\cdot \mathrm{rk}(\Sigma_Y)\big)}{n} \\
\le\ & c(K,\delta)\|\Sigma_Y\|^2 \frac{\mathrm{rk}(\Sigma_Y)\log\big(n\cdot \mathrm{rk}(\Sigma_Y)\big)}{n}\mathrm{rank}(\Sigma_Y) + c(K,\delta)\|\Sigma_Y\|^2 \frac{\mathrm{rk}(\Sigma_Y)^2 A \log(n)^3}{n}
\end{aligned} \tag{A.109}
$$

under the assumptions that $B^2 \le \frac{\delta\sqrt{A}}{16\sqrt{2n}}$ and $n \ge c(K)\mathrm{rk}(\Sigma_Y)\big[\log\big(n\cdot\mathrm{rk}(\Sigma_Y)\big)\big]$, where the last step in (A.109) follows from Lemma 2. Note that the assumption $B^2 \le \frac{\delta\sqrt{A}}{16\sqrt{2n}}$ is valid as long as $\mathrm{rank}(\Sigma_Y) \le c_1\delta^2 \cdot n\frac{A\lambda_2^2}{\lambda_1^2}$ for any constant $c_1 \le \frac{1}{4(6\sqrt{2}+6)^2}$. Finally, by the union bound over the events $\mathcal{E}$, $\mathcal{E}_1$ and $\mathcal{E}_2$, inequality (A.109) will hold with probability at least $1 - \frac{\left(\frac{8}{3}r_H+1\right)}{n^A} - \frac{4}{n}$. To this end, note that the condition $\mathrm{rank}(\Sigma_Y) \le c_1\delta^2 \cdot n\frac{A\lambda_2^2}{\lambda_1^2}$ is equivalent to

$$
\mathrm{rank}(\Sigma_Y) \le c(K)\cdot n \cdot \frac{\log(n)}{\log\big(n\cdot\mathrm{rk}(\Sigma_Y)\big)}
$$

when λ_1, λ_2 are chosen as (A.106) and (A.107) respectively. The upper bound on $\mathrm{rank}(\Sigma_Y)$ is in the order of n up to logarithmic factors.

A.5 Proofs ommitted in Section 5

A.5.1 Convergence analysis of the proximal gradient method (Theorem 7)

In this section we present the convergence analysis of the proximal gradient method (with matrix variables). It is worth noting that our analysis follows the argument in Beck (2017, Chapter 10). Recall that our loss function can be written in the form $L(S) = g(S) + h(S)$, where h is convex, and g is the average of N functions $g_{i,j}(S) = \mathrm{tr}\left(\rho_{\frac{\sqrt{N}\lambda_2}{2}}(\widetilde{Y}_{i,j}\widetilde{Y}_{i,j}^T - S)\right)$. Note that

$\nabla g_{i,j}(S) = -\rho'_{\frac{\sqrt{N}\lambda_2}{2}}(\widetilde{Y}_{i,j}\widetilde{Y}_{i,j}^T - S)$, and using the fact from Bhatia (2013, Lemma VII.5.5) we have that $\nabla g_{i,j}(S)$ is Lipschitz in Frobenius norm with $L = 1$, i.e.

$$\left\|\nabla g_{i,j}(U) - \nabla g_{i,j}(V)\right\|_F \le L \left\|U - V\right\|_F,$$

hence $g(S)$ is also Lipschitz in Frobenius norm with $L = 1$. We have the following matrix form of the descent lemma:

Lemma 10. *Assume that $g(S)$ is Lipschitz in Frobenius norm with constant $L > 0$. Then*

$$g(S_2) \le g(S_1) + \langle \nabla g(S_1), S_2 - S_1 \rangle + \frac{L}{2}\left\|S_2 - S_1\right\|_F^2.$$

Proof. First, denote $U_t = S_1 + t(S_2 - S_1)$, we have that

$$g(S_2) = g(S_1) + \int_0^1 \langle \nabla g(U_t), S_2 - S_1 \rangle \, dt,$$

hence

$$\begin{aligned}\left|g(S_2) - g(S_1) - \langle \nabla g(S_1), S_2 - S_1 \rangle\right| &= \left|\int_0^1 \langle \nabla g(U_t) - \nabla g(S_1), S_2 - S_1 \rangle \, dt\right| \\ &\le \int_0^1 \left\|\nabla g(U_t) - \nabla g(S_1)\right\|_F \left\|S_2 - S_1\right\|_F dt \le \frac{L}{2}\left\|S_2 - S_1\right\|_F^2.\end{aligned}$$

□

Now recall that the proximal gradient descent algorithm update is

$$S^{t+1} = prox_{\alpha_t, h}(S^t - \alpha_t \nabla g(S^t)).$$

Set $G_\alpha(S) = \frac{1}{\alpha}\big(S - prox_{\alpha,h}(S - \alpha\nabla g(S))\big)$, then $S^{t+1} = S^t - \alpha_t G_{\alpha_t}(S^t)$. The following lemma guarantees that the PGD makes progress at each step.

Lemma 11. *Assume that $0 \leq \alpha_t \leq L$ for all $t = 1, 2, \ldots$, then for any symmetric matrix U,*

$$L(S^{t+1}) \leq L(U) + \left\langle G_{\alpha_t}(S^t), S^t - U \right\rangle - \frac{\alpha_t}{2} \left\| G_{\alpha_t}(S^t) \right\|_F^2 .$$

Proof. Since $g(\cdot)$ is convex, we have that for any symmetric matrix U,

$$g(U) \geq g(S^t) + \left\langle \nabla g(S^t), U - S^t \right\rangle .$$

Combining this with Lemma 10, we have that

$$\begin{aligned} g(S^{t+1}) &\leq g(S^t) + \left\langle \nabla g(S^t), S^{t+1} - S^t \right\rangle + \frac{\alpha_t}{2} \left\| G_{\alpha_t}(S^t) \right\|_F^2 \\ &\leq g(U) - \left\langle \nabla g(S^t), U - S^t \right\rangle + \left\langle \nabla g(S^t), S^{t+1} - S^t \right\rangle + \frac{\alpha_t}{2} \left\| G_{\alpha_t}(S^t) \right\|_F^2 \\ &= g(U) + \left\langle \nabla g(S^t), S^{t+1} - U \right\rangle + \frac{\alpha_t}{2} \left\| G_{\alpha_t}(S^t) \right\|_F^2 . \end{aligned}$$

Since $h(\cdot)$ is convex, for any $V \in \partial h(S^{t+1})$,

$$h(U) \geq h(S^{t+1}) + \left\langle V, U - S^{t+1} \right\rangle .$$

Recall that

$$S^{t+1} = \operatorname*{argmin}_S \left\{ h(S) + \frac{1}{2\alpha_t} \left\| S - (S^t - \alpha_t \nabla g(S^t)) \right\|_F^2 \right\}.$$

By the optimality conditions,

$$0 \in \partial h(S^{t+1}) + \frac{1}{\alpha_t} (S^{t+1} - S^t + \alpha_t \nabla g(S^t)).$$

Therefore,

$$G_{\alpha_t}(S^t) - \nabla g(S^t) \in \partial h(S^{t+1}),$$

and

$$L(S^{t+1}) \leq g(U) + h(U) + \left\langle \nabla g(S^t), S^{t+1} - U \right\rangle + \frac{\alpha_t}{2} \left\| G_{\alpha_t}(S^t) \right\|_F^2$$

$$+ \left\langle G_{\alpha_t}(S^t) - \nabla g(S^t), S^{t+1} - U \right\rangle$$

$$= L(U) + \left\langle G_{\alpha_t}(S^t), S^{t+1} - U \right\rangle + \frac{\alpha_t}{2} \left\| G_{\alpha_t}(S^t) \right\|_F^2$$

$$\leq L(U) + \left\langle G_{\alpha_t}(S^t), S^t - U \right\rangle - \frac{\alpha_t}{2} \left\| G_{\alpha_t}(S^t) \right\|_F^2,$$

where in the last step we used the fact that $S^{t+1} = S^t - \alpha_t G_{\alpha_t}(S^t)$. □

Taking $U = S^t$ in Lemma 11, we have that

$$L(S^{t+1}) \leq L(S^t) - \frac{\alpha_t}{2} \left\| G_{\alpha_t}(S^t) \right\|_F^2,$$

i.e. the PGD method is making progress at each iteration. Taking $U = S^*$ in Lemma 11, where S^* is the true minimizer of $L(S)$, we have that

$$L(S^{t+1}) - L(S^*) \leq \left\langle G_{\alpha_t}(S^t), S^t - S^* \right\rangle - \frac{\alpha_t}{2} \left\| G_{\alpha_t}(S^t) \right\|_F^2$$

$$= \frac{1}{2\alpha_t} \Big(\left\langle 2\alpha_t G_{\alpha_t}(S^t), S^t - S^* \right\rangle - \left\| \alpha_t G_{\alpha_t}(S^t) \right\|_F^2 \Big)$$

$$= \frac{1}{2\alpha_t} \Big(\left\| S^t - S^* \right\|_F^2 - \left\| \alpha_t G_{\alpha_t} - S^t + S^* \right\|_F^2 \Big)$$

$$= \frac{1}{2\alpha_t} \Big(\left\| S^t - S^* \right\|_F^2 - \left\| S^{t+1} - S^* \right\|_F^2 \Big).$$

Assuming that the step size is fixed (i.e. $\alpha_t = \alpha$) or diminishing (i.e. $\alpha_t \geq \alpha_{T+1} = \alpha$), summing up both sides of the above inequality for $t = 0, 1, \ldots, T$, and recalling that $L(S^{t+1}) \leq L(S^t)$, we have that

$$(T+1)(L(S^{t+1} - L(S^*)) \leq \frac{1}{2\alpha} \Big(\left\| S^0 - S^* \right\|_F^2 - \left\| S^{t+1} - S^* \right\|_F^2 \Big),$$

hence

$$L(S^{t+1}) - L^* \leq \frac{\left\|S^0 - S^*\right\|_F^2}{2\alpha(T+1)},$$

as desired. Note that the convergence rate can be improved to $\mathcal{O}(\frac{1}{T^2})$, see Nesterov (1983, 2003), Tseng (2008) for details.

A.5.2 Numerical method of updating eigenvalues (solving equation (5.28))

In this section, we present the numerical method introduced by Bunch et al. (1978) which computes the roots of $\omega_i(\mu) = 0$ for $i = 1, \ldots, k \leq d$, where $\omega_i(\mu)$ is defined as

$$\omega_i(\mu) = 1 + \sum_{j=1}^{k} \frac{\zeta_j^2}{\delta_j - \mu}$$

and $\delta_j = (d_j - d_i)/\rho$. Recall that the eigenvalues of $C = D + \rho z z^T$, denoted as $\widetilde{d}_1, \ldots, \widetilde{d}_k$, and the eigenvalues of D, denoted as $d_1, \ldots, d_k$, satisfy the identity $\widetilde{d}_i = d_i + \rho\mu_i$ with $\omega_i(\mu_i) = 0$. Therefore, it remains to solve equations $\omega_i(\mu) = 0$, $i = 1, \ldots, k$. Fix $i \in \{1, \ldots, k\}$, and define

$$\psi_i(t) = \sum_{j=1}^{i} \frac{\zeta_j^2}{\delta_j - t}, \quad i = 1, \ldots, k,$$

and

$$\phi_i(t) = \begin{cases} 0, \quad i = k, \\ \sum_{j=i+1}^{k} \frac{\zeta_j^2}{\delta_j - t}, \quad 1 \leq i < k. \end{cases}$$

It is clear that $\omega_i(t) = 1 + \psi_i(t) + \phi_i(t)$. Without loss of generality, we shall assume that $\rho > 0$; otherwise, we can replace d_i by $-d_{k-i+1}$ and ρ by $-\rho$. Also, we assume that $k > 1$; otherwise, we have the trivial case $\mu_1 = \zeta_1^2$. We will deal with the case $i < k$ and $i = k$ separately.

1. Assume that $i \in \{1, \ldots, k-1\}$ is fixed. We are seeking μ_i such that $0 < \mu_i < \min\{1 - \sum_{j=1}^{i-1}, \delta_{i+1}\}$ (by Theorem 8) and

$$-\psi_i(\mu_i) = \phi_i(\mu_i) + 1.$$

Assume that we have an approximation t_1 to the root μ_i with $0 < t_1 < \mu_i$, and we want to get an updated approximation t_2. As suggested by Bunch et al. (1978), we shall consider the local approximation to the rational functions ϕ_i and ψ_i at t_1, namely,

$$\frac{p_1}{q_1 - t_1} = \psi_i(t_1), \qquad \frac{p_1}{(q_1 - t_1)^2} = \psi_i'(t_1), \tag{A.110}$$

$$r_1 + \frac{s_1}{\delta - t_1} = \phi_i(t_1), \quad \frac{s_1}{(\delta - t_1)^2} = \phi_i'(t_1). \tag{A.111}$$

where $\delta = \delta_{i+1}$. It can be easily verified that p_1, q_1, r_1, s_1 satisfies

$$p_1 = \psi_i(t_1)^2/\psi_i'(t_1), \quad q_1 = t_1 + \psi_i(t_1)/\psi_i'(t_1), \tag{A.112}$$

$$r_1 = \phi_i(t_1) - (\delta - t_1)\phi_i'(t_1), \quad s = (\delta - t_1)^2\phi_i'(t_1). \tag{A.113}$$

The updated approximation t_2 is then obtained by solving the following equation:

$$-\frac{p_1}{q_1 - t_2} = 1 + r_1 + \frac{s_1}{\delta - t_2}. \tag{A.114}$$

Direct computation shows that

$$t_2 = t_1 + 2b/(a + \sqrt{a^2 - 4b}),$$

where

$$a = \frac{(\delta - t_1)(1 + \phi_i(t_1)) + \psi_i(t_1)^2/\psi_i'(t_1)}{c} + \psi_i(t_1)/\psi_i'(t_1),$$

$$b = \frac{(\delta - t_1) w \psi_i(t_1)}{\psi_i'(t_1) c},$$

$$c = 1 + \phi_i(t_1) - (\delta - t_1)\phi_i'(t_1),$$

$$w = 1 + \phi_i(t_1) + \psi_i(t_1).$$

The following theorem shows that the update (A.114) is guaranteed to converge to μ_i:

Theorem 13 (Bunch et al. (1978))**.** *Let $t_0 \in (0, \mu_i)$ and t_{j+1} be the solution of $-\frac{p_j}{q_j - t} = 1 + r_j + \frac{s_j}{\delta - t}$, $j \geq 0$, where p_j, q_j, r_j, s_j are defined by* (A.112)*. Then we have that $t_j < t_{j+1} < \mu_i$ and $\lim_{j\to\infty} = \mu_i$. Moreover, the rate of convergence is quadratic, meaning that for any j sufficiently large, $|t_{j+1} - \mu_i| \leq C|t_j - \mu_i|^2$, where C is an absolute constant independent of iteration.*

It remains to determine an initial guess t_0 such that $t_0 \in (0, \mu_i)$. Recall that $\omega_i(\mu_i) = 0$, which is equivalent to

$$1 + \sum_{j=1, j\neq i, i+1}^{k} \frac{\zeta_j^2}{\delta_j - \mu_i} + \frac{\zeta_{i+1}^2}{\delta_{i+1} - \mu_i} = \frac{\zeta_i^2}{\mu_i}.$$

Since $\mu_i < \delta_{i+1}$, we can define t_0 to be the positive solution of the equation

$$1 + \sum_{j=1, j\neq i, i+1}^{k} \frac{\zeta_j^2}{\delta_j - \delta_{i+1}} + \frac{\zeta_{i+1}^2}{\delta_{i+1} - t_0} = \frac{\zeta_i^2}{t_0}.$$

By monotonicity, we have that $t_0 \in (0, \mu_i)$, as desired.

2. Now we assume that $i = k$. In this case, $\phi_k(t) = 0$ and we want to solve the equation $-\psi_k(t) = 1$. Theorem 13 is still valid, and the update (A.114) can be simplified as

$$t_{j+1} = t_j + \left(\frac{1 + \psi_k(t_j)}{\psi_k'(t_j)}\right)\psi_k(t_j). \tag{A.115}$$

To choose $t_0 \in (0, \mu_k)$, we again recall that $\omega_k(\mu_k) = 0$, which is equivalent to

$$1 - \frac{\zeta_k^2}{\mu_k} + \sum_{j=1}^{k-1} \frac{\zeta_j^2}{\delta_j - \mu_k} = 0.$$

Since $\mu_k < 1$, we define t_0 to be the solution of

$$1 - \frac{\zeta_k^2}{t_0} + \sum_{j=1}^{k-1} \frac{\zeta_j^2}{\delta_j - 1} = 0.$$

By monotonicity, we have that $t_0 < \mu_k$. Moreover, note that $\sum_{j=1}^{k-1} \zeta_j^2 \leq \|z\|_2 = 1$ and $\delta_j < 0$, $\forall j = 1, \ldots, k-1$, so $1 + \sum_{j=1}^{k-1} \frac{\zeta_j^2}{\delta_j - 1} > 0$. Therefore, $t_0 \in (0, \mu_k)$, as desired.

A.6 Auxiliary technical results

A.6.1 Detailed derivation of the claim of Remark 1

In this section we present the detailed derivation of Remark 1. First, consider the function as follows

$$F(S, U_1, \cdots, U_n) := \frac{1}{2} \sum_{i=1}^{n} \left\| Y_i Y_i^T - S - U_i \right\|_F^2 + \lambda_1 \|S\|_1 + \lambda_2 \sum_{i=1}^{n} \|U_i\|_1 .$$

For a fixed matrix S, the matrix $Y_i Y_i^T - S$ has a spectral decomposition

$$Y_i Y_i^T - S = \sum_{j=1}^{d} \lambda_j^{(i)} v_j^{(i)} v_j^{(i)T}, \text{ for all } i = 1, \ldots, n,$$

where $\lambda_j^{(i)}$ is the j-th eigenvalue of $Y_i Y_i^T - S$ and $v_j^{(i)}$ is the corresponding eigenvector. We claim that $F(S, \cdot)$ can be minimized by choosing

$$\widetilde{U}_i = \sum_{j=1}^{d} \operatorname{sign}(\lambda_j^{(i)}) \left(|\lambda_j^{(i)}| - \lambda_2 \right)_+ v_j^{(i)} v_j^{(i)T} = \gamma_{\lambda_2}(Y_i Y_i^T - S), \tag{A.116}$$

where $\gamma_\lambda(u) := \text{sign}(u)\left(|u| - \lambda\right)_+, \ \forall u \in \mathbb{R}, \lambda \in \mathbb{R}^+$, and $(x)_+ := \max(x, 0)$. Indeed, note that $(U_1, \ldots, U_n) \mapsto G_S(U_1, \ldots, U_n) := F(S, U_1, \ldots, U_n)$ is strictly convex, so a sufficient and necessary condition for $(\widetilde{U}_1, \ldots, \widetilde{U}_n)$ to be a point of minimum is

$$\mathbf{0} \in \partial G_S(\widetilde{U}_1, \ldots, \widetilde{U}_n) = \left(-\left(Y_1 Y_1^T - S - U_1\right) + \lambda_2 \widetilde{V}_1, \ldots, -\left(Y_n Y_n^T - S - U_n\right) + \lambda_2 \widetilde{V}_n \right),$$

where $\widetilde{V}_i \in \partial \left\| \widetilde{U}_i \right\|, i = 1, \ldots, n$. By choosing $\widetilde{V}_i := \sum_{j:|\lambda_j^{(i)}| > \lambda_2} \text{sign}(\lambda_j^{(i)}) v_j^{(i)} v_j^{(i)T} + \sum_{j:|\lambda|_j^{(i)} \leq \lambda_2} \frac{\lambda_j^{(i)}}{\lambda_2} v_j^{(i)} v_j^{(i)T} \in \partial \left\| \widetilde{U}_i \right\|_1$, it is easy to verify that $\partial G_S(\widetilde{U}_1, \ldots, \widetilde{U}_n) = \mathbf{0}$, hence $(\widetilde{U}_1, \ldots, \widetilde{U}_n)$ is the minimizer. Plugging in to $F(S, U_1, \ldots, U_n)$, we get that

$$\begin{aligned}
F(S, \widetilde{U}_1, \cdots, \widetilde{U}_n) &= \frac{1}{2} \sum_{i=1}^{n} \left\| Y_i Y_i^T - S - \widetilde{U}_i \right\|_F^2 + \lambda_1 \|S\|_1 + \lambda_2 \sum_{i=1}^{n} \left\| \widetilde{U}_i \right\|_1 \\
&= \frac{1}{2} \sum_{i=1}^{n} \left\| \sum_{j=1}^{d} \left[\lambda_j^{(i)} - \gamma_{\lambda_2}(\lambda_j^{(i)}) \right] v_j^{(i)} v_j^{(i)T} \right\|_F^2 + \lambda_2 \sum_{i=1}^{n} \sum_{j=1}^{d} \gamma_{\lambda_2}(\lambda_j^{(i)}) + \lambda_1 \|S\|_1 \\
&= \sum_{i=1}^{n} \left(\sum_{j:|\lambda_j^{(i)}| > \lambda_2} (\lambda_2 |\lambda_j^{(i)}| - \frac{\lambda_2^2}{2}) + \sum_{j:|\lambda_j^{(i)}| \leq \lambda_2} \frac{\lambda_j^{(i)2}}{2} \right) + \lambda_1 \|S\|_1 \\
&= \text{tr} \left(\sum_{i=1}^{n} \rho_{\lambda_2}(Y_i Y_i^T - S) \right) + \lambda_1 \|S\|_1, \qquad \text{(A.117)}
\end{aligned}$$

where

$$\rho_\lambda(u) = \begin{cases} \frac{u^2}{2}, & |u| \leq \lambda \\ \lambda |u| - \frac{\lambda^2}{2}, & |u| > \lambda \end{cases}$$

is the Huber's loss function. Note that our loss function $L(S, \boldsymbol{U_{I_n^2}})$ can be expressed as

$$\begin{aligned}
L(S, \boldsymbol{U_{I_n^2}}) &= \frac{1}{N} \sum_{i \neq j} \left\| \widetilde{Y}_{i,j} \widetilde{Y}_{i,j}^T - S - \sqrt{N} U_{i,j} \right\|_F^2 + \lambda_1 \|S\|_1 + \lambda_2 \sum_{i \neq j} \|U_{i,j}\|_1 \\
&= \frac{2}{N} \left[\frac{1}{2} \sum_{i \neq j} \left\| \widetilde{Y}_{i,j} \widetilde{Y}_{i,j}^T - S - \sqrt{N} U_{i,j} \right\|_{\mathrm{F}}^2 + \frac{\sqrt{N}\lambda_2}{2} \sum_{i \neq j} \left\| \sqrt{N} U_{i,j} \right\|_1 \right] + \lambda_1 \|S\|_1.
\end{aligned}$$

Therefore, (A.117) implies that

$$\min_{S,\boldsymbol{U}_{\boldsymbol{I}_n^2}} L(S,\boldsymbol{U}_{\boldsymbol{I}_n^2}) = \min_S \min_{\boldsymbol{U}_{\boldsymbol{I}_n^2}} L(S,\boldsymbol{U}_{\boldsymbol{I}_n^2}) = \min_S L(S,\widetilde{\boldsymbol{U}}_{\boldsymbol{I}_n^2})$$
$$= \min_S \left\{ \frac{2}{N} \mathrm{tr} \left(\sum_{i \neq j} \rho_{\frac{\sqrt{N}\lambda_2}{2}} (\widetilde{Y}_{i,j}\widetilde{Y}_{i,j}^T - S) \right) + \lambda_1 \|S\|_1 \right\},$$

where

$$\widetilde{U}_{i,j} = \frac{1}{\sqrt{N}} \sum_{k=1}^d \mathrm{sign}(\lambda_k^{(i,j)}) \left(|\lambda_k^{(i,j)}| - \frac{\sqrt{N}\lambda_2}{2} \right)_+ v_k^{(i,j)} v_k^{(i,j)T} = \gamma_{\frac{\sqrt{N}\lambda_2}{2}} (\widetilde{Y}_{i,j}\widetilde{Y}_{i,j}^T - S)$$

with $\widetilde{Y}_{i,j}\widetilde{Y}_{i,j}^T - S = \sum_{k=1}^d \lambda_k^{(i,j)} v_k^{(i,j)} v_k^{(i,j)T}$, for all $i = 1, \ldots, n$.

A.6.2 Proof of Lemma 9

We denote $\Sigma_Y = \Sigma$, which is valid throughout this proof only. The proof of relations (A.88,A.89) was presented in Mendelson and Zhivotovskiy (2020, Lemma 2.1) with constants $c(K) = K^3$ and $c(K) = K^4$ respectively. For (A.90), assume that $\Sigma_Z - \Sigma$ has eigenvalues $\lambda_1 \leq \cdots \leq \lambda_d$ with corresponding orthonormal eigenvector set $\{u_1, \ldots, u_d\}$. Define $T_1 = 0$ and $T_j = \sum_{l=1}^{j-1} \lambda_l, j = 2, \ldots, d+1$. Then $\lambda_j = T_{j+1} - T_j$, and we have that

$$\|\Sigma_Z - \Sigma\|_F^2 = \sum_{j=1}^d \lambda_j^2 = \sum_{j=1}^d \lambda_j (T_{j+1} - T_j).$$

Summation by parts implies that

$$\sum_{j=1}^d \lambda_j (T_{j+1} - T_j) = (\lambda_d T_{d+1} - \lambda_1 T_1) - \sum_{j=2}^d T_j (\lambda_j - \lambda_{j-1})$$
$$= \lambda_d T_{d+1} - \sum_{j=2}^d T_j (\lambda_j - \lambda_{j-1}) \leq |\lambda_d||T_{d+1}| + \left| \sum_{j=2}^d T_j (\lambda_j - \lambda_{j-1}) \right|.$$

Since $\lambda_j - \lambda_{j-1} \geq 0$, we have that

$$\left|\sum_{j=2}^{d} T_j(\lambda_j - \lambda_{j-1})\right| \leq \max_{2\leq j\leq d} |T_j| \sum_{j=2}^{d}(\lambda_j - \lambda_{j-1}) = (\lambda_d - \lambda_1) \max_{2\leq j\leq d} |T_j|,$$

hence

$$\|\Sigma_Z - \Sigma\|_F^2 \leq |\lambda_d||T_{d+1}| + (\lambda_d - \lambda_1) \max_{2\leq j\leq d} |T_j| \leq 2\,\|\Sigma_Z - \Sigma\| \max_{2\leq j\leq d+1} |T_j|. \tag{A.118}$$

It remains to bound $\max_{2\leq j\leq d+1} |T_j|$. Note that for $j = 2, \ldots, d+1$,

$$\begin{aligned}
|T_j| &= \left|\sum_{l=1}^{j-1} \lambda_l\right| = \left|\sum_{l=1}^{j-1} \langle(\Sigma_Z - \Sigma)u_l, u_l\rangle\right| = \left|\sum_{l=1}^{j-1} \langle\Sigma_Z u_l, u_l\rangle - \langle\Sigma u_l, u_l\rangle\right| \\
&= \left|\sum_{l=1}^{j-1} \mathbb{E}\big[\langle Y, u_l\rangle^2\, \mathbb{1}\{\|Y\|_2 \leq R\}\big] - \mathbb{E}\big[\langle Y, u_l\rangle^2\big]\right| = \left|\sum_{l=1}^{j-1} \mathbb{E}\big[\langle Y, u_l\rangle^2\, \mathbb{1}\{\|Y\|_2 > R\}\big]\right|.
\end{aligned}$$

Applying Cauchy-Schwartz inequality and $L_4 - L_2$ norm equivalence, we deduce that

$$\begin{aligned}
|T_j| &\leq \sum_{l=1}^{j-1} \mathbb{E}\big[\langle Y, u_l\rangle^4\big]^{\frac{1}{2}} P(\|Y\|_2 > R)^{\frac{1}{2}} \leq \sum_{l=1}^{j-1} K^2 \mathbb{E}\big[\langle Y, u_l\rangle^2\big] P(\|Y\|_2 > R)^{\frac{1}{2}} \\
&\leq K^2 P(\|Y\|_2 > R)^{\frac{1}{2}} \sum_{l=1}^{d} \mathbb{E}\big[\langle Y, u_l\rangle^2\big].
\end{aligned} \tag{A.119}$$

Observe that $\{u_1, \ldots, u_d\}$ is an orthonormal set on $\mathbb{R}^d$, so Parseval's identity implies that

$$\sum_{l=1}^{d} \mathbb{E}\big[\langle Y, u_l\rangle^2\big] = \mathbb{E}\big[\|Y\|_2^2\big] = \operatorname{tr}(\mathbb{E}\Big[Y^T Y\Big]) = \mathbb{E}\Big[\operatorname{tr}(YY^T)\Big] = \operatorname{tr}(\mathbb{E}\Big[YY^T\Big]) = \operatorname{tr}(\Sigma). \tag{A.120}$$

On the other hand, applying Cauchy-Schwartz inequality and $L_4 - L_2$ norm equivalence again, we have that

$$\mathbb{E}\big[\|Y\|_2^4\big] = \mathbb{E}\left[\Big(\sum_{j=1}^{d} \langle Y, e_j\rangle^2\Big)^2\right] = \mathbb{E}\left[\sum_{i,j} \langle Y, e_i\rangle^2 \langle Y, e_j\rangle^2\right]$$

$$\leq \sum_{i,j} \mathbb{E}\left[\langle Y, e_i\rangle^4\right]^{\frac{1}{2}} \mathbb{E}\left[\langle Y, e_j\rangle^4\right]^{\frac{1}{2}} \leq K^4 \sum_{i,j} \mathbb{E}\left[\langle Y, e_i\rangle^2\right] \mathbb{E}\left[\langle Y, e_j\rangle^2\right] = K^4 \sum_{i,j} \Sigma_{i,i}\Sigma_{j,j} = K^4 \mathrm{tr}\,(\Sigma)^2.$$

Markov's inequality implies that

$$P(\|Y\|_2 > R)^{\frac{1}{2}} \leq \left(\frac{\mathbb{E}\left[\|Y\|_2^4\right]}{R^4}\right)^{\frac{1}{2}} \leq K^2 \frac{\mathrm{tr}\,(\Sigma)}{R^2}. \tag{A.121}$$

Combining (A.119,A.120,A.121) together, we have that

$$|T_j| \leq K^2 \cdot K^2 \frac{\mathrm{tr}\,(\Sigma)}{R^2} \cdot \mathrm{tr}\,(\Sigma) = K^4 \frac{\mathrm{tr}\,(\Sigma)^2}{R^2}$$

for $j = 2, \ldots, d+1$. Therefore,

$$\|\Sigma_Z - \Sigma\|_F^2 \leq 2\|\Sigma_Z - \Sigma\| \max_{2 \leq j \leq d+1} |T_j| \leq 2 \cdot K^3 \frac{\|\Sigma\|\,\mathrm{tr}\,(\Sigma)}{R^2} \cdot K^4 \frac{\mathrm{tr}\,(\Sigma)^2}{R^2} = 2K^7 \frac{\|\Sigma\|\,\mathrm{tr}\,(\Sigma)^3}{R^4},$$

hence

$$\|\Sigma_Z - \Sigma\|_F \leq \sqrt{2} K^{\frac{7}{2}} \frac{\|\Sigma\|^{\frac{1}{2}}\,\mathrm{tr}\,(\Sigma)^{\frac{3}{2}}}{R^2} = c(K) \frac{\|\Sigma\|^2\,\mathrm{rk}(\Sigma)^{\frac{3}{2}}}{R^2},$$

as desired.