Master's Paper of the Department of Statistics, the University of Chicago
(Internal document only, not for circulation)

Robust Estimation of Covariance Matrix and Precision Matrix

Wanrong Zhu

Advisor(s):Wei Biao Wu

Approved ______________________________

Date ______________________________

Feb-14, 2018

Abstract

Given n i.i.d. observations of a random vector $\mathbf{X} \in \mathbb{R}^p$, we consider the estimation of its covariance matrix $\mathbf{\Sigma}$ and its precision matrix $\mathbf{\Omega}$ when $\mathbf{X}$ is heavy-tailed distributed. To deal with the heavy-tailed issue, we propose a robust estimator for $\mathbf{\Sigma}$ using Huber loss and a modified robust Dentizig type estimator for $\mathbf{\Omega}$. We show that, with proper choice of truncating parameter in Huber loss, the robust estimator is consistent under the infinite norm, and dimension p is allowed to grow exponentially with the sample size n. We derive the explicit exponential-type tail bound for it in both cases with existence of only the fourth moment and only the second moment. When the true precision matrix is sparse, we derive the explicit tail bound for the modified robust Dentizig type estimators under the spectral norm. We also establish the rates of convergence under different norms for the estimator. Numerical performance of these two estimators is investigated by simulation studies. We compare our robust estimator for $\mathbf{\Sigma}$ with sample covariance matrix and the modified robust Dentizig type estimator for $\mathbf{\Omega}$ with existing regularized estimator (**CLIME**). The procedure of robust estimation is easy to implement and we propose a data driven method to choose truncating parameter in Huber loss. Simulation results show that the robust estimators have comparable performance with existing estimators when distribution of $\mathbf{X}$ is Gaussian. When $\mathbf{X}$ is heavy-tail distributed, for example t_3 distribution, the robust estimators have better performance and the advantage become more significant as p increases.

Contents

1 Introduction **3**

2 Robust estimator **5**

3 Main results **7**

3.1 Rates of convergence for robust covariance matrix 7

3.2 Rates of convergence for RCLIME 8

4 Proofs of main results **11**

5 Numerical results **15**

5.1 Choice of t 15

5.2 Simulation results 16

6 Discussion **20**

A Appendix **20**

1 Introduction

Estimation of a covariance matrix and its inverse is of great importance in variance areas of statistical analysis, including principle component analysis (PCA), linear or quadratic discriminant analysis (LDA and QDA), graphical models and functional estimation. Accurate and stable estimation is important especially for the analysis in ultrahigh dimensional settings, where the number of features p greatly surpasses the sample size n. Some classical methods and results may no longer be feasible in the high-dimensional settings. In addition, assumptions on the tail distributions and other restrictions may bring us unwanted limit. To solve above problems, alternative methods have been studied and developed.

Let $\mathbf{X} = (X_1, ...X_p)^T$ be a p-variate random vector with covariance matrix $\mathbf{\Sigma}$ and precision matrix $\mathbf{\Omega} \coloneqq \mathbf{\Sigma}^{-1}$. Given an independent and identically distributed random sample $\{\mathbf{X}_1, ...\mathbf{X}_n\}$ from distribution of $\mathbf{X}$, a natural pivotal estimator for covariance matrix $\mathbf{\Sigma}$ is its sample covariance matrix $\mathbf{S} = n^{-1}\sum_{i=1}^{n}\left(\mathbf{X}_i - \bar{\mathbf{X}}\right)\left(\mathbf{X}_i - \bar{\mathbf{X}}\right)^T$, where $\bar{\mathbf{X}} = n^{-1}\sum_{i=1}^{n}\mathbf{X}_i$. But it is well known that the empirical sample covariance matrix is not a good estimator if p is large since neither the eigenvalues nor the eigenvectors of $\mathbf{S}$ are consistently converge to those of $\mathbf{\Sigma}$ (see, e.g. [1, 2]). In addition, sample covariance matrix $\mathbf{S}$ is singular if $p > n$ and thus its inverse is not well defined, let alone to use its inverse to estimate precision matrix. To address this problem, regularized estimators have been studied extensively under different model assumptions. For data with ordering structures, Bickel and Levina [3] introduced the banding approach to estimate covariance matrix and the Cholesky factor of its inverse. They showed that banding the sample covariance matrix leads to a well-behaved estimator. Cai et al. [4, 5] present the minimax rate and optimal rates of banding or tapering estimators. Shrinkage estimators also have been proposed. Bickel and Levina [6] showed consistency of the thresholded estimator in terms of operator norm. To estimate the precision matrix, Cholesky decomposition based method has been studied by several authors like Wu and Pourahmadi [7], Bickel and Levina [3]. Penalized likelihood methods also have been studied during the last decade. Different penalties were considered from convex l_1 penalty to non-convex penalties like SCAD and MCP. Theoretical properties have been thoroughly studied by Rothman et al. [8] and Lam and Fan [9]. Column-by-column estimation method was then proposed to make computation simpler. Yuan [10] proposed the graphical Dantzig selector and Cai, Liu and Luo [11] proposed CLIME, both can be solved by linear programming.

Although methods above tackled some problems in high-dimensional settings, they still need light tail assumptions. In the heavy tail situation, the traditional techniques are infeasible in that the deviations may be difficult to diagnose. But the thing is that sometimes such deviations are not important. A traditional way is to appeal to robustness. In this paper, we propose a robust estimator using Huber loss for covariance matrix. The concentration results for the robust estimator does not require light tail assumptions. If there exists the fourth moment, we can get the concentration rate $\|\mathbf{\Sigma}_n - \mathbf{\Sigma}\|_\infty = O_p\left(\sqrt{\log(p)/n}\right)$ with proper choice of truncating parameter. This rate is same as the concentration rate for sample covariance matrix in case of Gaussian distribution. If the tail is even heavier such that there only exists $(2+\delta)$th moment, we can still get the concentration rate $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_\infty =$

$O_p\left((\log(p)/n)^{1-1/(1+\delta/2)}\right)$. We also present the explicit tail bounds in Theorem 1 and 2.

We then focus on the problem of estimating precision matrix under sparsity conditions. We modify existing Dantzig-type estimator for precision matrix by replacing the sample covariance matrix with our robust estimator. We call this method of estimating precision matrix robust constrained l_1-minimization for inverse matrix estimation (RCLIME). One advantage of RCLIME is that the RCIME estimator can be used for both light-tail and heavy-tail cases. The requirement for tail condition can be relaxed. We show that the RCLIME can achieve the same convergence rates under different norms as those of constrained l_1-minimization for inverse matrix estimation (CLIME) in [11] when only the fourth moment exists. And p can still grow exponentially with the sample size n. We also derive the exponential-type bound when there only exists $(2+\delta)$th moments. This bound shows that RCLIME has good concentration property even when the tail is heavier. Besides, the simulation experiment shows that the robust estimator for covariance matrix performs much better than sample covariance matrix both in light tail cases and heavy tail cases. And numerical performance of RCLIME is slightly better than that of CLIME under the light tail distribution. The advantage become significant as the tail of distribution becomes heavier and p increases.

The rest of this paper is organized as follows. In section 2, we introduce the robust estimator for covariance matrix and RCLIME estimator for precision matrix. We present the theoretical properties in section 3. Section 4 provides an outline of the proofs, with the more technical details deferred to appendices. In section 5, we propose a quantile method to select truncating parameter in Huber loss when constructing robust estimator for covariance matrix and investigate numerical performance of two robust estimators. We provide further discussion in section 6.

2 Robust estimator

In this section, we introduce a robust estimator for covariance matrix and a robust Dentzig-type estimator for precision matrix. Before turning to the estimators, we introduce some notations first. For a vector $\mathbf{a} = (a_1, ..., a_n) \in \mathbb{R}^n$, we define vector l_1 norm $\|\mathbf{a}\|_1 = \sum_{i=1}^n |a_i|$, vector l_2 norm $\|\mathbf{a}\|_2 = \sqrt{\sum_{i=1}^n a_i^2}$, vector l_∞ norm $\|\mathbf{a}\|_\infty = \max_{1\leq i\leq n} |a_i|$. For a matrix $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{n\times n}$, we define the elementwise l_1 norm $\|\mathbf{A}\|_1 = \sum_{i=1}^n \sum_{j=1}^n |a_{ij}|$, elementwise l_∞ norm $\|\mathbf{A}\|_\infty = \max_{1\leq i\leq n, 1\leq j\leq n} |a_{ij}|$, Frobenius norm $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^n a_{ij}^2}$, matrix L_1 norm $\|\mathbf{A}\|_{L_1} = \max_{1\leq j\leq n} \sum_{i=1}^n |a_{ij}|$, and operator norm $\|\mathbf{A}\|_2 = \sup_{\|x\|_2\leq 1} \|\mathbf{A}x\|_2$. We use $\mathbf{I}$ to denote a $p \times p$ identity matrix.

Suppose $\mathbf{X} = (X_1, ...X_p)^T$ is a p-variate random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Let $\boldsymbol{\Omega}$ denote $\boldsymbol{\Sigma}^{-1}$. Suppose $\{\mathbf{X}_i\}_{i=1}^n$ is an independent and identically distributed random sample from the distribution of $\mathbf{X}$. Write $\boldsymbol{\mu} = (\mu_1, ..., \mu_p)^T$, $\boldsymbol{\Sigma} = (\sigma_{ij})$, $\boldsymbol{\Omega} = (\omega_{ij})$.

We now recall how we define sample covariance matrix $\mathbf{S}$. Covariance σ_{ij} between component X_i and X_j is defined as $\mathbb{E}X_iX_j - \mathbb{E}X_i\mathbb{E}X_j$. Sample covariance $\tilde{\sigma}_{ij}$ is defined as $\tilde{\sigma}_{ij} = n^{-1}\sum_{k=1}^n X_{ik}X_{jk} - (n^{-1}\sum_{k=1}^n X_{ik})(n^{-1}\sum_{k=1}^n X_{jk})$. It can be viewed as combination of least square mean estimators for $\mathbb{E}X_iX_j$, $\mathbb{E}X_i$ and $\mathbb{E}X_i$. Sample version estimators are obtained through minimizing quadratic loss. However, in heavy tail case, there tend to be more extreme values far away from the mean value. Quadratic loss of those values will be large and affect the estimation results. To underweight effect caused by extreme values and robustify the estimation, we propose to use the Huber loss[12]:

$$l_t(x) = \begin{cases} t|x| - \frac{1}{2}t^2 & \text{if } |x| > t; \\ \frac{1}{2}x^2 & \text{if } |x| \leq t. \end{cases} \tag{1}$$

In this paper, we regard t as a thresholding parameter, whose optimal value will be further discussed in next section. The Huber loss is quadratic for small values of x and linear for large values of x. The parameter t balances quadratic and linear penalty.

We now define robust estimator $\boldsymbol{\Sigma}_n$ for $\boldsymbol{\Sigma}$ and robust Dentzig-type estimator $\hat{\boldsymbol{\Omega}}$ for $\boldsymbol{\Omega}$. We first define robust mean estimator using the Huber loss. For $1 \leq i \leq p$, the robust estimator $\hat{\mu}_i$ for $\mathbb{E}X_i$ is defined by solving the following optimization problem:

$$\hat{\mu}_i = \arg\min_{\mu_i\in\mathbb{R}} \sum_{k=1}^n l_{t_{ni}}\left(X_{ik} - \mu_i\right). \tag{2}$$

The thresholding parameter t depends on $\{X_{ik}\}_{k=1}^n$. It should be noted that solving the above optimization problem is equivalent to solving the following equation:

$$\sum_{k=1}^n \rho_{t_{ni}}(X_{ik} - \mu_i) = 0. \tag{3}$$

where $\rho_t(x) = x$ if $|x| \leq t$, $\rho_t(x) = t$ if $|x| > t$ and $\rho_t(x) = -t$ if $|x| < t$. We can see that solution to equation (3) is unique because $\rho_t(x)$ is monotonous. Therefore, $\hat{\mu}_i$ is solution to equation (3). Similarly, for $1 \leq i \leq p$, $1 \leq j \leq p$, the robust estimator $\hat{\mu}_{ij}$ for $\mathbb{E}X_iX_j$ is solution of the following equation:

$$\sum_{k=1}^{n} \rho_{t_{nij}} \left(X_{ki}X_{kj} - \mu_{ij}\right) = 0. \tag{4}$$

It should be noted that $\hat{\mu}_{ij} = \hat{\mu}_{ji}$. With robust mean estimators $\hat{\mu}_i$ and $\hat{\mu}_{ij}$ defined above, the robust covariance matrix $\mathbf{\Sigma}_n$ for $\mathbf{\Sigma}$ is defined as

$$\begin{aligned} &\mathbf{\Sigma}_n = (\hat{\sigma}_{ij}), \text{ where} \\ &\hat{\sigma}_{ij} = \hat{\mu}_{ij} - \hat{\mu}_i\hat{\mu}_j. \end{aligned} \tag{5}$$

Clearly, $\mathbf{\Sigma}_n$ is symmetric since $\hat{\mu}_{ij} = \hat{\mu}_{ji}$. Let $\hat{\mathbf{\Omega}}_1 = (\hat{w}^1_{ij})$ be solution of the following optimization problem:

$$\begin{aligned} &\min \|\mathbf{\Omega}\|_1 \text{ subject to :} \\ &\|\mathbf{\Sigma}_n\mathbf{\Omega} - \mathbf{I}\|_\infty \leq \lambda_n, \quad \mathbf{\Omega} \in \mathbb{R}^{p\times p}, \end{aligned} \tag{6}$$

where λ_n is a tuning parameter, whose optimal value will be discussed in next section. Since we do not require symmetric condition on $\hat{\mathbf{\Omega}}_1$ in (6), we need to symmetrize $\hat{\mathbf{\Omega}}_1$ in next step. The final estimator $\hat{\mathbf{\Omega}}$ for $\mathbf{\Omega}$ is defined as

$$\begin{aligned} &\hat{\mathbf{\Omega}} = (\hat{w}_{ij}), \quad \text{where} \\ &\hat{w}_{ij} = \hat{w}^1_{ij} I_{\{|\hat{w}^1_{ij}| \leq |\hat{w}^1_{ji}|\}} + \hat{w}^1_{ji} I_{\{|\hat{w}^1_{ji}| < |\hat{w}^1_{ij}|\}}. \end{aligned} \tag{7}$$

Now, $\hat{\mathbf{\Omega}}$ is symmetric and positive definite with high probability from convergence results in section 3.

3 Main results

In this section, we state our main results. We begin in Section 3.1 by stating some moment conditions on $\mathbf{X}$. In rest of Section 3.1, we illustrate Theorem 1 on convergence rate for our robust covariance matrix estimator with existence of fourth moment and Theorem 2 on convergence rate for $\mathbf{\Sigma}_n$ when there only exists second moment. Section 3.2 is devoted to illustrate rates of convergence for the RCLIME estimator under different moment conditions and different norms.

3.1 Rates of convergence for robust covariance matrix

This paper is focused on heavy tail cases. We shall specify the tail conditions before illustrating our main results. Write $\mathbf{X} = (X_1, ..., X_p)^T$. We split the analysis into two cases.

(A1): (Fourth moment case) Suppose that there exist some $\nu < \infty$ such that

$$\max\left\{\max_{1\le i,j\le p}\sqrt{\mathrm{var}(X_iX_j)},\ \max_{1\le i\le p}\mathrm{var}(X_i)\right\} \le \nu.$$

(A2): (Second moment case) Suppose that for some $0 < \delta < 2$,

$$\max_{1\le i\le p}\mathbb{E}|X_i|^k = m_k \le K,\ \text{for all } 0 < k < 2+\delta,$$

where $= m_k$ is constant for all k.

In both two cases, $\mathbf{X}$ has polynomial-type tails. Note that (A1) requires finite fourth moment for $\mathbf{X}$. Under (A2), there requires existence of only the second moment. Following Theorems, Theorem 1 and Theorem 2, give us rates of convergence for robust covariance matrix $\mathbf{\Sigma}_n$ under infinite norm loss for above two cases.

Theorem 1. *Assume that (A1) holds. For some* $a > 2$ *such that* $\log(p)/n \le 1/(8a)$*, let* $t_{ni} = t_{nij} = \sqrt{\nu^2/(a\log(p)/n)}$*. Then*

$$\|\mathbf{\Sigma}_n - \mathbf{\Sigma}\|_\infty < (4\nu + 8m\nu + 8\nu^2)\sqrt{\frac{a\log(p)}{n}}, \tag{8}$$

with probability greater than $1 - 10p^{2-a}$*, where* $m = \max_{1\le i\le p}\mathbb{E}|X_i|$*,* $\mathbf{\Sigma}_n$ *is defined as in (5).*

This concentration result shows that

$$\|\mathbf{\Sigma}_n - \mathbf{\Sigma}\|_\infty = O_p\left(\left(\frac{\log(p)}{n}\right)^{1/2}\right), \tag{9}$$

which is the same rate as that of sample covariance matrix in sub-Gaussian case.

We can see that (A1) still requires the fourth moment to be bounded. To further relax the tail condition, we consider (A2), where only the second moment is required. The result is stated as Theorem 2.

Theorem 2. *Under condition (A2), let* $\frac{\log(p)}{n} = o(1)$*, for* $1 \leq i, j \leq p$*,* $t_{ni} = \sqrt{m_{j2}/\{a_4 \log(p)/n\}}$*,* $t_{nij} = \left[m_{ij(1+\delta/2)}/\left\{a_2 \log(p)/n\right\}\right]^{1/(1+\delta/2)}$*. Then*

$$\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_\infty \leq 2a_1 m_{2+\delta}\left(a_2 \frac{\log(p)}{nm_{2+\delta}}\right)^{1-\frac{1}{1+\delta/2}} + 8m_1\sqrt{a_4 m_2 \frac{\log(p)}{n}} + 16a_4 m_2 \frac{\log(p)}{n}, \qquad (10)$$

with probability greater than $1 - (2p^{-1} + 2p^{a_3} + 8p^{2-a_4})$*, where* $m_{j2} = \mathbb{E}|X_j|^2$*,* $m_{ij(1+\delta/2)} = \mathbb{E}|X_i X_j|^{1+\delta/2}$*, for* $1 \leq i, j \leq p$*,* a_1 *is some constant associated with* δ*,* $a_2 = 8/\left(a_1 - \delta 2^{\delta/2+1} - 2\right)$*,* $a_3 = 2 - 2^{-\delta/2-2}\left(a_1 - \delta 2^{\delta/2+1} - 2\right) < 0$*,* $a_4 > 2$ *such that* $\log(p)/n \leq 1/(8a_4)$*.*

The explicit tail bound in Theorem 2 is complicated. It indicates that when we choose thresholding parameters $t_{ni} = O\left(\left(\frac{\log(p)}{n}\right)^{-1/2}\right)$, $t_{nij} = O\left(\left(\frac{\log(p)}{n}\right)^{-\frac{1}{1+\delta/2}}\right)$, the convergence rate is

$$\|\mathbf{\Sigma}_n - \mathbf{\Sigma}\|_\infty = O_p\left(\left(\frac{\log(p)}{n}\right)^{1-\frac{1}{1+\delta/2}}\right). \qquad (11)$$

We can see that when $\delta \to 2$, bound in equation (11) approaches $(\log(p)/n)^{1/2}$. In both two heavy-tail cases, we show that our robust covariance matrix estimator $\mathbf{\Sigma}_n$ converge to true covariance matrix $\mathbf{\Sigma}$ under infinite norm. And we allow dimension p to grow exponentially with sample size n.

3.2 Rates of convergence for RCLIME

We begin by considering the uniformity class of precision matrix defined by

$$\mathcal{U}\left(M, s_0(p), q\right) = \left\{\mathbf{\Omega} : \mathbf{\Omega} \succ 0, \|\mathbf{\Omega}\|_{L_1} \leq M, \max_{1\leq i\leq p}\sum_{j=1}^{p}|\omega_{ij}|^q \leq s_0(p)\right\}, \qquad (12)$$

for $0 \leq q < 1$. If $q = 0$, $\mathcal{U}\left(M, s_0(p), 0\right)$ is a class of sparse matrices. The uniformity class of matrices $\mathcal{U}(M, q, s_0(p))$ is commonly used when the sparsity assumption is needed in analysis. For example, Bickel and Levina (2008b) used it in covariance thresholding. Following analysis of precision matrix is uniformly on $\mathcal{U}\left(M, s_0(p), q\right)$.

Theorem 3. *Under conditions in Theorem 1, let* $\lambda_n = C_0 M\sqrt{\log(p)/n}$*. Then we have*

$$\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\|_2 \leq C_1 M^{2-2q} s_0(p)\left(\frac{\log p}{n}\right)^{(1-q)/2}, \qquad (13)$$

with probability greater than $1 - 10p^{2-a}$ *, where* $C_0 = (4\nu + 8m\nu + 8\nu^2)\sqrt{a}$*,* $C_1 = 2C_0^{1-q}4^{1-q}(1 + 2^{1-q} + 3^{1-q})$*.*

Theorem 3 shows that

$$\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\|_2 = O_p\left(s_0(p)\left(\frac{\log(p)}{n}\right)^{(1-q)/2}\right). \qquad (14)$$

This is the same rate of convergence as CLIME estimator in [11] and the inverse of thresholding estimator in [6]. It is reasonable that we yield the same rate as previous results. That is because, in previous studies, the key of proof is the exponential-type bound of sample covariance matrix under infinite norm in exponential-tail type case. And Theorem 1 shows that our robust covariance estimator achieves the same exponential bound. But in [6], Guassian assumption is required to achieve this rate. And in the case of polynomial-type tails in [11], p cannot grow exponentially with sample size p. Here we consider polynomial-type tail case with existence of only fourth moment and allow p to grow exponentially with sample size n.

Next we consider the case if there only exists the second moment. Based on theorem 2 above we have the following Theorem.

Theorem 4. *Under conditions in Theorem 2, let* $\lambda_n = C_3 M(log(p)/n)^{1-1/(1+\delta/2)}$. *Then we have*

$$\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\|_2 \leq C_2 M^{2-2q} s_0(p) \left(\frac{\log p}{n}\right)^{(1-q)/(1+2/\delta)}, \tag{15}$$

with probability greater than $1 - (2p^{-1} + 2p^{a_3} + 8p^{2-a_4})$, *where* $C_3 = 8m_1\sqrt{a_4 m_2} + 16a_4 m_2 + a_1 m_{2+\delta} \left(a_2/m_{2+\delta}\right)^{1-1/(1+\delta)}$ *and* $C_2 = 2C_3^{1-q}4^{1-q}\left(1 + 2^{1-q} + 3^{1-q}\right)$, a_1, a_2, a_3 *and* a_4 *are defined in Theorem 2.*

We can see that Theorem 4 yields a convergence rate

$$\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\|_2 = O_p\left(s_0(p)\left(\frac{\log(p)}{n}\right)^{(1-q)/(1+2/\delta)}\right), \tag{16}$$

for $0 < \delta < 2$. This rate is slower than that in equation (14) but it still shows that $\hat{\mathbf{\Omega}}$ has a good convergence property. And the rate approaches the rate in equation (14) when δ approaching 2.

Next we will study the convergence rates for $\sup_{\mathbf{\Omega}\in\mathcal{U}} \mathbb{E}\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\|_{\mathbf{2}}$. This expectation does not always exist. We will make some modification to $\hat{\mathbf{\Omega}}$ to make the expectation well defined. We first replace $\mathbf{\Sigma}_n$ with $\mathbf{\Sigma}_{n,\rho} = \mathbf{\Sigma}_n + \rho\mathbf{I}$. Through adding a small value we can ensure $\mathbf{\Sigma}_{n,\rho}^{-1}$ is a feasible point. Then we replace $\hat{\mathbf{\Omega}}_1$ with $\hat{\mathbf{\Omega}}_{1,\rho}$, where $\hat{\mathbf{\Omega}}_{1,\rho}$ minimize $\|\mathbf{\Omega}\|_1$ subject to $\|\mathbf{\Sigma}_{n,\rho} - \mathbf{I}\|_\infty \leq \lambda_n$. We use $\hat{\mathbf{\Omega}}_\rho$ to denote the final estimator after same symmetrization scheme in section 2. We can see that $\|\hat{\mathbf{\Omega}}_\rho\|_{L1} \leq \|\mathbf{\Sigma}_{n,\rho}^{-1}\|_{L1} < \infty$. Then $\sup_{\mathbf{\Omega}\in\mathcal{U}} \mathbb{E}\|\hat{\mathbf{\Omega}}_\rho - \mathbf{\Omega}\|_2$ is well defined. We derive the following convergence rate for fourth moment case.

Theorem 5. *Under conditions in Theorem 3, if* $\rho = \sqrt{\frac{\log(p)}{n}}$, *then we have*

$$\sup_{\mathbf{\Omega}\in\mathcal{U}} \mathbb{E}\|\hat{\mathbf{\Omega}}_\rho - \mathbf{\Omega}\|_2^2 = O\left(M^{4-4q}s_0^2(p)\left(\frac{\log p}{n}\right)^{1-q}\right). \tag{17}$$

Lastly, we consider convergence rates under Frobenius norm and matrix L1 norm. We have Theorem 6 for fourth moment case.

Theorem 6. *Under conditions of Theorem 3, we have*

1.

$$\frac{1}{p}\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\|_F^2 \leq 4C_1 M^{4-2q} s_0(p) \left(\frac{\log(p)}{n}\right)^{1-q/2}, \tag{18}$$

$$\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\|_\infty \leq 4C_0 M^2 \sqrt{\frac{\log(p)}{n}}, \tag{19}$$

with probability greater than $1 - 10p^{2-a}$.

2.

$$\sup_{\mathbf{\Omega}\in\mathcal{U}} \mathbb{E}\|\hat{\mathbf{\Omega}}_\rho - \mathbf{\Omega}\|_F^2 = O\left(M^{4-2q} s_0^2(p) \left(\frac{\log(p)}{n}\right)^{1-q/2}\right) \tag{20}$$

$$\sup_{\mathbf{\Omega}\in\mathcal{U}} \mathbb{E}\|\hat{\mathbf{\Omega}}_\rho - \mathbf{\Omega}\|_\infty^2 = O\left(M^4 \frac{\log(p)}{n}\right) \tag{21}$$

4 Proofs of main results

In this section, we work through proofs of main results in section 3. Our proofs are based on several lemmas, with some of the more technical aspects deferred to appendices.
Lemma 1. *Let* $\{y_i\}_{i=1}^n$ *be an independent and identically distributed sample with* $\mathbb{E}(y_i) = \mu$ *and* $\mathrm{var}(y_i) = \sigma^2$. *Let* $\hat{\mu}$ *be the solution of* $\sum_{i=1}^n \rho_{t_n}(y_i - \mu) = 0$. *Assume that* $\log(p)/n \leq 1/(8a)$, *where* $a > 2$, $t_n = \sqrt{\nu^2/\left(a\log(p)/n\right)}$, *where* $\nu \geq \sigma$.

$$P\left(|\hat{\mu} - \mu| \geq 4\nu\sqrt{\frac{a\log(p)}{n}}\right) \leq 2p^{-a}. \tag{22}$$

Lemma 1 illustrate the concentration property of robust mean estimator when variance exists. We will use Lemma 1 to proof Theorem 1.

Proof of Theorem 1. For some i and j, we can split $|\hat{\sigma}_{ij} - \sigma_{ij}|$ into two part as following.

$$\begin{aligned} |\hat{\sigma}_{ij} - \sigma_{ij}| &= |\hat{\mu}_{ij} - \hat{\mu}_i\hat{\mu}_j\mathbb{E}X_iX_j + \mathbb{E}Xi\mathbb{E}X_j| \\ &\leq |\hat{\mu}_{ij} - \mathbb{E}X_iX_j| + |\hat{\mu}_i\hat{\mu}_j - \mathbb{E}X_i\mathbb{E}X_j| \\ &= I + II, \end{aligned} \tag{23}$$

moreover,

$$\begin{aligned} II &\leq |\mathbb{E}X_j||\hat{\mu}_i - \mathbb{E}X_i| + |\hat{\mu}_i - \mathbb{E}X_i||\hat{\mu}_j - \mathbb{E}X_j| + |\mathbb{E}X_i||\hat{\mu}_j - \mathbb{E}X_j| \\ &\leq m|\hat{\mu}_i - \mathbb{E}X_i| + m|\hat{\mu}_j - \mathbb{E}X_j| + |\hat{\mu}_i - \mathbb{E}X_i||\hat{\mu}_j - \mathbb{E}X_j|. \end{aligned} \tag{24}$$

Choice of t_{nij}, t_{ni} and t_{ni} satisfy the condition in Lemma 1. By Lemma 1 we have

$$P\left(I = |\hat{\mu}_{ij} - \mathbb{E}X_iX_j| \geq 4\nu\sqrt{\frac{a\log(p)}{n}}\right) \leq 2p^{-a}, \tag{25}$$

and

$$P\left(m|\hat{\mu_i} - \mathbb{E}X_i| \geq 4m\nu\sqrt{\frac{a\log{(p)}}{n}}\right) \leq 2p^{-a}, \tag{26}$$

$$P\left(m|\hat{\mu_j} - \mathbb{E}X_j| \geq 4m\nu\sqrt{\frac{a\log{(p)}}{n}}\right) \leq 2p^{-a}. \tag{27}$$

Combining (26) and (27), we have

$$\begin{aligned} &P\left(|\hat{\mu_i} - \mathbb{E}X_i||\hat{\mu_j} - \mathbb{E}X_j| \geq 16\nu^2\frac{a\log{(p)}}{n}\right) \\ &\leq P\left(|\hat{\mu_i} - \mathbb{E}X_i| \geq 4\nu\sqrt{\frac{a\log{(p)}}{n}}\right) + P\left(|\hat{\mu_j} - \mathbb{E}X_j| \geq 4\nu\sqrt{\frac{a\log{(p)}}{n}}\right) \\ &\leq 4p^{-a}. \end{aligned} \tag{28}$$

Note that $a\log(p)/n \leq 1/8$. Therefore, $16\nu^2 a\log(p)/n \leq 8\nu^2\sqrt{a\log(p)/n}$. Combining (26), (27), (28), we have

$$\begin{aligned} &P\left(II \geq \left(8m\nu + 8\nu^2\right)\sqrt{\frac{a\log(p)}{n}}\right) \\ &\leq P\left(II \geq 8m\nu\sqrt{\frac{a\log(p)}{n}} + 16\nu^2\frac{a\log(p)}{n}\right) \\ &\leq 8p^{-a}. \end{aligned} \tag{29}$$

From (25) and (29) we have

$$\begin{aligned} &P\left(|\hat{\sigma}_{ij} - \sigma_{ij}| \geq \left(4\nu + 8m\nu + 8\nu^2\right)\sqrt{\frac{a\log(p)}{n}}\right) \\ &\leq P\left(I \geq 4\nu\sqrt{\frac{a\log(p)}{n}}\right) + P\left(II \geq \left(8m\nu + 8\nu^2\right)\sqrt{\frac{a\log(p)}{n}}\right) \\ &\leq 10p^{-a}. \end{aligned} \tag{30}$$

Using union bound we can get

$$\begin{aligned} &P\left(\max_{1\leq i,j\leq p}|\hat{\sigma}_{ij} - \sigma_{ij}| \geq \left(4\nu + 8m\nu + 8\nu^2\right)\sqrt{\frac{a\log(p)}{n}}\right) \\ &\leq \sum_{1\leq i,j\leq p} P\left(|\hat{\sigma}_{ij} - \sigma_{ij}| \geq \left(4\nu + 8m\nu + 8\nu^2\right)\sqrt{\frac{a\log(p)}{n}}\right) \\ &\leq 10p^{2-a}. \end{aligned} \tag{31}$$

Then we get our final results that

$$\|\mathbf{\Sigma}_n - \mathbf{\Sigma}\|_\infty < \left(4\nu + 8m\nu + 8\nu^2\right)\sqrt{\frac{a\log(p)}{n}}, \tag{32}$$

with probability greater than $1 - 10p^{2-a}$. □

Next, we introduce Lemma 2 to show that we can still have exponential type bound for robust mean estimator when we just have finite $(1+\epsilon)th$ moment. In this case, even the variance is not exist. The proof of Lemma 2 is quite different from that of Lemma 1. We put the proof in appendix.

Lemma 2. *Let $\{x_i\}_{i=1}^n$ be independently identically distributed sample from the same distribution as random variable X. For $1 < \theta < 2$, $\mathbb{E}|X|^\theta = m_\theta < \infty$. Let $\hat{\mu}$ be the solution of $\sum_{i=1}^n \rho_k(x_i - \mu) = 0$. Then we have*

$$P\left(|\hat{\mu} - \mu| \geq t\right) \leq 2\left[\exp\left\{-\frac{3n(t - c_1 m_\theta k^{1-\theta})}{8k}\right\} + \exp\left\{-\frac{n(t - c_1 m_\theta k^{1-\theta})^2}{c_2 k^{2-\theta} m_\theta}\right\}\right], \tag{33}$$

where $c_1 = (2^{\theta-1})/(\theta-1) + 2$, $c_2 = 2^{4+\theta}$, k satisfies $t - c_1 m_\theta k^{1-\theta} > 0$ and $k > 2t$.

Proof of Theorem 2. Let $t = \tilde{l}_1 + 2m_1 l_2 + l_3^2$, then from equation (23), (24) we have

$$
\begin{aligned}
&P\left(|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}|_\infty \geq t\right) \\
\leq& p^2 P\left(|\hat{\sigma}_{ij} - \sigma_{ij}| \geq \tilde{l}_1 + 2m_1 l_2 + l_3^2\right) \\
\leq& p^2 P\left(|\hat{\mu}_{ij} - \mu_{ij}| \geq \tilde{l}_1\right) + 2p^2 P\left(|\hat{\mu}_i - \mu_i| \geq l_2\right) + 2p^2 P\left(|\hat{\mu}_i - \mu_i| \geq l_3\right) \\
=& I + II + III
\end{aligned}
\tag{34}
$$

For part I:

$\hat{\mu}_{ij}$ is robust mean estimator for $X_i X_j$. Since X_i has finite $(2+\delta)th$ moment, $X_i X_j$ has finite $(1+\delta/2)th$ moment. And

$$\mathbb{E}(X_i X_j)^{1+\delta/2} = m_{ij(1+\delta/2)} < \mathbb{E}(X_i)^{2+\delta} + \mathbb{E}(X_j)^{2+\delta} < 2m_{2+\delta}. \tag{35}$$

Using Lemma 2, we have

$$
\begin{aligned}
&p^2 P\left(|\hat{\mu}_{ij} - \mu_{ij}| \geq l_1\right) \\
\leq& 2\left[\exp\left\{2\log(p) - \frac{3n(l_1 - c_1 m_{ij(1+\delta/2)} t_{nij}^{-\delta/2})}{8t_{nij}}\right\} + \exp\left\{2\log(p) - \frac{n(l_1 - c_1 m_{ij(1+\delta/2)} t_{nij}^{-\delta/2})^2}{c_2 m_{ij(1+\delta/2)} t_{nij}^{1-\delta/2}}\right\}\right],
\end{aligned}
\tag{36}
$$

where $c_1 = \frac{2^{\delta/2}}{\delta/2} + 2$, $c_2 = 2^{5+\delta/2}$.

Let $l_1 = a_1 m_{ij(1+\delta/2)} t_{nij}^{-\delta/2}$, where $a_1 > c_1$. Then we let

$$t_{nij} = \left[\frac{a_2 \log(p)}{m_{ij(1+\delta/2)} n}\right]^{-1/(1+\delta/2)}, \tag{37}$$

where $a_2 = 8/(a_1 - \delta 2^{\delta/2+1} - 2)$ and equation $\left\{3n\left(l_1 - c_1 m_{ij(1+\delta/2)} t_{nij}^{-\delta/2}\right)\right\}/(8t_{nij}) = 3$ is satisfied.Let $\tilde{l}_1 = 2a_1 m_{2+\delta} t_{nij}^{-\delta/2}$. Since $m_{ij(1+\delta/2)} < 2m_{2+\delta}$, we have $\tilde{l}_1 > l_1$ and

$$I \leq p^2 P\left(|\hat{\mu}_{ij} - \mu_{ij}| \geq l_1\right) \leq 2p^{-1} + 2p^{a_3}, \tag{38}$$

Where $a_3 = 2 - \left(a_1 - \delta 2^{\delta/2+1} - 2\right) / \left(2^{\delta/2+2}\right)$. Here we can choose a_1 such that $a_3 < 0$.

For part II and III:

$\hat{\mu}_i$ is robust mean for X_i. We can use Lemma 1 here since $\mathbb{E}(X_i)$ has finite second moment. Assume that $\log(p)/n \leq 1/(8a_4)$, where $a_4 > 2$. Let $t_i = \sqrt{m_{j2}/\left(a_4 \log(p)/n\right)}$, $l_2 = l_3 = 4\sqrt{a_4 m_2 \log(p)/n} > 4\sqrt{a_4 m_{j2} \log(p)/n}$. Then using Lemma 1 we have

$$II = III \leq 4p^{2-a_4}. \tag{39}$$

Combining (34), (38), (39) we can get our final bound. □

Lemma 3. *Suppose that* $\boldsymbol{\Omega} \in \mathcal{U}\left(M, s_0(p), 0\right)$*,* $\rho \ge 0$*. We have*

$$\|\hat{\boldsymbol{\Omega}}_\rho - \boldsymbol{\Omega}\|_2 \le 2\left(1 + 2^{1-q} + 3^{1-q}\right)\left(4\|\boldsymbol{\Omega}\|_{L1}\right)^{1-q} s_0(p)\lambda_n^{1-q}, \tag{40}$$

with $\lambda_n \ge \|\boldsymbol{\Omega}\|_{L1}\left(\max_{i,j}|\hat{\sigma}_{ij} - \sigma_{ij}| + \rho\right)$*.*

This lemma is stated as Theorem 6 in Cai, Liu, Luo(2011). We will not repeat the proof in this paper. This lemma can be applied here to proof Theorem 4 and Theorem 5 since the proof of this lemma just uses the property of l_1 minimization not the structure of sample covariance.

Proof of Theorem 3. Let $\lambda_n = C_0 M\sqrt{\frac{\log p}{n}}$ and $\rho = 0$, where $C_0 = (4\nu + 8m\nu + 8\nu^2)\sqrt{a}$. Then under event $\mathcal{A} = \left\{\max_{i,j}|\hat{\sigma}_{ij} - \sigma_{ij}| < C_0\sqrt{\frac{\log p}{n}}\right\}$, $\lambda_n \ge \|\boldsymbol{\Omega}\|_{L1}(\max_{i,j}|\hat{\sigma}_{ij} - \sigma_{ij}|)$ which satisfies conditions in Lemma 3.

So under $\mathcal{A}$, by Theorem 1, with probability $\ge 1 - 10p^{2-a}$,

$$\begin{aligned}
\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\|_2 &\le 2\left(1 + 2^{1-q} + 3^{1-q}\right)\left(4\|\boldsymbol{\Omega}\|_{L1}\right)^{1-q} s_0(p)\lambda_n^{1-q} \\
&\le 2\left(1 + 2^{1-q} + 3^{1-q}\right)\left(4M\right)^{1-q} s_0(p)\left(C_0 M\sqrt{\frac{\log(p)}{n}}\right)^{1-q} \\
&= C_1 M^{2-2q} s_0(p)\left(\frac{\log p}{n}\right)^{(1-q)/2},
\end{aligned} \tag{41}$$

where $C_1 = 2C_0^{1-q}4^{1-q}\left(1 + 2^{1-q} + 3^{1-q}\right)$. □

Proof of Theorem 4 is the same as proof of Theorem 3 and we will not repeat the proof here.

Proof of Theorem 5. We have that $\|\hat{\boldsymbol{\Omega}}_\rho\|_{L1} \le \|\boldsymbol{\Sigma}_{n,\rho}^{-1}\|_{L1} \le p\rho^{-1} = p\sqrt{\frac{n}{log(p)}}$ and $\|\hat{\boldsymbol{\Omega}}_\rho\|_2 \le p\|\hat{\boldsymbol{\Omega}}_\rho\|_{L1} \le p^2\sqrt{\frac{n}{log(p)}}$. By theorem 1 and because a is large enough, we have

$$\begin{aligned}
\sup_{\boldsymbol{\Omega}\in\mathcal{U}} \mathbb{E}\|\hat{\boldsymbol{\Omega}}_\rho - \boldsymbol{\Omega}\|_2^2 &= \sup_{\boldsymbol{\Omega}\in\mathcal{U}} \mathbb{E}\|\hat{\boldsymbol{\Omega}}_\rho - \boldsymbol{\Omega}\|_2^2 \times I\left\{\max_{ij}|\hat{\sigma}_{ij} - \sigma_{ij}| + \rho \le C_0\sqrt{\log p/n}\right\} \\
&\quad + \sup_{\boldsymbol{\Omega}\in\mathcal{U}} \mathbb{E}\|\hat{\boldsymbol{\Omega}}_\rho - \boldsymbol{\Omega}\|_2^2 \times I\left\{\max_{ij}|\hat{\sigma}_{ij} - \sigma_{ij}| + \rho > C_0\sqrt{\log p/n}\right\} \\
&= O\left(M^{4-4q}s_0^2(p)\left(\frac{\log p}{n}\right)^{1-q}\right) + O\left(p^4\left(\frac{n}{\log p}\right)p^{1-a/2}\right) \\
&= O\left(M^{4-4q}s_0^2(p)\left(\frac{\log p}{n}\right)^{1-q}\right).
\end{aligned} \tag{42}$$

□

Proof of Theorem 6 is similar to that of Theorem 1 and Theorem 5.

5 Numerical results

In this section, we investigate the numerical performance of the RCLIME estimator. We first introduce a quantile method for threshold selection. This data driven method is shown to work well. And the procedure is easy to complete. An R package for computing threshold t_n and robust estimator for covariance matrix has been developed. It is available at xxx. In simulation studies, we compare numerical performance of RCLIME and CLIME both under cases of light-tail distribution and heavy-tail distribution.

5.1 Choice of t

The question of threshold selection is hard to answer analytically. Cross validation is frequently used as a standard way in literature. For example, Bickel and Levina(2008a,b) used the sample covariance matrix of the validating sample as target when minimizing the expectation of norm loss which is estimated by

$$\hat{R}(t) = \frac{1}{N} \sum_{\nu=1}^{N} \|B_t(\hat{\mathbf{\Sigma}}_1^{(\nu)}) - \hat{\mathbf{\Sigma}}_2^{(\nu)}\|,$$

where $\hat{\mathbf{\Sigma}}_1^{(\nu)}$ and $\hat{\mathbf{\Sigma}}_2^{(\nu)}$ are sample covariance matrices of training and validating sample from the νth split. But here, sample covariance matrix is not a good target due to the heavy-tail condition. And we find the cross validation method doesn't work in simulation studies. So instead of using cross validation method, we propose another data driven method called quantile method.

For a given dataset $\{y_i\}_{i=1}^n$, let y_q denote the $q\%$-quantile of the dataset, we define quantile parameter t_q for dataset $\{y_i\}_{i=1}^n$ as

$$t_q = \frac{y_q - y_{1-q}}{2}. \tag{43}$$

We need to solve $p(p+1)$ equations in the form of $\sum_{k=1}^n \rho_{t_n}(y_k - \mu) = 0$ when estimating means for $X_i's$ and $X_iX_j's$ in the estimation process. For each step we are supposed to obtain a threshold t_q using data points $\{X_{ki}\}'s$ or $\{X_{ki}X_{kj}\}'s$ (or $(X_i - \hat{\mu}_i)(X_j - \hat{\mu}_j)'s$ in modification version) and then estimate the corresponding mean.

To show the quantile method is appropriate and also to make a good choice of q, we tried three different values $q = 0.90, 0.95, 0.99$. We generate a random sample distributed as multivariate $t_{2.5}$ with $\mu = 0$, $\omega_{ij} = |i-j|^{0.6}$ and of size $n = 100$, dimension $p = 200$. Figure 1 and Figure 2 show us how the choice of q will effect the estimation of mean. All the three choices are fine since the estimated values are all almost zero, which is true value for mean. Also $q = 0.95$ is a little bit better than other two choices since the truncating bounds look more reasonable. Figure 3 shows how the choice of q effect the estimation of covariances. We plot the true values of $\{\sigma_{1i}\}_{i=1}^p$, robust estimated values and sample values. It can be shown that large value of t (or a large value of q) perform better at the point where the true value is large. When the value is small, small t will bring better estimation results. We also compute the Frobenius norm losses for three estimators computed with different q and the estimated covariance matrix with $q = 0.95$ has the smallest norm loss. As a whole, $q = 0.95$ is a good choice. In the following estimation, we will use the quantile method and use $q = 0.95$.

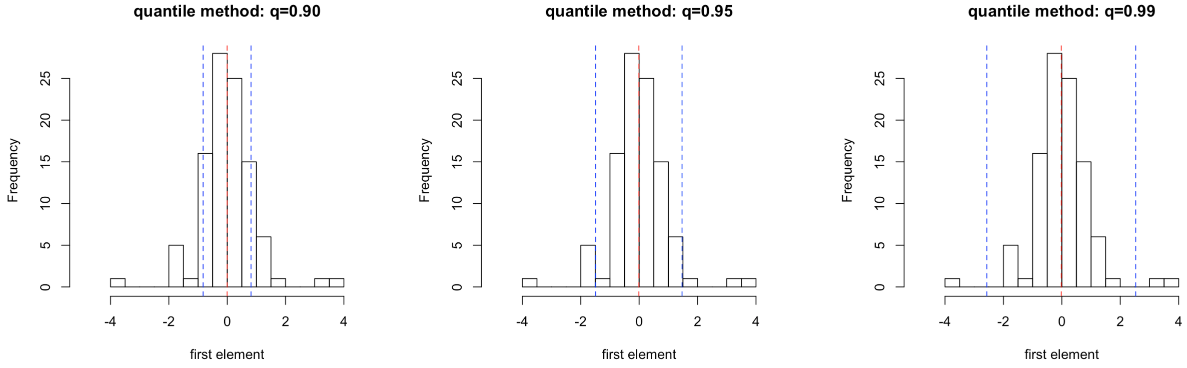

Figure 1: *The data is $t_{2.5}$ distributed with $n = 100, p = 200$. The histogram is for the first element $X'_{1k}s, k = 1, ...100$. Red dashed line is robust mean. Blue dashed lines represent bounds of truncating area.*

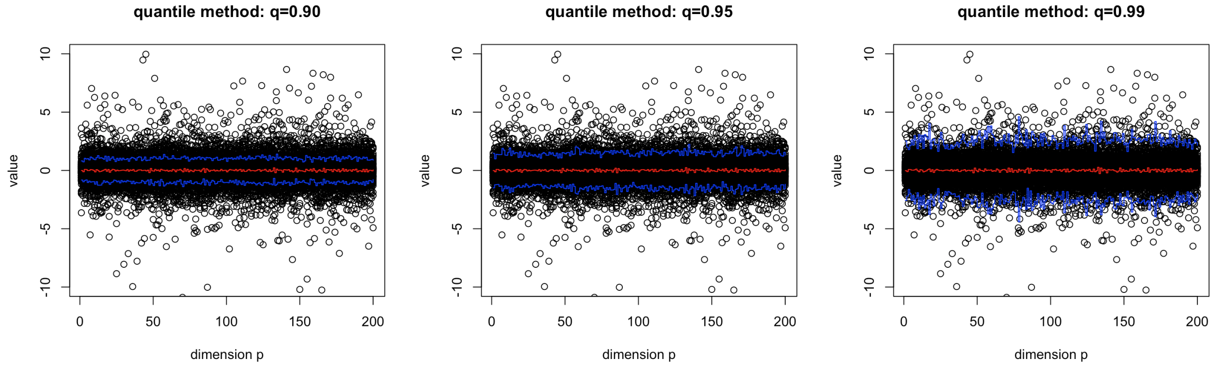

Figure 2: *Plot of data. The data is $t_{2.5}$ distributed with $n = 100, p = 200$. Red line is robust mean. Blue lines represent bounds of truncating area.*

5.2 Simulation results

In this section we investigate numerical performance of the robust estimator for covariance matrix and RCLIME. We estimator $\hat{\mathbf{\Omega}}_{RCLIME}$ and the CLIME estimator $\hat{\mathbf{\Omega}}_{CLIME}$.

In simulation study, we consider five types of distribution.

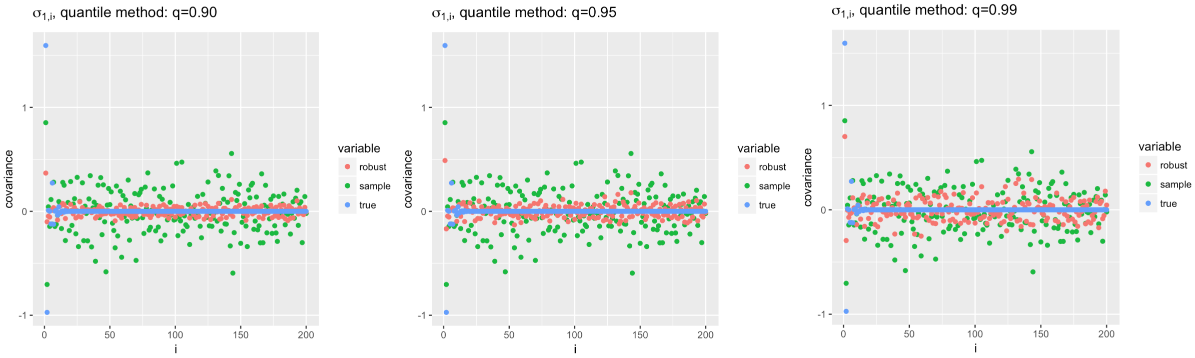

Figure 3: *Plot of covariance cov*(X_1, X_i). *The data is* $t_{2.5}$ *distributed with* $n = 100, p = 200$. *Frobenius norm losses for robust estimators are 5.250974(q = 0.90); 5.230138(q = 0.95); 9.733198 (q = 0.99).*

Type 1. Normal with $\mu = 0$, $\omega_{ij} = |i - j|^{0.6}$

Type 2. Lognormal with $\mu = 0$, $\omega_{ij} = |i - j|^{0.6}$

Type 3. $T_{2.5}$ with $\mu = 0$, $\omega_{ij} = |i - j|^{0.6}$

Type 4. T_3 with $\mu = 0$, $\omega_{ij} = |i - j|^{0.6}$

Type 5. T_5 with $\mu = 0$, $\omega_{ij} = |i - j|^{0.6}$

For each model, we generate a training sample of size n=100. We compute robust covariance estimator $\hat{\mathbf{\Sigma}}$ and RCLIME estimator $\hat{\mathbf{\Omega}}_\lambda$ with 50 different values of λ based on training sample. To validating tuning parameter λ, we generate an independent validating sample of size 100 from the same distribution as training sample. We compute robust estimator $\hat{\mathbf{\Sigma}}_n^v$ for covariance matrix using quantile method discussed above based on the validating sample and compute likelihood loss for the 50 different $\hat{\mathbf{\Omega}}_\lambda$, where the likelihood loss is defined by

$$L(\mathbf{\Sigma}, \hat{\mathbf{\Omega}}_\lambda) =< \hat{\mathbf{\Omega}}_\lambda, \mathbf{\Sigma} > - \log\det(\hat{\mathbf{\Omega}}_\lambda).$$

In practical, we estimate it with

$$\hat{L}(\lambda) =< \hat{\mathbf{\Omega}}_\lambda, \hat{\mathbf{\Sigma}}_n^v > - \log\det(\hat{\mathbf{\Omega}}_\lambda).$$

We use the $\hat{\mathbf{\Omega}}_\lambda$ with the smallest likelihood loss as the final estimator. We compute $\hat{\mathbf{\Omega}}_{CLIME}$ on the same training and validating sample using the same validation scheme. For each type of distribution, we replicate 100 times and consider different values of p=100, 150, 200. We record different types of norm losses for estimated covariance matrix and estimated precision matrix.

From table 4 we can see that robust estimator for covariance matrix perform uniformly better than sample covariance matrix in terms of norm loss. And the norm losses of robust estimator are much smaller than those of sample covariance matrix under heavy tail conditions. From table 3 we see that $\hat{\mathbf{\Omega}}_{RCLIME}$ is comparable with $\hat{\mathbf{\Omega}}_{CLIME}$ in normal case and perform better

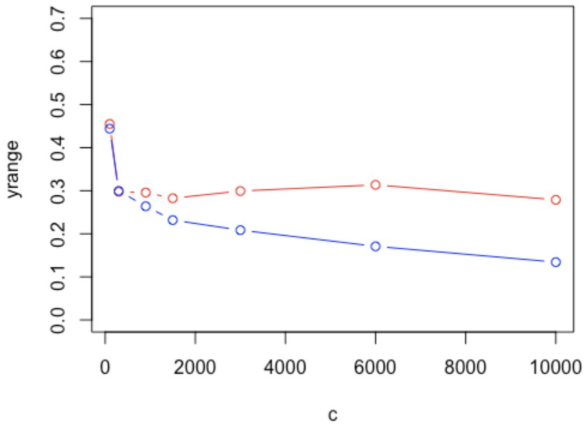

Figure 4: *This plot shows the relationship between dimension and maximum error of inverse of variance. x-axis is dimension p, y-axis is maximum loss of inverse. Here the distribution is multivariate t_3 and we let $n = p$.*

than $\hat{\boldsymbol{\Omega}}_{CLIME}$ in heavy tail cases. The advantage is not as significant as that of robust covariance matrix. It is because in our simulation study the values of elements in precision matrix are small. So even the difference between two estimated covariance matrix is big, the difference between two estimated precision matrix is small. But we can see that advantage of $\hat{\boldsymbol{\Omega}}_{RCLIME}$ tend to be significant when the tail becomes heavier and dimension p gets larger. When p is larger than 200, RCLIME performs uniformly better than CLIME in all type of norm loss. In terms of operator norm, RCLIME reduced about 25% norm loss in $t_{2.5}$ distribution. We can also noticed that the standard error of RCLIME is smaller than that of CLIME, which means robust estimation can produce a more consistent result.

To show the promise of our robust estimators when p is much larger than n, we investigated the variances σ_{ii} in the covariance matrix $\boldsymbol{\Sigma}$ and their inverse σ_{ii}^{-1} instead of investigating the entire covariance matrix due to the cost of computation. We generate samples for model 4 of size n=100 and let p change from 100 to 10000. For each sample we compute and record the diagonals of robust covariance matrix $\hat{\sigma}_{ii,R}$ and sample covariance matrix $\hat{\sigma}_{ii,S}$ and compute their inverse $\hat{\sigma}_{ii,R}^{-1}$, $\hat{\sigma}_{ii,S}^{-1}$. We record the largest error of estimated variance and largest error of estimated inverse of variance for both robust estimator and sample estimator. From figure 4 we can see that the difference between errors of $\hat{\sigma}_{ii,R}^{-1}$ and $\hat{\sigma}_{ii,S}^{-1}$ becomes quite significant when p is larger than 500. So we can expect the advantage of RCILME to be significant when p is large enough.

To better illustrate the performance of robust estimators, figure 5 shows structures of covariance matrix and heatmaps of the nonzeros in the precision matrix. The plots suggest that the structure recovered by robust covariance matrix has a closer resemblance to the true model than sample covariance matrix. For precision matrix, RCLIME can include more true nonzero entries than CLIME on the off-diagonals.

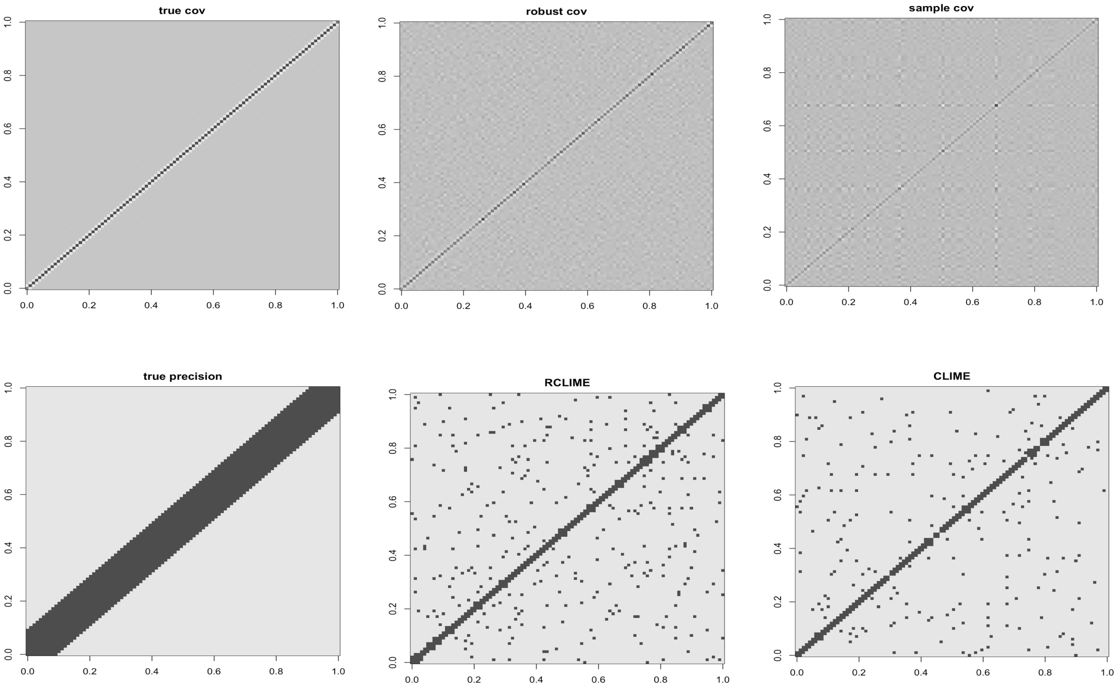

Figure 5: *The distribution here is $t_{2.5}$, n=100, p=100. In plots for covariance matrix White represents for zeros and black represents for 1. In plots for inverse covariance matrix White represents for zeros and black represents for nonzeros.*

6 Discussion

This paper introduces a robust estimator for covariance matrix and a RCLIME estimator for precision matrix. From analysis above we can see that both the robust estimator for covariance matrix and RCLIME estimator have desirable properties. Compared with other estimators, our robust estimators require existence of only $(2+\delta)$th moment and dimension p can still grow exponentially with sample size n. So our estimators can be more widely applied.

Our method and the robust idea can be applied in many other estimation procedures. For estimators based on sample covariance matrix like banded and thresholding estimator for covariance matrix and Glasso estimator for precison matrix, we can replace the sample covariance matrix with our robust estimator. Such extension would be very interesting.

A Appendix

Proof of Lemma 1. The function $\rho_{t_n}(y)$ satisfies:

$$-\log\left(1-\frac{y}{t_n}+\left(\frac{y}{t_n}\right)^2\right)\le\frac{\rho_{t_n}(y)}{t_n}\le\log\left(1+\frac{y}{t_n}+\left(\frac{y}{t_n}\right)^2\right). \tag{44}$$

Define $r(\theta)=n^{-1}\sum_{i=1}^n\rho_{t_n}(y_i-\theta)$. Due to definition of $\hat{\mu}$, $\hat{\mu}$ is solution of $r(\theta)=0$. Using independence of $\{y_i\}_{i=1}^n$, we have

$$\begin{aligned}\mathbb{E}[\exp\{nr(\theta)/t_n\}]&=\Pi_{i=1}^n\mathbb{E}\left[\exp\left\{\rho_{t_n}(y_i-\theta)/t_n\right\}\right]\\&\le\Pi_{i=1}^n\mathbb{E}\left\{1+(y_i-\theta)/t_n+(y_i-\theta)^2/t_n^2\right\}\\&=\left[1+(\mu-\theta)/t_n+\left\{\sigma^2+(\mu-\theta)^2\right\}/t_n^2\right]^n\\&\le\exp\left[n(\mu-\theta)/t_n+n\left\{\sigma^2+(\mu-\theta)^2\right\}/t_n^2\right].\end{aligned}\tag{45}$$

Similarly,

$$\begin{aligned}\mathbb{E}\left[\exp\{-nr(\theta)/t_n\}\right]&=\Pi_{i=1}^n\mathbb{E}[\exp\{\rho_{t_n}(\theta-y_i)/t_n\}]\\&\le\Pi_{i=1}^n\mathbb{E}\left\{1+(\theta-y_i)/t_n+(\theta-y_i)^2/t_n^2\right\}\\&=\left[1+(\theta-\mu)/t_n+\left\{\sigma^2+(\theta-\mu)^2\right\}/t_n^2\right]^n\\&\le\exp\left[-n(\mu-\theta)/t_n+n\left\{\sigma^2+(\mu-\theta)^2\right\}/t_n^2\right].\end{aligned}\tag{46}$$

Define

$$B_+(\theta)=(\mu-\theta)+\frac{\sigma^2+(\mu-\theta)^2}{t_n}+\frac{at_n\log(p)}{n}, \tag{47}$$

$$B_-(\theta)=(\mu-\theta)-\frac{\sigma^2+(\mu-\theta)^2}{t_n}-\frac{at_n\log(p)}{n}. \tag{48}$$

By Chebyshev inequality we have,

$$\begin{aligned} P\left(r(\theta) > B_+(\theta)\right) &= P\left(\exp\{nr(\theta)/t_n\} > \exp\left\{nB_+(\theta)/t_n\right\}\right) \\ &\leq \frac{\mathbb{E}\left[\exp\{nr(\theta)/t_n\}\right]}{\exp\left[n(\mu-\theta)/t_n + n\left\{\sigma^2+(\mu-\theta)^2\right\}/t_n^2 + a\log(p)\right]} \\ &\leq p^{-a}, \end{aligned} \tag{49}$$

and

$$\begin{aligned} P(r(\theta) < B_-(\theta)) &= P(-r(\theta) > -B_-(\theta)) \\ &\leq \frac{\mathbb{E}\left[\exp\{nr(\theta)/t_n\}\right]}{\exp\left[-n(\mu-\theta)/t_n + n\left\{\sigma^2+(\mu-\theta)^2\right\}/t_n^2 + a\log(p)\right]} \\ &\leq p^{-a}. \end{aligned} \tag{50}$$

So we can get,

$$P\left(B_-(\theta) \leq r(\theta) \leq B_+(\theta)\right) > 1 - 2p^{-a}. \tag{51}$$

Let θ_+ be the smaller solution of $B_+(\theta) = 0$, θ_- be the larger solution of $B_-(\theta) = 0$. Under assumption of $\frac{a\log(p)}{n} \leq \frac{1}{8}$, $\nu > \sigma$ and choice of $t_n = \sqrt{\frac{\nu^2}{a\log(p)/n}}$, we have $1 - 4(\sigma^2/t_n^2 + a\log(p)/n) > 1 - 8a\log(p)/n > 0$. Then we have

$$\begin{aligned} \theta_+ &= \mu - 2\left\{\frac{\sigma^2}{t_n} + \frac{at_n\log(p)}{n}\right\}\left[-1 - \sqrt{1 - 4\left\{\frac{\sigma^2}{t_n^2} + \frac{a\log(p)}{n}\right\}}\right]^{-1} \\ &< \mu + 2\left\{\frac{\sigma^2}{t_n} + \frac{at_n\log(p)}{n}\right\} \\ &< \mu + 4\nu\sqrt{\frac{a\log(p)}{n}}. \end{aligned} \tag{52}$$

Similarly,

$$\begin{aligned} \theta_+ &= \mu + 2\left\{\frac{\sigma^2}{t_n} + \frac{at_n\log(p)}{n}\right\}\left[-1 - \sqrt{1 - 4\left\{\frac{\sigma^2}{t_n^2} + \frac{a\log(p)}{n}\right\}}\right]^{-1} \\ &> \mu - 2\left\{\frac{\sigma^2}{t_n} + \frac{at_n\log(p)}{n}\right\} \\ &> \mu - 4\nu\sqrt{\frac{a\log(p)}{n}}. \end{aligned} \tag{53}$$

$r(\theta)$ is monotone decreasing function of θ, so under event $\{B_-(\theta) \leq r(\theta) \leq B_+(\theta)\}$, $\theta_- \leq \hat{\mu} \leq \theta_+$. So we have

$$\begin{aligned} P\left(\mu - 4\nu\sqrt{\frac{a\log(p)}{n}} < \hat{\mu} < \mu + 4\nu\sqrt{\frac{a\log(p)}{n}}\right) &> P\left(\theta_- \leq \hat{\mu} \leq \theta_+\right) \\ &> P\left(B_-(\theta) \leq r(\theta) \leq B_+(\theta)\right) \\ &> 1 - 2p^{-a}. \end{aligned} \tag{54}$$

Then we get result

$$P\left(|\hat{\mu}-\mu| \geq 4\nu\sqrt{\frac{a\log p}{n}}\right) \leq 2p^{-a}. \tag{55}$$

□

Proof of Lemma 2. Note that $P(|\hat{\mu}-\mu| \geq t) \leq P(\hat{\mu}-\mu \geq t) + P(\hat{\mu}-\mu \leq -t)$. Let us focus on $P(\hat{\mu}-\mu \geq t)$ first. It can be noticed that $\rho_k(x)$ is a nondecreasing function. So, event $\{\hat{\mu}-\mu \geq t\}$ implies $\{\sum_{i=1}^n \rho_k(x_i-(\mu+t)) \geq 0\}$. Thus,

$$P(\hat{\mu}-\mu \geq t) \leq P\left(\sum_{i=1}^n \left[\rho_k\left\{x_i-(\mu+t)\right\}+y\right] \geq ny\right), \tag{56}$$

where $y = \mathbb{E}\rho_k\left[X-(\mu+t)\right]$. Since $|\rho_k\left\{x_i-(\mu+t)\right\}+y| \leq 2k$, we hope to use Bernstein inequality. If we can choose k and t such that $y \geq 0$ and bound $V = \operatorname{var}\left(\rho_k\left\{X-(\mu+t)\right\}\right)$, we have the following inequality,

$$P\left(\sum_{i=1}^n \left[\rho_k\left\{x_i-(\mu+t)\right\}+y\right] \geq ny\right) \leq \exp\left\{-\frac{(ny)^2/2}{\frac{2}{3}nyK+nV}\right\}. \tag{57}$$

Next we need to check if $y > 0$ and bound V. We can assume that $\mu = 0$ w.l.g. Then we can write y as

$$y = \mathbb{E}\rho_k(X) - \mathbb{E}\rho_k(X-t) - \mathbb{E}(\rho_k(X)). \tag{58}$$

Note that

$$\begin{aligned}|\mathbb{E}\rho_k(X)| =& |\mathbb{E}\left\{X1_{\{|X|\leq k\}} + \operatorname{sign}(X)k1_{\{|X|\geq k\}}\right\}| \\ \leq& \mathbb{E}|X|\frac{k^{1-\theta}}{|X|^{1-\theta}} + k\mathbb{E}\frac{|X|^\theta}{k^\theta} \\ =& 2k^{1-\theta}m_\theta,\end{aligned} \tag{59}$$

and

$$\begin{aligned}\mathbb{E}\rho_k(X) - \mathbb{E}\rho_k(X-t) =& \mathbb{E}\int_{X-t}^{X} 1_{|u|\leq k}du \\ =& \mathbb{E}\int_R 1_{|u|\leq k}1_{\{X-t<u<X\}}du \\ =& \int_{-k}^{k} P\left(u<X<u+t\right)du \\ =& \int_{-k+t}^{k+t} F(v)dv - \int_{-k}^{k} F(v)dv \\ =& \int_{k}^{k+t} F(v)dv - \int_{-k}^{-k+t} F(v)dv \\ =& t - \left(\int_{k}^{k+t} P(X>u)du - \int_{-k}^{-k+t} P(X\leq u)du\right).\end{aligned} \tag{60}$$

If $t < \frac{k}{2}$, we have

$$\begin{aligned}0 \leq &\int_k^{k+t} P(X > u)du - \int_{-k}^{-k+t} P(X \leq u)du \\ \leq &\int_{k/2}^{\infty} P(x \geq u)du + \int_{-\infty}^{-k/2} P(x \leq u)du \\ = &\int_{k/2}^{\infty} \mathbb{E}1_{\{|X| \geq u\}} du \\ \leq &\int_{k/2}^{\infty} \mathbb{E}\frac{|X|^\theta}{u^\theta} du \\ = &\frac{m_\theta}{\theta - 1} 2^{\theta-1} k^{1-\theta}.\end{aligned} \tag{61}$$

Combining (44)-(47) y can be bounded as

$$t - \left(\frac{2^{\theta-1}}{\theta - 1} + 2\right) m_\theta k^{1-\theta} \leq y \leq t + 2m_\theta k^{1-theta} \tag{62}$$

Hence, if $t - (\frac{2^{\theta-1}}{\theta-1} + 2)m_\theta k^{1-\theta} > 0$then we can get $y > 0$. Under this condition we can use Bernstein Inequality to get (56). Next we deal with the variance term V. Let X' be a random variable independent of X and has same distribution as X. Then,

$$\begin{aligned}V = &\text{var}\left(\rho_k(X - t)\right) \\ = &\frac{1}{2}\mathbb{E}\left(\rho_k(X - t) - \rho_k(X' - t)\right)^2 \\ \leq &\frac{1}{2}\mathbb{E}\left[\min\left\{|X - X'|, 2k\right\}\right]^2 \\ = &2k^2\mathbb{E}\left[\min\left\{\frac{|X - X'|}{k}, 1\right\}\right]^2 \\ \leq &2k^{2-\theta}\mathbb{E}|X - X'|^\theta \\ \leq &2k^{2-\theta}\mathbb{E}\left\{(2|X|)^\theta + (2|X'|)^\theta\right\} \\ = &2^{2+\theta} m_\theta k^{2-\theta}.\end{aligned} \tag{63}$$

If $\theta \geq 2$, we can have $V \leq 16m_\theta$. Combining (42), (43), (48) and (49), we can bound $P(\hat{\mu} - \mu \geq t)$ as follows,

$$\begin{aligned}&P(\hat{\mu} - \mu \geq t) \\ \leq &\exp\left\{-\frac{(ny)^2/2}{\frac{2}{3}nyk + nV}\right\} \\ \leq &\exp\left\{-\frac{3ny}{8k}\right\} + \exp\left\{-\frac{ny^2}{4V}\right\} \\ \leq &\exp\left\{-\frac{3n\left(t - \left(\frac{2^{\theta-1}}{\theta-1} + 2\right) m_\theta k^{1-\theta}\right)}{8k}\right\} + \exp\left\{-\frac{n\left(t - \left(\frac{2^{\theta-1}}{\theta-1} + 2\right) m_\theta k^{1-\theta}\right)^2}{2^{4+\theta} m_\theta k^{2-\theta}}\right\}.\end{aligned} \tag{64}$$

Similarly we can get the same bound for $P(\hat{\mu} - \mu \leq -t)$ and get the final exponential-type tail bound for $P(|\hat{\mu} - \mu| \geq t)$.

□

References

[1] Iain M Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of statistics*, pages 295–327, 2001.

[2] Iain M Johnstone and Arthur Yu Lu. Sparse principal components analysis. *Unpublished manuscript*, 7, 2004.

[3] Peter J Bickel and Elizaveta Levina. Regularized estimation of large covariance matrices. *The Annals of Statistics*, pages 199–227, 2008.

[4] T Tony Cai, Cun-Hui Zhang, Harrison H Zhou, et al. Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4):2118–2144, 2010.

[5] T Tony Cai, Harrison H Zhou, et al. Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics*, 40(5):2389–2420, 2012.

[6] Peter J Bickel, Elizaveta Levina, et al. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008.

[7] Wei Biao Wu and Mohsen Pourahmadi. Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 90(4):831–844, 2003.

[8] Adam J Rothman, Peter J Bickel, Elizaveta Levina, Ji Zhu, et al. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

[9] Clifford Lam and Jianqing Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of statistics*, 37(6B):4254, 2009.

[10] Ming Yuan. High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11(Aug):2261–2286, 2010.

[11] Tony Cai, Weidong Liu, and Xi Luo. A constrained 1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

[12] Peter J Huber. Robust estimation of a location parameter. *The annals of mathematical statistics*, pages 73–101, 1964.

Table 1: *Comparison of average(SE) losses for precision matrix over 100 replications*

	Normal		Lognormal		$T_{2.5}$	
p	$\hat{\Omega}_{RCLIME}$	$\hat{\Omega}_{CLIME}$	$\hat{\Omega}_{RCLIME}$	$\hat{\Omega}_{CLIME}$	$\hat{\Omega}_{RCLIME}$	$\hat{\Omega}_{CLIME}$
			Operator norm			
100	2.616(0.085)	2.603(0.133)	2.176(0.187)	2.808(0.114)	2.237(0.392)	2.490(0.404)
150	2.666(0.029)	2.696(0.027)	2.270(0.217)	2.898(0.090)	2.312(0.477)	2.692(0.527)
200	2.784(0.025)	2.832(0.021)	2.319(0.190)	2.955(0.092)	2.211(0.298)	2.939(0.555)
			Frobenius norm			
100	9.168(0.334)	9.054(0.551)	9.261(0.296)	9.933(0.433)	10.554(1.099)	9.764(0.987)
150	11.508(0.093)	11.576(0.092)	11.652(0.286)	12.637(0.477)	13.223(1.809)	12.922(2.567)
200	14.002(0.115)	14.222(0.071)	13.391 (0.231)	14.931(0.544)	14.78(1.279)	16.047(2.966)
			Matrix l_∞ norm			
100	3.108(0.069)	3.096(0.064)	3.800(0.376)	3.334(0.086)	4.585(1.049)	3.708(0.996)
150	3.175(0.073)	3.153(0.054)	4.027(0.406)	3.378(0.064)	4.837(2.048)	4.116(1.961)
200	3.191(0.067)	3.192(0.040)	3.907 (0.331)	3.457(0.206)	4.441(1.398)	4.877(3.368)
	T_3		T_5			
			Operator norm			
100	2.218(0.159)	2.700(0.338)	2.477(0.111)	2.696(0.129)		
150	2.308(0.315)	2.793(0.276)	2.606(0.086)	2.789(0.145)		
200	2.408(0.179)	2.930(0.303)	2.674(0.086)	2.844(0.100)		
			Frobenius norm			
100	8.748(0.205)	9.840(1.228)	8.790(0.319)	9.502(0.523)		
150	11.125(0.931)	12.454(1.304)	11.347(0.347)	12.143(0.726)		
200	13.116(0.397)	15.440(1.863)	2.674(0.086)	2.844(0.100)		
			Matrix l_∞ norm			
100	3.456(0.396)	3.420(0.467)	3.156(0.167)	3.255(0.110)		
150	3.646(1.343)	3.4662(0.603)	3.187(0.160)	3.365(0.395)		
200	3.506(0.357)	4.408(2.515)	3.302(0.202)	3.382(0.199)		

Table 2: *Comparison of average(SE) losses for covariance matrix over 100 replications*

	Normal		Lognormal		$T_{2.5}$	
p	$\hat{\Sigma}_n$	S	$\hat{\Sigma}_n$	S	$\hat{\Sigma}_n$	S
			Operator norm			
100	5.519(0.478)	7.577(0.593)	4.492(0.109)	14.761(6.798)	4.795(0.115)	87.120(188.185)
150	7.396(0.412)	9.936(0.507)	4.736(0.074)	19.038(9.602)	5.054(0.178)	128.529(452.040)
200	9.059(0.463)	11.945(0.622)	4.930(0.063)	24.844(17.321)	5.373(0.719)	179.857(479.199)
			Frobenius norm			
100	19.575(0.373)	22.772(0.548)	18.414(0.277)	29.266(5.610)	20.440(0.537)	97.836(185.926)
150	28.951(0.378)	34.078(0.5492)	23.701(0.203)	41.933(9.344)	26.374(0.391)	146.610(450.839)
200	38.233(0.418)	45.256(0.593)	28.658(0.169)	54.475(13.733)	31.851(0.774)	200.610(475.144)
			Matrix l_∞ norm			
100	19.995(1.150)	24.648(1.612)	11.442(0.614)	46.247(15.940)	12.806(1.534)	192.655(412.38)
150	29.216(1.340)	35.637(1.632)	15.788(0.805)	68.878(24.886)	17.888(2.569)	305.676(1099.71)
200	38.706(1.367)	47.237(1.992)	20.414(0.984)	97.094(45.120)	23.007(3.483)	437.932(1168.327)

	T_3		T_5	
			Operator norm	
100	4.580(0.175)	58.214(102.134)	4.854(0.615)	25.477(35.258)
150	5.230(0.664)	57.414(89.184)	6.697(1.077)	27.066(19.161)
200	6.226(1.123)	117.439(173.185)	8.454(1.158)	33.702(20.324)
			Frobenius norm	
100	19.138(0.324)	71.173(99.993)	19.100(0.687)	39.041(32.689)
150	26.001(0.765)	76.794(86.946)	27.539(1.428)	49.393(17.889)
200	32.486(1.213)	141.098(169.215)	35.851(1.948)	63.915(17.349)
			Matrix l_∞ norm	
100	15.792(1.499)	130.219(207.031)	18.508(1.725)	62.983(78.757)
150	22.237(2.815)	135.143(176.667)	27.021(2.683)	72.174(45.386)
200	28.582(3.096)	285.023(393.643)	35.871(3.359)	91.916(42.768)