

ROBUST ESTIMATION OF THE MEAN AND COVARIANCE MATRIX FOR HIGH DIMENSIONAL TIME SERIES

Danna Zhang

University of California, San Diego

Abstract: High-dimensional nonGaussian time series data are becoming increasingly common. However, the conventional methods used to estimate mean vectors and second-order characteristics are inadequate for ultrahigh-dimensional and heavy-tailed data. Therefore, we use a framework of functional dependence measures to establish a Bernstein-type inequality under dependence. Then, we investigate a Huber estimator for the mean for a high-dimensional time series with $(1 + \epsilon)$ th moments, for some $0 < \epsilon \leq 1$, and establish a phase transition for Huber estimators. The transition admits nearly subGaussian concentration around the unknown mean for $\epsilon = 1$, and a slower convergence rate if $0 < \epsilon < 1$. We also investigate Huber-type estimators for the covariance and precision matrices of the process with $(2 + 2\epsilon)$ th moments, for some $0 < \epsilon \leq 1$, and present the convergence rates for robust modifications of the regularized estimators. Similarly, a phase transition occurs between $\epsilon = 1$ and $0 < \epsilon < 1$. As a significant improvement, the dimension can be allowed to increase exponentially with the sample size to ensure consistency under very mild moment conditions. Numerical results indicate that the Huber estimates perform well.

Key words and phrases: Bernstein-type inequality, covariance and precision matrix, heavy tailed data, high dimensional time series, Huber estimation, phase transition, regularized estimation. 

1. Introduction

The recent widespread increase in the collection of high-dimensional data has led to numerous methodologies and theories for analyzing such data. Suppose we have identically distributed observations $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^p$. We wish to estimate the mean vector $\mu = (\mu_1, \dots, \mu_p)^\top = \mathbb{E}\mathbf{x}_i$ when the dimension p can be much larger than the sample size n . A simple, natural, and popular method of doing so is to calculate the sample mean vector $\bar{\mathbf{x}} = n^{-1} \sum_{i=1}^n \mathbf{x}_i$. When $\mathbf{x}_1, \dots, \mathbf{x}_n$ are independent and identically distributed (i.i.d.) Gaussian or subGaussian, nice

Corresponding author: Danna Zhang, Department of mathematics, University of California, San Diego, La Jolla, CA 92093, USA. E-mail: daz076@ucsd.edu.

performance bounds can be derived with the help of concentration inequalities; see Chapter 14 of Bühlmann and Van De Geer (2011) for a review of many useful inequalities.

The covariance matrix and inverse covariance (precision) matrix play a fundamental role in characterizing the second-order properties of high-dimensional data. Denote the covariance matrix and precision matrix by $\Sigma_{\mathbf{x}} = \mathbb{E}[(\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^\top]$ and $\Omega_{\mathbf{x}} = \Sigma_{\mathbf{x}}^{-1}$, respectively. It is well known that the sample covariance matrix

$$\hat{\Sigma}_{\mathbf{x}} = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top \quad (1.1)$$

is not a consistent estimator of $\Sigma_{\mathbf{x}}$, and that we cannot use its inverse to estimate the precision matrix $\Omega_{\mathbf{x}}$, owing to its singularity when $p > n$. Theories related to estimating covariance matrices and their inverses for high-dimensional i.i.d. data have developed significantly. For example, various regularization methods have been investigated for estimating $\Sigma_{\mathbf{x}}$, starting from the sample covariance matrix $\hat{\Sigma}_{\mathbf{x}}$. Such methods include thresholding (Bickel and Levina (2008b), El Karoui (2008)) and its variants (Rothman, Levina and Zhu (2009), Cai and Liu (2011)), banding (Bickel and Levina (2008a)), and tapering (Cai and Zhou (2012)), among others. In additions many alternatives to regularized estimates have been considered; see Cai, Liu and Zhou (2016) for a review.

Most theoretical investigations assume that $\mathbf{x}_1, \dots, \mathbf{x}_n$ are i.i.d. Gaussian or subGaussian random vectors, which is quite restrictive. On the one hand, the assumption of independence may not be valid for temporally observed data in many fields, including finance, signal processing, neuroimaging, meteorology, and seismology. As a result, regularized estimations were later generalized to include high-dimensional time series; see Chen, Xu and Wu (2013), McMurry and Politis (2015) and Basu and Michailidis (2015), among many others. On the other hand, high-dimensional time series data are often drawn from non-subGaussian or even heavy-tailed distributions. For example, being able to estimate the covariance and precision matrices for high-dimensional time series drawn from non-subGaussian distributions is becoming a crucial problem in fields such as portfolio allocation (Kim et al. (2012)), risk management (Koopman and Lucas (2008)), and brain network (Friston (2011)) and geophysical dynamic studies (Kondrashov et al. (2005)). Here, Chen, Xu and Wu (2013) attempt to do so by quantifying the convergence rates of covariance and precision matrix estimators, and Zhang and Wu (2017) provide Gaussian approximations for the sample mean vector and sample covariance matrix. Both assume the underlying process has finite q th mo-

ments, for some $q > 4$, and allow the dimension p to increase polynomially with the sample size n as a natural requirement of consistency.

If the process is not Gaussian, Huber (1964) remarked that “the sample mean then may have a catastrophically bad performance...” Motivated by Huber (1964) and Huber (1973), Huber’s estimator of the mean μ_j , for $1 \leq j \leq p$, based on the observations $\mathbf{x}_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{ip})^\top$, for $1 \leq i \leq n$, is defined as the solution to the equation

$$\sum_{i=1}^n \varphi_\kappa(\mathbf{x}_{ij} - \theta) = 0, \tag{1.2}$$

where $\varphi_\kappa(x) = (x \wedge \kappa) \vee (-\kappa)$ is the Huber function with the robustification parameter $\kappa > 0$. The properties of Huber estimators with a fixed robustification parameter have been well studied in regression settings; see, for example, Huber (1973), Catoni (2012), and Fan, Li and Wang (2017), among others. Robust estimation has also been applied to matrices such as the covariance matrix for the i.i.d. case. Here, notable works include those of Catoni (2016), Minsker (2016), Fan, Li and Wang (2017), and Avella-Medina et al. (2018).

There is limited research on the theoretical properties of robust estimators for high-dimensional time series with finite q th moments. To the best of our knowledge, whether the moment order $q \leq 4$ and the dimension p can be ultrahigh with $\log p = o(n^c)$ for some $c > 0$ remains an open problem. In this study, we solve this problem by establishing the consistency and deriving the convergence rates for Huber estimators of the mean vector, covariance matrix, and precision matrix for a large class of time series, taking into account the following: (i) the complex dynamics of the data-generating system; (ii) temporal dependence; (iii) high-dimensional data; and (iv) mild moment conditions. The latter are new features that are quite distinct from those of classical problems. We consider p -dimensional stationary processes of the form

$$\mathbf{x}_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{ip})^\top = \mathbf{h}(\varepsilon_i, \varepsilon_{i-1}, \dots), \tag{1.3}$$

where ε_i , for $i \in \mathbb{Z}$, are i.i.d. random elements, and $\mathbf{h} = (h_1, \dots, h_p)^\top$ is an \mathbb{R}^p -valued measurable function, such that \mathbf{x}_i is well defined. In the univariate case, where $p = 1$, the framework defined in (1.3) provides a natural paradigm for both linear and nonlinear time series models, and represents a large class of stationary processes that appear frequently in practice; see Wiener (1958), Rosenblatt (1971), Priestley (1988), Tong (1990), and Wu (2005), among many others. By allowing the data-generating function to be \mathbb{R}^p -valued, where p may

diverge to infinity, we can extend many existing low dimensional stationary processes to their high-dimensional counterparts in a natural way; see Chen, Xu and Wu (2013), Wu and Wu (2016), and Zhang and Wu (2017)) for examples.

Analyzing such data presents a great challenge and requires new statistical methods and tools. In Section 2, we establish a sharp Bernstein-type inequality under dependence, which is the main tool used to obtain the performance bounds of Huber estimators. We expect that our inequality to be useful in other high-dimensional inference problems that involve dependent data. In Section 3, we consider a Huber estimation for means with $(1 + \epsilon)$ th ($0 < \epsilon \leq 1$) moments. A phase transition can be observed for Huber estimators that admits nearly subGaussian concentration around the unknown mean for $\epsilon = 1$, and a slower convergence rate if $0 < \epsilon < 1$. In Section 4, we consider a Huber-type estimator for a covariance matrix, and establish the convergence rates under an element-wise maximum norm. As a significant improvement, $\log p = o(n/(\log n)^2)$ can be allowed for consistency under very mild moment conditions on the underlying process. In contrast, previous results allowed p to increase only polynomially with n under finite polynomial moments; see, for example, Bickel and Levina (2008a), Cai and Liu (2011), and Chen, Xu and Wu (2013). Using the Huber-type covariance matrix estimator as a pilot estimator, we investigate regularized estimators of the covariance and precision matrix, and verify the nice performance of the spectral norm convergence. In Section 5, we conduct a simulation study to assess the empirical performance of the Huber mean estimators. All proofs are relegated to the online Supplementary Material.

We now introduce some notation. For a random variable X and $q \geq 1$, we define $\|X\|_q = (\mathbb{E}|X|^q)^{1/q}$. For a vector $v = (v_1, \dots, v_p)^\top \in \mathbb{R}^p$, we define $|v|_\infty = \max_j |v_j|$ and $|v|_1 = \sum_{j=1}^p |v_j|$. For a matrix $A = (a_{ij})_{i,j=1}^p \in \mathbb{R}^{p \times p}$, define the matrix ℓ_1 -norm $\|A\|_{\ell_1} = \max_{1 \leq j \leq p} \sum_{i=1}^p |a_{ij}|$, element-wise ℓ_∞ -norm $|A|_\infty = \max_{i,j} |a_{ij}|$, element-wise ℓ_1 -norm $|A|_1 = \sum_{i,j=1}^p |a_{ij}|$, and spectral norm $\rho(A) = \sqrt{\lambda_{\max}(A^\top A)}$, where λ_{\max} denotes the largest eigenvalue. Write the $p \times p$ identity matrix as I_p . We use C, C', \dots to denote positive constants, with values that may differ between contexts.

2. Bernstein-type Inequality under Dependence

The well-known Bernstein inequality (Bernstein (1946)) provides an exponential concentration result for sums of uniformly bounded independent random variables. Let X_1, \dots, X_n be independent random variables, such that $\mathbb{E}X_i = 0$,

$\sigma_i^2 = \text{Var}(X_i) < \infty$, and $|X_i| \leq M$, for all i . Denote $S_n = \sum_{i=1}^n X_i$. Then, for any $x > 0$, we have

$$\mathbb{P}(S_n \geq x) \leq \exp \left\{ -\frac{x^2}{2 \sum_{i=1}^n \sigma_i^2 + 2Mx/3} \right\}. \tag{2.1}$$

It is well known that the Bernstein-type inequality also holds if, rather than being uniformly bounded, X_i has finite exponential moments. Inequality (2.1) suggests two types of bounds for the tail probability: a subGaussian-type tail $\exp\{-x^2/(C \sum_{i=1}^n \sigma_i^2)\}$, in terms of the variance of S_n ; and a sub-exponential-type tail $\exp\{-x/(CMx)\}$, involving the uniform bound M .

Establishing exponential-type tail probability inequalities for dependent sequences is a challenging problem. Here, relevant works include, for example, exponential-type inequalities derived for sums of Markov chains by Douc, Guillin and Moulines (2008) (Theorem 10) under some drift condition, and by Adamczak (2008) (Theorem 6) under the minorization condition. Merlevède, Peligrad and Rio (2009, 2011) derived Bernstein-type bounds for sums of strong mixing processes. In this section, we consider stationary processes of the form given in (1.3) when $p = 1$, which we abbreviate as

$$X_i = h(\varepsilon_i, \varepsilon_{i-1}, \dots), \tag{2.2}$$

where h is a real-valued function. To establish a concentration inequality, we need to introduce appropriate dependence measures. Following Wu (2005), we adopt the following framework of functional dependence measures: If $\|X_i\|_q < \infty$, for some $q \geq 1$, define the dependence measure at lag $i \geq 0$ as

$$\delta_{i,q} = \|X_i - X_{i,\{0\}}\|_q = \|h(\varepsilon_i, \varepsilon_{i-1}, \dots) - h(\varepsilon_i, \dots, \varepsilon_1, \varepsilon'_0, \varepsilon_{-1}, \dots)\|_q, \tag{2.3}$$

where ε'_i is an i.i.d. copy of ε_i . By convention, $\delta_{i,q} = 0$ for all $i < 0$. The dependence measure $\delta_{i,q}$ quantifies the q th moment of the difference between the original process X_i and the decoupled process $X_{i,\{0\}}$, with ε_0 replaced by ε'_0 , and other innovations kept the same. Thus, it measures the effect of ε_0 on the process X_i , which can be interpreted as a possibly nonlinear impulse response function. Assume that there exists some constant $\rho \in (0, 1)$, such that

$$\|X.\|_q := \sup_{m \geq 0} \rho^{-m} \sum_{i=m}^{\infty} \delta_{i,q} < \infty. \tag{2.4}$$

Here, $\|X\|_q$ is called the q th dependence-adjusted moment (DAM) of the process, and the property in (2.4) is called a geometric moment contraction (GMC(q)). Note that in the special case of independent sequences, $\|X\|_q$ is equivalent to the q th moment $\|X_i\|_q$. In this sense, we can interpret the DAM as the moment accounting for dependence.

Theorem 1 below provides a Bernstein-type inequality for the process in (2.2), assuming boundedness and a finite second DAM. The exponential inequality (2.6) is characterized by the DAM $\|X\|_2$, uniform bound M , and dependence parameter ρ , which determines the values of the constants C_1 and C_2 in the inequality.

Theorem 1. *Let (X_i) be the process in (2.2), and let $S_n = \sum_{i=1}^n X_i$. Assume $\mathbb{E}X_i = 0$, $|X_i| \leq M$ for all i , and $\|X\|_2 < \infty$ for some $\rho \in (0, 1)$. In addition, assume $n \geq 4 \vee (\log(\rho^{-1})/2)$. For any $t > 0$, such that $t < (C_2 M)^{-1}(\log n)^{-2}$, we have*

$$\log \mathbb{E} \exp(tS_n) \leq \frac{C_1 t^2 (n\|X\|_2^2 + M^2)}{1 - C_2 t M (\log n)^2}, \quad (2.5)$$

which further implies the Bernstein-type inequality: for $x > 0$,

$$\mathbb{P}(S_n \geq x) \leq \exp \left\{ - \frac{x^2}{4C_1(n\|X\|_2^2 + M^2) + 2C_2 M (\log n)^2 x} \right\}, \quad (2.6)$$

where $C_1 = 2 \max\{(e^4 - 5)/4, [\rho(1 - \rho) \log(\rho^{-1})]^{-1}\} \cdot (8 \vee \log(\rho^{-1}))^2$, $C_2 = \max\{(c \log 2)^{-1}, [1 \vee (\log(\rho^{-1})/8)]\}$ with $c = [\log(\rho^{-1})/8] \wedge \sqrt{(\log 2) \log(\rho^{-1})/4}$.

Remark 1 (Sharpness of Theorem 1). If $\|X\|_2 = O(1)$, compared with the classical Bernstein inequality (2.1) for independent processes, our result (2.6) is not far off with an additional $(\log n)^2$ order in the sub-exponential-type tail. Theorem 6 in Adamczak (2008) provides a slightly sharper inequality involving only an additional $\log n$ order:

$$\mathbb{P}(S_n \geq x) \leq C \exp \left\{ - \frac{1}{C} \min \left(\frac{x^2}{n\nu^2}, \frac{x}{\log n} \right) \right\},$$

where $S_n = \sum_{i=1}^n X_i$, $X_i = \sum_{j=1}^n f(Y_j)$, (Y_i) is a Markov chain satisfying some minorization condition, f is a bounded function, and $\nu^2 = \lim_{n \rightarrow \infty} \text{Var}(S_n/\sqrt{n})$. Our result is as sharp as that established in Theorem 2 of Merlevède, Peligrad and Rio (2009), up to a multiplicative constant in the exponential function:

$$\mathbb{P}(S_n \geq x) \leq \exp \left\{ - \frac{Cx^2}{n\nu^2 + M^2 + M(\log n)^2 x} \right\}, \quad (2.7)$$

where (X_i) is a strong mixing process with mean zero, bounded by M .

Conveniently, our framework provides a neat closed form of the upper bound of the long-run variance ν^2 in terms of the DAM. Define the projection operator $\mathcal{P}_j \cdot = \mathbb{E}(\cdot | \varepsilon_j, \varepsilon_{j-1}, \dots) - \mathbb{E}(\cdot | \varepsilon_{j-1}, \varepsilon_{j-2}, \dots)$. Then, we can write $X_i = \sum_{h=0}^{\infty} \mathcal{P}_{i-h} X_i$. By the orthogonality of \mathcal{P}_j , triangle inequality, and Hölder inequality, we have

$$\begin{aligned} |\text{Cov}(X_0, X_k)| &= \left| \sum_{h=0}^{\infty} \mathbb{E}[(\mathcal{P}_{-h} X_0)(\mathcal{P}_{-h} X_k)] \right| \leq \sum_{h=0}^{\infty} \left| \mathbb{E}[(\mathcal{P}_{-h} X_0)(\mathcal{P}_{-h} X_k)] \right| \\ &\leq \sum_{h=0}^{\infty} \|\mathcal{P}_{-h} X_0\|_2 \|\mathcal{P}_{-h} X_k\|_2 \leq \sum_{h=0}^{\infty} \delta_{h,2} \delta_{h+k,2}, \end{aligned}$$

where the final step follows from $\|\mathcal{P}_j X_i\|_2 \leq \delta_{i-j,2}$, given Jensen's inequality. Hence, it follows that

$$\nu^2 = \sum_{k=-\infty}^{\infty} |\text{Cov}(X_0, X_k)| \leq 2 \sum_{k=0}^{\infty} \sum_{h=0}^{\infty} \delta_{h,2} \delta_{h+k,2} \leq 2 \|X\|_2^2. \tag{2.8}$$

A few comments on the conditions of Theorem 1 are in order. First, we require the GMC condition in (2.4) to depict the dependence. This is an easily verified condition, satisfied by many linear and nonlinear time series models; see Wu (2005) and Shao and Wu (2007) for examples. As noted in Section 5 of the latter paper, the contraction conditions widely used to check the stationarity of Markov chains (Elton (1990), Diaconis and Freedman (1999), Jarner and Tweedie (2001), Wu and Shao (2004)) typically imply a GMC, under some mild assumptions. In contrast, the mixing conditions for probabilistic dependence measures are, in general, not easy to verify, because the calculation involves taking the supremum over two sigma algebras. This creates overwhelming difficulties if the process is high dimensional. In addition, many well-known processes are not strong mixing. For example, Andrews (1984) showed that a simple autoregressive process with innovations as i.i.d. Bernoulli shifts is not strong mixing. In view of these features, we employ the GMC rather than the mixing condition as an underlying assumption for the dependent process.

In the framework of functional dependence measures, a commonly used condition, weaker than the GMC, assumes polynomially decaying dependence mea-

tures; that is

$$\sum_{i=m}^{\infty} \delta_{i,q} = O(m^{-\alpha}) \text{ for some } \alpha > 1 \text{ and } q \geq 1.$$

This was adopted in Chen, Xu and Wu (2013), Wu and Wu (2016), and Zhang and Wu (2017), among others. However, we can show that an exponential-type probability inequality does not, in general, hold with polynomially decaying dependence measures, even if the process is uniformly bounded. For example, consider the moving average process $e_i = \sum_{j=0}^{\infty} a_j \varepsilon_{i-j}$, where $a_j = O(j^{-\alpha})$, $j \geq 1$, for some $\alpha > 1$, and ε_i are i.i.d. symmetric, with tail probability

$$\mathbb{P}(\varepsilon_i \geq x) = x^{-q}(\log x)^{-2}, \text{ for } x \geq x_0, q > 2. \quad (2.9)$$

Let $X_i(t) = \mathbf{1}\{e_i \leq t\}$ and $S_n(t) = \sum_{i=1}^n X_i(t)$. From Theorem 14 in Chen and Wu (2018), we have the following precise order for the tail probability: for $\sqrt{n} \log n \leq x \leq n/\log n$,

$$\mathbb{P}(S_n(t) \geq x) = \frac{C(1 + o(1))n}{x^{q\alpha}(\log x)^2}, \quad (2.10)$$

where C is a constant that depends on t , q , and α . Compared with the exponential bound in (2.6), the algebraic decay in (2.10) is much larger.

3. Robust Estimation of Mean Vectors

Starting from this section, we consider high-dimensional stationary process $\mathbf{x}_i \in \mathbb{R}^p$, generated from (1.3). For high-dimensional time series, it is challenging to depict the dependence structure, because both the temporal and the cross-sectional dependence need be considered. A main advantage of the representation in (1.3) is that it lets us define physically meaningful and easily workable dependent measures, even for high-dimensional cases. Similarly, as in (2.3), we define the functional dependence measure for each component process $(\mathbf{x}_{\cdot j})$, for $1 \leq j \leq p$, as follows: If $\|\mathbf{x}_{ij}\|_q < \infty$, for some $q \geq 1$, define

$$\delta_{i,q,j} = \|\mathbf{x}_{ij} - \mathbf{x}_{ij,\{0\}}\|_q = \|h_j(\varepsilon_i, \varepsilon_{i-1}, \dots) - h_j(\varepsilon_i, \dots, \varepsilon_1, \varepsilon'_0, \varepsilon_{-1}, \dots)\|_q,$$

which measures the temporal dependence at lag i . Because each component \mathbf{x}_{ij} is dependent on the p -variate vectors $\mathbf{x}_{i-1}, \mathbf{x}_{i-2}, \dots$, $\delta_{i,q,j}$ incorporates the cross-sectional dependence as well.

Assume that the GMC is satisfied for each component process. For each j , there exists a constant $\rho_j \in (0, 1)$, such that

$$\|\mathbf{x}_{\cdot j}\|_q := \sup_{m \geq 0} \rho_j^{-m} \sum_{i=m}^{\infty} \delta_{i,q,j} < \infty. \tag{3.1}$$

To account for high dimensionality, let

$$\|\mathbf{x}_{\cdot}\|_q := \max_{1 \leq j \leq p} \|\mathbf{x}_{\cdot j}\|_q$$

be the uniform DAM, which may increase with p . Let $\rho := \max_{1 \leq j \leq p} \rho_j$ be the uniform dependence parameter. In the following example, we bound the dependence measure and the uniform DAM of a high-dimensional time series. This is a key step in applying our theorems.

Example 1 (High-dimensional Linear Models). Let ε_{ij} , for $i, j \in \mathbb{Z}$, be i.i.d. random variables with mean zero and $\|\varepsilon_{ij}\|_q < \infty$ for some $q \geq 2$. Write $\varepsilon_i = (\varepsilon_{i1}, \dots, \varepsilon_{ip})^\top$, and define the p -dimensional linear process

$$\mathbf{x}_i = \sum_{k=0}^{\infty} A_k \varepsilon_{i-k}, \tag{3.2}$$

where A_k , for $k \in \mathbb{N}$, are $p \times p$ real coefficient matrices, such that $\sum_{k=0}^{\infty} \text{tr}(A_k A_k^\top) < \infty$. Then, by Kolmogorov’s three-series theorem, the process in (3.2) is well defined. Denote the k th row of A_j by $A_j(k, \cdot)$. By Rosenthal’s inequality (Rosenthal (1970)), we have

$$\delta_{i,q,j} = \|A_i(j, \cdot) \varepsilon_0\|_q \leq (q - 1)^{1/2} |A_i(j, \cdot)|_2 \|\varepsilon_{00}\|_q.$$

If there exist $\rho_j \in (0, 1)$ and $K_p > 0$, which may depend on p , such that $|A_i(j, \cdot)|_2 \leq K_p \rho_j^i$, for all $i \geq 0$ and $1 \leq j \leq p$, with $\rho = \max_{1 \leq j \leq p} \rho_j$, we then have

$$\|\mathbf{x}_{\cdot}\|_q = \max_{1 \leq j \leq p} \sup_{m \geq 0} \rho_j^{-m} \sum_{i=m}^{\infty} \delta_{i,q,j} \leq \frac{K_p (q - 1)^{1/2} \|\varepsilon_{00}\|_q}{1 - \rho}. \tag{3.3}$$

Before proceeding, we state the main assumptions required in studying the properties of the Huber mean estimator $\hat{\mu}^H = (\hat{\mu}_1^H, \dots, \hat{\mu}_p^H)^\top$, where $\hat{\mu}_j^H$ is the solution to the equation (1.2) with robustification parameter $\kappa > 0$.

Assumption 1. Assume $n \geq 4 \vee (\log(\rho^{-1})/2)$ and $p \geq 3$.

Assumption 2.

- (a) Assume $\sigma_2 := \max_{1 \leq j \leq p} \sqrt{\text{Var}(\mathbf{x}_{ij})} < \infty$ and $\|\mathbf{x}\cdot\|_2 < \infty$, for some $\rho \in (0, 1)$.
- (b) Assume $\sigma_{1+\epsilon} = \max_{1 \leq j \leq p} (\mathbb{E}|\mathbf{x}_{ij} - \mu_j|^{1+\epsilon})^{1/(1+\epsilon)} < \infty$, for some $\epsilon \in (0, 1)$,

$$\|\mathbf{x}\cdot^*\|_{1+\epsilon} := \sup_{m \geq 0} \rho^{-m} \sum_{i=m}^{\infty} \delta_{i,1+\epsilon,j}^{(1+\epsilon)/2} < \infty,$$

for some $\rho \in (0, 1)$. Denote $\|\mathbf{x}\cdot^*\|_{1+\epsilon} = \max_{1 \leq j \leq p} \|\mathbf{x}\cdot^*\|_{1+\epsilon}$.

Assumption 1 is a very mild condition on the sample size n and the dimension p , which is noninformative and used purely for technical reasons. Assumption 2 imposes moment and dependence conditions on the underlying process. Theorem 2 is established under Assumption 2 (a), with a finite variance and a second DAM for each component process. Theorem 3 adheres to Assumption 2 (b), relaxing the order to just $(1 + \epsilon)$, for some $0 < \epsilon < 1$.

In the rest of the paper, let C_1 and C_2 be the constants as in Theorem 1. Define $\mathcal{C}_1 = 4\sqrt{\sqrt{C_1} + C_2}$, $\mathcal{C}_2 = \sqrt{C_1}$, and $\mathcal{C}_3 = C_1/2 + \sqrt{C_1} + C_2$.

Theorem 2. *Let Assumptions 1 and 2 (a) be satisfied. Let $\hat{\mu}_j^H$ be the Huber estimator of $\mu_j = \mathbb{E}\mathbf{x}_{ij}$, with the robustification parameter*

$$\kappa = \frac{8\sigma^*}{\mathcal{C}_1} \cdot \sqrt{\frac{n}{(\log n)^2 \log(1/x)}}, \tag{3.4}$$

for $\sigma^* \geq \sigma_2$, where $0 < x \leq 1/e$ satisfies

$$\mathcal{C}_1 \log\left(\frac{1}{x}\right) \left[\mathcal{C}_1 (\log n)^2 + \frac{\mathcal{C}_2 (\log n) \|\mathbf{x}\cdot\|_2}{\sigma_2} \right] \leq 4n. \tag{3.5}$$

Then, for $1 \leq j \leq p$, we have

$$\mathbb{P} \left(|\hat{\mu}_j^H - \mu_j| \geq \frac{(\mathcal{C}_1 \sigma^* \log n + \mathcal{C}_2 \|\mathbf{x}\cdot\|_2) \sqrt{\log(1/x)}}{\sqrt{n}} \right) \leq 2e^{-1/4} x. \tag{3.6}$$

In particular, letting $x = p^{-\tau-1}$, for some $\tau > 0$, if (3.5) is satisfied,

$$\mathbb{P} \left(|\hat{\mu}^H - \mu|_\infty \geq \sqrt{\tau + 1} (\mathcal{C}_1 \sigma^* \log n + \mathcal{C}_2 \|\mathbf{x}\cdot\|_2) \sqrt{\frac{\log p}{n}} \right) \leq 2e^{-1/4} p^{-\tau}. \tag{3.7}$$

Remark 2. Theorem 2 indicates that the Huber estimator admits a nearly sub-Gaussian deviation bound with second moments. In particular, by (3.6), the constructed robust mean estimator $\hat{\mu}_j^H$ deviates from the true mean μ_j logarithmically.

mically in $1/x$. However, we cannot expect such behavior by the sample mean under the same moment condition, even for the special case of i.i.d. random variables. In particular, consider i.i.d. symmetric random variables \mathbf{x}_{ij} , with the same tail probability as in (2.9). Then, we have $\mu_j = 0$, for all j . Let $\Phi(\cdot)$ be the cdf of a standard normal distribution. If we consider the empirical mean $\hat{\mu}_j = n^{-1} \sum_{i=1}^n \mathbf{x}_{ij}$, by Theorem 1.9 of Nagaev (1979), for $x \geq n^{-1/2}$,

$$\mathbb{P}(\hat{\mu}_j - \mu_j \geq x) = (1 + o(1)) \left(1 - \Phi(\sqrt{nx}) + \frac{1}{n^{q-1}(\log nx)^2 x^q} \right),$$

indicating that it may deviate from the true mean polynomially in $1/x$.

Remark 3. When it applies to the high-dimensional case, the result in (3.7) provides the rate of element-wise maximum norm convergence for the mean estimator $\hat{\mu}^H$. If σ_2 and $\|\mathbf{x}\|_2$ are both of a constant order, it follows that

$$|\hat{\mu}^H - \mu|_\infty = O_{\mathbb{P}} \left(\log n \sqrt{\frac{\log p}{n}} \right),$$

under the scaling condition $\log n \sqrt{\log p/n} \rightarrow 0$. As a natural requirement for consistency, $\log p = o(n/(\log n)^2)$ can be allowed for the dimension p . Theorem 5 in Fan, Li and Wang (2017) addresses the i.i.d. case under Assumption 2 (a), and shows that $|\hat{\mu}^H - \mu|_\infty = O_{\mathbb{P}}(\sqrt{\log p/n})$. By comparison, there is an additional multiplicative $\log n$ term in the convergence rate for the dependent case. This term is induced by the additional order $(\log n)^2$ in the Bernstein-type inequality (cf.-Theorem 1), the main tool used in the proof of Theorem 2.

Remark 4. Letting $x = p^{-\tau-1}$, condition (3.5) quantifies the relationship between n and p required in the high-dimensional case. It explicitly includes the quantity $\|\mathbf{x}\|_2/\sigma_2$, the ratio of the uniform second-order DAM to the largest component-wise standard deviation. This quantity can be used to depict the strength of dependence, and may diverge as p grows. Note too that the robustification parameter κ may diverge to infinity by adapting to the sample size n and the dimension p , which departs from the findings of Huber (1964) with a fixed parameter.

In the previous discussion, we assume finite second moments: $\sigma_2 < \infty$ and $\|\mathbf{x}\|_2 < \infty$. Theorem 3 adheres to Assumption 2 (b), relaxing the moment order to $1 + \epsilon$, where $0 < \epsilon < 1$.

Theorem 3. *Let Assumptions 1 and 2 (b) be satisfied. Let $\hat{\mu}_j^H$ be the Huber estimator of $\mu_j = \mathbb{E}\mathbf{x}_{ij}$, with the robustification parameter*

$$\kappa = K_\epsilon \left(\frac{n}{(\tau + 1)\mathcal{C}_3(\log n)^2 \log p} \right)^{1/(1+\epsilon)}, \tag{3.8}$$

where $K_\epsilon \geq (2^{-\epsilon} \|\mathbf{x}^*\|_{1+\epsilon}^2 + (2 + 2^\epsilon/\epsilon)\sigma_{1+\epsilon}^{1+\epsilon})^{1/(1+\epsilon)}$, and τ is a positive constant satisfying $(\tau + 1)\mathcal{C}_3 n^{-1}(\log n)^2 \log p \leq 1/4$. Then, we have

$$\mathbb{P} \left(|\hat{\mu}^H - \mu|_\infty \geq 2K_\epsilon \left(\frac{(\tau + 1)\mathcal{C}_3(\log n)^2 \log p}{n} \right)^{\epsilon/(1+\epsilon)} \right) \leq 2e^{-1/4} p^{-\tau}. \tag{3.9}$$

Theorem 3 delivers a slower convergence rate in the regime $0 < \epsilon < 1$. A phase transition at $\epsilon = 1$ is easily observed from Theorem 2 and Theorem 3. If $\sigma_{1+\epsilon}$ and $\|\mathbf{x}^*\|_{1+\epsilon}$ are both of a constant order, for $0 < \epsilon \leq 1$, the phase transition is smooth.

4. Robust Estimation of Covariance and Precision Matrices

4.1. Huber estimation of covariance matrix and regularization

Robust estimation is applied to matrices such as the covariance matrix; see, for example, Catoni (2016), Minsker (2016), Fan, Li and Wang (2017), and Avella-Medina et al. (2018). For $\Sigma_{\mathbf{x}} = (\sigma_{\mathbf{x},jk})_{j,k=1}^p = \mathbb{E}[(\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^\top]$, the Huber type estimator is given by

$$\hat{\Sigma}_{\mathbf{x}}^H = (\hat{\sigma}_{\mathbf{x},jk}^H)_{j,k=1}^p = (\hat{\mu}_{jk}^H - \hat{\mu}_j^H \hat{\mu}_k^H)_{j,k=1}^p, \tag{4.1}$$

where $\hat{\mu}_j^H$ and $\hat{\mu}_{jk}^H$ are the Huber estimators of $\mu_j = \mathbb{E}\mathbf{x}_{ij}$ and $\mu_{jk} = \mathbb{E}(\mathbf{x}_{ij}\mathbf{x}_{ik})$, respectively; that is, they are the solution to the equations

$$\sum_{i=1}^n \varphi_{\kappa_1}(\mathbf{x}_{ij} - \theta) = 0 \quad \text{and} \quad \sum_{i=1}^n \varphi_{\kappa_2}(\mathbf{x}_{ij}\mathbf{x}_{ik} - \theta) = 0,$$

respectively, with robustification parameters $\kappa_1, \kappa_2 > 0$. The convergence results under the element-wise maximum norm are established in Corollary 1 and Corollary 2, which rely on the following additional assumption.

Assumption 3.

- (a) Assume $\omega_4 := \max_{1 \leq j \leq p} \|\mathbf{x}_{ij}\|_4 < \infty$ and $\|\mathbf{x}\|_4 < \infty$, for some $\rho \in (0, 1)$.
- (b) Assume $\omega_{2+2\epsilon} := \max_{1 \leq j \leq p} \|\mathbf{x}_{ij}\|_{2+2\epsilon} < \infty$, for some $\epsilon \in (0, 1)$, and $\|\mathbf{x}\|_{2+2\epsilon}$

$< \infty$, for some $\rho \in (0, 1)$.

We claim that Assumption 3 and Assumption 2 correspond to the same ρ , although they adhere to moments of different orders. This holds owing to an interesting property of the GMC: If $\|X_i\|_{q^*} < \infty$ for some $q^* > 0$, and $\text{GMC}(q_0)$ holds for the process (X_i) , for some $0 < q_0 \leq q^*$ and $\rho \in (0, 1)$, then $\text{GMC}(q)$ holds with the same ρ , for all $q \in (0, q^*]$. The above property of the GMC follows from Lemma 2 in Wu and Min (2005).

Corollary 1. *Let Assumptions 1 and 3 (a) be satisfied. Denote $\mu^o = \max_{1 \leq j \leq p} |\mu_j|$. Let $\hat{\Sigma}_{\mathbf{x}}^H = (\hat{\mu}_{jk}^H - \hat{\mu}_j^H \hat{\mu}_k^H)_{j,k=1}^p$, where $\hat{\mu}_j^H$ and $\hat{\mu}_{jk}^H$ are the Huber estimators of μ_j and μ_{jk} , respectively, with robustification parameters chosen as*

$$\kappa_1 = \frac{8\sigma^*}{C_1\sqrt{\tau+2}} \cdot \sqrt{\frac{n}{(\log n)^2 \log p}}, \quad \kappa_2 = \frac{8\omega^*}{C_1\sqrt{\tau+2}} \cdot \sqrt{\frac{n}{(\log n)^2 \log p}},$$

respectively, for $\sigma^* \geq \sigma_2$, $\omega^* \geq \omega_4^2$, and τ a positive constant satisfying

$$(\tau + 2)C_1 \log p \left(C_1(\log n)^2 + C_2 \log n \max \left\{ \frac{\|\mathbf{x}\cdot\|_2}{\sigma_2}, \frac{\|\mathbf{x}\cdot\|_4}{\omega_4} \right\} \right) \leq 4n. \quad (4.2)$$

Then, we have

$$\mathbb{P}(|\hat{\Sigma}_{\mathbf{x}}^H - \Sigma_{\mathbf{x}}|_{\infty} \geq \Delta_{n,p}) \leq \frac{8e^{-1/4}}{3} p^{-\tau}, \quad (4.3)$$

with

$$\begin{aligned} \Delta_{n,p} = & \sqrt{\tau+2} [C_1(2\mu^o\sigma^* + \omega^*) \log n + C_2(2\mu^o\|\mathbf{x}\cdot\|_2 + \omega_4\|\mathbf{x}\cdot\|_4)] \sqrt{\frac{\log p}{n}} \\ & + (\tau+2)(C_1\sigma^* \log n + C_2\|\mathbf{x}\cdot\|_2)^2 \cdot \frac{\log p}{n}. \end{aligned}$$

Remark 5. From Corollary 1, we have

$$|\hat{\Sigma}_{\mathbf{x}}^H - \Sigma_{\mathbf{x}}|_{\infty} = O_{\mathbb{P}} \left((\mu^o\sigma_2 + \omega_4^2) \log n \sqrt{\frac{\log p}{n}} + (\mu^o\|\mathbf{x}\cdot\|_2 + \omega_4\|\mathbf{x}\cdot\|_4) \sqrt{\frac{\log p}{n}} \right),$$

under the scaling condition $\log n \sqrt{\log p/n} \rightarrow 0$ and condition (4.2). The convergence rate is the sum of two terms, where the former incorporates the moments σ_2 and ω_4 , and the order $\log n \sqrt{\log p/n}$. Compared with Proposition 3 of Avella-Medina et al. (2018), which addresses the Huber-type estimator of $\Sigma_{\mathbf{x}}$ for i.i.d. vectors, we include an additional $\log n$ factor when it is generalized to the time series setting. This factor is characterized by DAMs $\|\mathbf{x}\cdot\|_2$ and $\|\mathbf{x}\cdot\|_4$, which would not arise in the i.i.d. case. Note that $\|\mathbf{x}\cdot\|_2/\sigma_2$ and $\|\mathbf{x}\cdot\|_4/\omega_4$ may

diverge as p grows, especially when the strength of dependence is strong. Hence, we cannot tell which term dominates in general without extra information.

Among the extensive literature on covariance matrix estimation in the high-dimensional case, the sample covariance matrix $\hat{\Sigma}_{\mathbf{x}}$ defined in (1.1) is widely used as a pilot estimator for $\Sigma_{\mathbf{x}}$. Various regularized (banded, tapered, thresholded) estimators can then be constructed based on $\hat{\Sigma}_{\mathbf{x}}$, after imposing some structural assumptions on the true covariance matrix. See, for example, Bickel and Levina (2008a,b), El Karoui (2008), Cai, Zhang and Zhou (2010), Cai and Liu (2011), and Cai and Zhou (2012) for independent data, and Wu and Pourahmadi (2009), McMurry and Politis (2010), and Chen, Xu and Wu (2013) for temporally dependent data. In these theoretical investigations, either subGaussianity of the data is assumed, or the derived deviation bound increases with p polynomially.

Our result represents a significant improvement by relaxing the subGaussian assumption to the existence of fourth moments, while retaining a nearly subGaussian deviation bound. Regularized estimators based on $\hat{\Sigma}^H$ can exhibit such nice performance. We illustrate this by discussing the property of a robust modification of the thresholded estimator

$$T_u(\hat{\Sigma}_{\mathbf{x}}^H) = (\hat{\sigma}_{\mathbf{x},jk}^H \mathbf{1}\{\hat{\sigma}_{\mathbf{x},jk}^H \geq u\})_{1 \leq j,k \leq p},$$

where $\hat{\Sigma}_{\mathbf{x}}^H$ is the Huber estimator of the covariance matrix defined in (4.1). Here, we consider the following uniform class of sparse matrices:

$$\mathcal{U}_r(M_1, s_0(p)) = \left\{ \Sigma = (\sigma_{jk})_{j,k=1}^p : \max_j \sigma_{jj} \leq M_1, \max_j \sum_{k=1}^p |\sigma_{jk}|^r \leq s_0(p) \right\},$$

for some $0 \leq r < 1$. The above class, defined in terms of a strong ℓ^r -ball, was also considered by Bickel and Levina (2008a), Rothman, Levina and Zhu (2009), Cai, Liu and Luo (2011), Cai and Zhou (2012), and Chen, Xu and Wu (2013). Imposing such a structural assumption on the true covariance matrix, we obtain a convergence result under the spectral norm, in a similar way to the proof of Theorem 1 in Bickel and Levina (2008a). Assume $\Sigma_{\mathbf{x}}$ belongs to $\mathcal{U}_r(M_1, s_0(p))$, and that μ^o , σ_2 , ω_4 , $\|\mathbf{x}\|_2$, and $\|\mathbf{x}\|_4$ are of a constant order. Then, we have that if $\log n \sqrt{\log p/n} = o(1)$,

$$\rho(T_u(\hat{\Sigma}_{\mathbf{x}}^H) - \Sigma_{\mathbf{x}}) = O_{\mathbb{P}} \left(s_0(p) \left(\frac{(\log n)^2 \log p}{n} \right)^{(1-r)/2} \right), \quad (4.4)$$

where $u = C \log n \sqrt{\log p/n}$ and C is a sufficiently large constant. We now compare our result with those of existing work on thresholded covariance estimation for the uniform class $\mathcal{U}_r(M_1, s_0(p))$. First, for vectors $\mathbf{x}_1, \dots, \mathbf{x}_n$ that are i.i.d. and Gaussian, Bickel and Levina (2008a) derive the following thresholded estimator, based on the sample covariance matrix:

$$\rho(T_u(\hat{\Sigma}_{\mathbf{x}}) - \Sigma_{\mathbf{x}}) = O_{\mathbb{P}} \left(s_0(p) \left(\frac{\log p}{n} \right)^{(1-r)/2} \right), \tag{4.5}$$

for $u = C \sqrt{\log p/n}$. The same rate as that of (4.5) is achievable for some variants; see, for example, Rothman, Levina and Zhu (2009) for generalized thresholding, and Cai and Liu (2011) for adaptive thresholding. The latter work shows that the rate is minimax optimal. Our result concerns high-dimensional time series and relaxes the Gaussian/subGaussian assumption to the existence of fourth moments, at the cost of a logarithmic factor in the convergence rate.

We next compare our result with those of works that assume finite polynomial moments. In Section 2.3 of Bickel and Levina (2008a), when $\|\mathbf{x}_{ij}\|_q$ is bounded for some $q \geq 4$, by taking $u = Cp^{4/q}n^{-1/2}$, they obtained

$$\rho(T_u(\hat{\Sigma}_{\mathbf{x}}) - \Sigma_{\mathbf{x}}) = O_{\mathbb{P}} \left(s_0(p) \left(\frac{p^{8/q}}{n} \right)^{(1-r)/2} \right).$$

To ensure consistency, $p = o(n^{8/q})$ is required. Cai and Liu (2011) assumed $q > 4 + \epsilon$, and suggested $p \leq Cn^{\epsilon/4}$ in Theorem 1 (ii). Chen, Xu and Wu (2013) quantified the convergence rate for high-dimensional time series when $q > 4$. By Theorem 2.3 and Corollary 2.7 therein, p can still only be allowed to increase polynomially with n . In contrast, we can allow $\log p = o(n/(\log n)^2)$, and require only $q = 4$. We can further relax the moment condition by imposing finite $(2 + 2\epsilon)$ th moments, for some $0 < \epsilon < 1$; see Corollary 2.

Corollary 2. *Let Assumptions 1, 2 (b), and 3 (b) be satisfied. Denote $\mu^o = \max_{1 \leq j \leq p} |\mu_j|$. Let $\hat{\Sigma}_{\mathbf{x}}^H = (\hat{\mu}_{jk}^H - \hat{\mu}_j^H \hat{\mu}_k^H)_{j,k=1}^p$, where $\hat{\mu}_j^H$ and $\hat{\mu}_{jk}^H$ are the Huber estimators of μ_j and μ_{jk} , respectively, with robustification parameters chosen as*

$$\begin{aligned} \kappa_1 &= K_{\epsilon} \left(\frac{n}{C_3(\tau + 2)(\log n)^2 \log p} \right)^{1/(1+\epsilon)}, \\ \kappa_2 &= K'_{\epsilon} \left(\frac{n}{C_3(\tau + 2)(\log n)^2 \log p} \right)^{1/(1+\epsilon)}, \end{aligned}$$

respectively, for $K_\epsilon \geq (2^{-\epsilon} \|\mathbf{x}^*\|_{1+\epsilon}^2 + (2 + 2^\epsilon/\epsilon) \sigma_{1+\epsilon}^{1+\epsilon})^{1/(1+\epsilon)}$,
 $K'_\epsilon \geq (2^{-\epsilon} \omega_{2+2\epsilon}^{1+\epsilon} \|\mathbf{x}^*\|_{1+\epsilon}^2 + (2 + 2^\epsilon/\epsilon) \omega_{2+2\epsilon}^{2+2\epsilon})^{1/(1+\epsilon)}$, and τ a positive constant satisfying $(\tau + 2)C_3(\log n)^2 \log p \leq n/4$. We then have

$$\mathbb{P}(|\hat{\Sigma}_{\mathbf{x}}^H - \Sigma_{\mathbf{x}}|_\infty \geq \Delta_{n,p}^*) \leq \frac{8e^{-1/4}}{3} p^{-\tau}, \quad (4.6)$$

with

$$\begin{aligned} \Delta_{n,p}^* &= (4\mu^o K_\epsilon + 2K'_\epsilon) \left(\frac{(\tau + 2)C_3(\log n)^2 \log p}{n} \right)^{\epsilon/(1+\epsilon)} \\ &\quad + 4K_\epsilon^2 \left(\frac{(\tau + 2)C_3(\log n)^2 \log p}{n} \right)^{2\epsilon/(1+\epsilon)}. \end{aligned}$$

As Corollary 2 shows, under a weaker moment condition, the convergence rate is slower, which is in line with the phase transition phenomenon of Huber mean estimators. With the element-wise maximum norm convergence result, we can establish the rates of convergence under the spectral norm for various regularized estimators based on $\hat{\Sigma}_{\mathbf{x}}^H$; the discussion is thus omitted here for brevity.

4.2. Robust estimation of precision matrices

A precision matrix is a powerful tool used to encode the relationships between a large number of random variables in graphical models. For the nonGaussian case, the matrix is associated with partial correlation graphs (e.g., Peng et al. (2009)). The problem of estimating a large precision matrix and recovering its support has drawn considerable attention in the i.i.d. case; see Meinshausen and Bühlmann (2006), Rothman et al. (2008), Lam and Fan (2009), Yuan (2010), Ravikumar et al. (2011), Cai, Liu and Luo (2011), Xue and Zou (2012), and Cai, Liu and Zhou (2016), among many others. We consider a modified procedure of the CLIME (Cai, Liu and Luo (2011)) to estimate $\Omega_{\mathbf{x}}$ within the framework given in (1.3). Let $\hat{\Sigma}_{\mathbf{x}}^H$ be the Huber-type estimator of $\Sigma_{\mathbf{x}}$. Our procedure for estimating $\Omega_{\mathbf{x}}$ consists of two steps.

Step I: Solve the optimization problem

$$\tilde{\Omega}^H = \arg \min |\Omega|_1 \quad \text{subject to} \quad |\hat{\Sigma}^H \Omega - I_p|_\infty \leq \lambda_n,$$

where $\lambda_n > 0$ is a tuning parameter.

Step II: Obtain the symmetric estimator

$$\hat{\Omega}^H = (\hat{\omega}_{jk}^H) \text{ where } \hat{\omega}_{jk}^H = \tilde{\omega}_{jk}^H \mathbf{1}\left\{|\tilde{\omega}_{jk}^H| \leq |\tilde{\omega}_{kj}^H|\right\} + \tilde{\omega}_{kj}^H \mathbf{1}\left\{|\tilde{\omega}_{jk}^H| > |\tilde{\omega}_{kj}^H|\right\}.$$

It is known that Step I is equivalent to solving the following p -vector minimization problems in parallel:

$$\tilde{\omega}_j^H = \arg \min |w|_1 \text{ subject to } |\hat{\Sigma}^H \omega - u_j|_\infty \leq \lambda_n, \text{ for } 1 \leq j \leq p,$$

where u_j is the unit vector in \mathbb{R}^p , with one in the j th coordinate, and zero otherwise. Then, we construct our estimator as $\hat{\Omega}^H = (\tilde{\omega}_1^H, \dots, \tilde{\omega}_p^H)$. We consider the uniform class of matrices

$$\mathcal{V}_r(M_2, s_0(p)) = \left\{ \Omega = (\omega_{jk})_{j,k=1}^p : \|\Omega\|_{\ell_1} \leq M_2, \max_j \sum_{k=1}^p |\sigma_{jk}|^r \leq s_0(p) \right\},$$

for some $0 \leq r < 1$. The following theorem gives the rates of convergence for the modified CLIME estimator $\hat{\Omega}^H$ under the element-wise maximum norm and the spectral norm.

Theorem 4. *Suppose $\Omega_{\mathbf{x}} \in \mathcal{V}_r(M_2, s_0(p))$. (i) Let the assumptions of Corollary 1 be satisfied, and let $\hat{\Sigma}_{\mathbf{x}}^H$ be the Huber-type estimator therein. Let $\hat{\Omega}_{\mathbf{x}}^H$ be obtained as*

$$\lambda_n = \mathcal{L}_1 M_2 \sqrt{\frac{\log p}{n}} + \mathcal{L}_2 M_2 \frac{\log p}{n},$$

for $\mathcal{L}_1 \geq \sqrt{\tau + 2} [\mathcal{C}_1(2\mu^o\sigma^* + \omega^*) \log n + \mathcal{C}_2(2\mu^o\|\mathbf{x}\|_2 + \omega_4\|\mathbf{x}\|_4)]$, and $\mathcal{L}_2 \geq (\tau + 2)(\mathcal{C}_1\sigma^* \log n + \mathcal{C}_2\|\mathbf{x}\|_2)^2$. Then, we have,

$$\mathbb{P} \left(|\hat{\Omega}_{\mathbf{x}}^H - \Omega_{\mathbf{x}}|_\infty \geq 4\|\Omega\|_{\ell_1} \lambda_n \right) \leq \frac{8e^{-1/4}}{3} p^{-\tau}, \tag{4.7}$$

$$\mathbb{P} \left(\rho(\hat{\Omega}_{\mathbf{x}}^H - \Omega_{\mathbf{x}}) \geq \mathcal{C}_4 s_0(p) (\|\Omega\|_{\ell_1} \lambda_n)^{1-r} \right) \leq \frac{8e^{-1/4}}{3} p^{-\tau}, \tag{4.8}$$

where $\mathcal{C}_4 = 2(1 + 2^{1-q} + 3^{1-q})4^{1-q}$. (ii) Let the assumptions of Corollary 2 be satisfied, and let $\hat{\Sigma}_{\mathbf{x}}^H$ be the Huber-type estimator therein. Then, (4.7) and (4.8) hold for

$$\lambda_n = \mathcal{L}_3 M_2 \left(\frac{(\tau + 2)\mathcal{C}_3(\log n)^2 \log p}{n} \right)^{\epsilon/(1+\epsilon)} + \mathcal{L}_4 M_2 \left(\frac{(\tau + 2)\mathcal{C}_3(\log n)^2 \log p}{n} \right)^{2\epsilon(1+\epsilon)},$$

where $\mathcal{L}_3 \geq 4\mu^\circ K_\epsilon + 2K'_\epsilon$ and $\mathcal{L}_4 \geq 4K_\epsilon^2$.

If $M_2 = O(1)$ and $\log n \sqrt{\log p/n} = o(1)$, and if μ° , σ_2 , ω_4 , $\|\mathbf{x}\|_2$, and $\|\mathbf{x}\|_4$ are of a constant order, we have

$$|\hat{\Omega}^H - \Omega_{\mathbf{x}}|_\infty = O\left(\log n \sqrt{\frac{\log p}{n}}\right), \quad (4.9)$$

with probability greater than $1 - O(p^{-\tau})$, for some $\tau > 0$.

We now compare our result with earlier results under the assumption of finite q th moments. For convenience of notation, we denote all other estimators by $\hat{\Omega}_{\mathbf{x}}$. For i.i.d. p -variate vectors, Ravikumar et al. (2011) studied the graphical Lasso estimator, with off-diagonal entries penalized by the ℓ_1 -norm, and Cai, Liu and Luo (2011) investigated the CLIME constructed using the sample covariance matrix $\hat{\Sigma}_{\mathbf{x}}$. Corollary 2 of the former paper showed that for some $q \geq 4$ and $\tau > 2$, if $p = O([n/s_0(p)]^{q/(4\tau)})$, $|\hat{\Omega}_{\mathbf{x}} - \Omega_{\mathbf{x}}|_\infty = O(p^{2\tau/q}/n^{1/2})$ with probability greater than $1 - O(p^{2-\tau})$. Furthermore, their Theorem 4 shows that for $p = O(n^\gamma)$ and $q = 4 + 4\gamma + \delta$, where $\gamma > 0$ and $\delta > 0$, with probability greater than $1 - O(n^{-\delta} + p^{-\tau/2})$, $|\hat{\Omega}_{\mathbf{x}} - \Omega_{\mathbf{x}}|_\infty = O(\sqrt{\log p/n})$.

The nice property of our estimator comprises three aspects: (i) We can relax the moment condition by allowing $2 < q \leq 4$. (ii) A nearly subGaussian deviation bound is attained. (iii) The dimension p can be allowed to increase exponentially with n , and the range $\log p = o(n/(\log n)^2)$ is much wider than when allowing for a polynomial increase only.

5. Numerical Results

5.1. Simulation study

We conduct a simulation study to compare the empirical performance of the Huber mean estimator with that of the sample mean. We consider the following linear process with fat-tailed errors: let ε_{ij} , for $i, j \in \mathbb{Z}$, be i.i.d. random variables distributed as $t(d)/\sqrt{d/(d-2)}$, where $t(d)$ is Student's t -distribution with degrees of freedom $d = 2.5$; let $\varepsilon_i = (\varepsilon_{i1}, \dots, \varepsilon_{ip})^\top$ and

$$X_i = \sum_{k=0}^{\infty} A_k \varepsilon_{i-k}. \quad (5.1)$$

Here, the coefficient matrix $A_k = \rho^k M$, where $M = (m_{ij})_{i,j=1}^p$ is a Toeplitz matrix with $m_{ij} = \rho^{|i-j|+1}$. The parameter $\rho \in (0, 1)$ controls the decay rate

Table 1. Uniform deviations of Huber mean estimate and sample mean.

(n, p)	$\rho = 0.2$		$\rho = 0.5$		$\rho = 0.8$	
	$ \hat{\mu}^H - \mu _\infty$	$ \hat{\mu} - \mu _\infty$	$ \hat{\mu}^H - \mu _\infty$	$ \hat{\mu} - \mu _\infty$	$ \hat{\mu}^H - \mu _\infty$	$ \hat{\mu} - \mu _\infty$
(20, 50)	0.019 (0.004)	0.035 (0.025)	0.257 (0.060)	0.369 (0.234)	2.362 (0.843)	2.588 (1.319)
(20, 100)	0.021 (0.004)	0.047 (0.061)	0.293 (0.059)	0.456 (0.320)	2.771 (0.823)	3.015 (1.504)
(20, 200)	0.023 (0.004)	0.056 (0.057)	0.324 (0.059)	0.574 (0.415)	0.933 (0.933)	3.847 (1.926)
(50, 50)	0.012 (0.002)	0.022 (0.013)	0.161 (0.036)	0.234 (0.140)	1.477 (0.439)	1.658 (0.678)
(50, 100)	0.013 (0.002)	0.028 (0.024)	0.183 (0.034)	0.273 (0.135)	1.785 (0.434)	2.147 (1.294)
(50, 200)	0.014 (0.002)	0.033 (0.024)	0.199 (0.033)	0.358 (0.269)	2.071 (0.468)	2.635 (1.691)
(100, 50)	0.008 (0.001)	0.015 (0.008)	0.115 (0.025)	0.159 (0.071)	1.084 (0.323)	1.273 (0.690)
(100, 100)	0.009 (0.001)	0.018 (0.012)	0.127 (0.022)	0.210 (0.201)	1.277 (0.320)	1.500 (0.833)
(100, 200)	0.010 (0.001)	0.024 (0.035)	0.140 (0.022)	0.235 (0.140)	1.483 (0.310)	1.846 (0.908)

of the functional dependence measures for the generated process. We consider the following numerical setups: ρ is set to 0.2, 0.5, 0.8; $n = 20, 50, 100$; and $p = 50, 100, 200$. In each simulation, we truncate the sum in the linear process in (5.1) to $\sum_{k=0}^{2000}$. For each case, we report the average (as an entry) and standard deviation (in parentheses) of the uniform deviation $|\hat{\mu}^H - \mu|_\infty$, and of $|\hat{\mu} - \mu|_\infty$ based on 1,000 repetitions.

In particular, to obtain the Huber mean estimate μ_j^H for the j th component process, $1 \leq j \leq p$, we solve (1.2) in parallel with the robustification parameter κ_j . Motivated by Bickel (1975) and the theoretical suggestion in (3.4) on the choice of κ_j , in practice, we take

$$\kappa_j = \hat{\sigma}_j \cdot \sqrt{\frac{n}{(\log n)^2 \log p}}, \tag{5.2}$$

where $\hat{\sigma}_j = \text{median}\{|\mathbf{x}_{ij} - m_j|\} / \Phi^{-1}(3/4)$ is the symmetrized interquartile range for the j th component process, for $1 \leq j \leq p$; m_j is the sample median of $\mathbf{x}_{.j}$, for $1 \leq i \leq n$; and Φ is the standard normal cdf.

Table 1 shows that, as expected from our theoretical results, the Huber mean

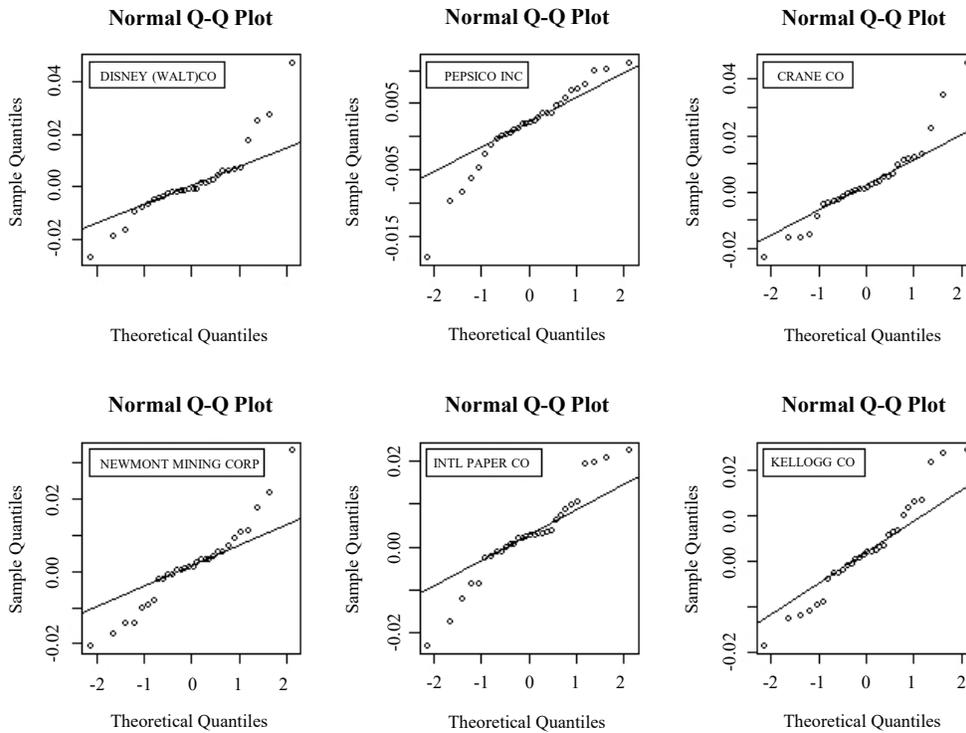


Figure 1. Normal QQplots of daily stock returns for six firms.

estimator outperforms the sample mean, with a smaller uniform deviation and a smaller standard deviation in all cases. Moreover, the uniform deviation is larger when the dependence is stronger, the sample size is smaller, and the dimension is higher, supporting the result in (3.7). A similar claim can be made in the case of Theorem 3; the results are not reported here.

5.2. Real-data analysis

In this section, we compare the aforementioned estimators using a real data set, taken from the CRSP. The data matrix contains daily returns of S&P 500 stocks between December 8, 2009, and December 29, 2017. We chose this period to avoid the effects of the financial crises in 2001 and 2008, which could make the time series of stock returns nonstationary. Stocks with missing price data are excluded, and prices on weekends are not observed. In total, there are 2,030 time points and 94 stocks in the data set. We estimate the true means of the daily returns for the 94 stocks using the empirical means based on the first 2,000 time

points, denoted by $\mu_0 \in \mathbb{R}^{94}$. Then, we use the data matrix $\mathbf{x} \in \mathbb{R}^{30 \times 94}$ of the remaining 30 time points to compare the Huber mean estimate $\hat{\mu}^H$ and the sample mean $\hat{\mu}$. In computing each $\hat{\mu}_j^H$ by solving (1.2), the robustification parameter κ_j is chosen as in (5.2). Figure 1 presents the normal qqplots of the daily stock returns for six firms selected from \mathbf{x} , all indicating heavier tails than that of the normal distribution. We have the following output: $|\hat{\mu}^H - \mu_0|_\infty = 0.0098$ and $|\hat{\mu} - \mu_0|_\infty = 0.0141$. If we apply the Huber mean estimate μ_0^H on the first 2,000 time points in place of μ_0 , we have similar results: $|\hat{\mu}^H - \mu_0^H|_\infty = 0.0106$ and $|\hat{\mu} - \mu_0^H|_\infty = 0.0143$. In both comparisons, the Huber mean estimator shows better performance in terms of accuracy than that of the sample mean for heavy-tailed data.

We further estimate the covariance matrix for the daily returns of the 94 stocks. Similarly, we estimate the true covariance matrix using the sample covariance matrix based on the first 2,000 time points, which we denote by $\Sigma_0 \in \mathbb{R}^{94 \times 94}$. We then work on the data matrix $\mathbf{x} \in \mathbb{R}^{30 \times 94}$ of the final 30 time points to obtain the sample covariance matrix $\hat{\Sigma}_{\mathbf{x}}$ given in (1.1) and the Huber type estimator $\hat{\Sigma}_{\mathbf{x}}^H$ given in (4.1). The output $|\hat{\Sigma}_{\mathbf{x}}^H - \Sigma_0|_\infty = 0.0016$ and $|\hat{\Sigma}_{\mathbf{x}} - \Sigma_0|_\infty = 0.0027$ shows that the Huber estimator also performs better in terms of covariance matrix estimation. Comparisons of various regularized estimators starting from the two pilot estimators $\hat{\Sigma}_{\mathbf{x}}, \hat{\Sigma}_{\mathbf{x}}^H$ can be further conducted, and are not discussed here.

Supplementary Material

The online Supplementary Material provides all technical proofs.

Acknowledgments

We thank the two anonymous referees, Associate Editor, and Editor for their helpful comments.

References

- Adamczak, R. (2008). A tail inequality for suprema of unbounded empirical processes with applications to markov chains. *Electronic Journal of Probability* **13**, 1000–1034.
- Andrews, D. (1984). Non-strong mixing autoregressive processes. *Journal of Applied Probability* **21**, 930–934.
- Avella-Medina, M., Battey, H., Fan, J., and Li, Q. (2018). Robust estimation of high-dimensional covariance and precision matrices. *Biometrika* **105**, 271–284.
- Basu, S. and Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics* **43**, 1535–1567.

- Bernstein, S. (1946). *The Theory of Probabilities*. Gostehizdat Publishing House, Moscow.
- Bickel, P. (1975). One-step huber estimates in the linear model. *Journal of the American Statistical Association* **70**, 428–434.
- Bickel, P. and Levina, E. (2008a). Covariance regularization by thresholding. *The Annals of Statistics* **36**, 2577–2604.
- Bickel, P. and Levina, E. (2008b). Regularized estimation of large covariance matrices. *The Annals of Statistics* **36**, 199–227.
- Bühlmann, P. and Van De Geer, S. (2011). *Statistics for High-dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media, Berlin.
- Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association* **106**, 672–684.
- Cai, T., Liu, W. and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* **106**, 594–607.
- Cai, T., Liu, W. and Zhou, H. (2016). Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *The Annals of Statistics* **44**, 455–488.
- Cai, T., Zhang, C. and Zhou, H. (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics* **38**, 2118–2144.
- Cai, T. and Zhou, H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics* **40**, 2389–2420.
- Catoni, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* **48**, 1148–1185.
- Catoni, O. (2016). PAC-bayesian bounds for the gram matrix and least squares regression with a random design. *arXiv preprint arXiv:1603.05229* .
- Chen, L. and Wu, W. (2018). Concentration inequalities for empirical processes of linear time series. *Journal of Machine Learning Research* **18**, 1–46.
- Chen, X., Xu, M. and Wu, W. (2013). Covariance and precision matrix estimation for high-dimensional time series. *The Annals of Statistics* **41**, 2994–3021.
- Diaconis, P. and Freedman, D. (1999). Iterated random functions. *SIAM Review* **41**, 45–76.
- Douc, R., Guillin, A. and Moulines, E. (2008). Bounds on regeneration times and limit theorems for subgeometric markov chains. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* **44**, 239–257.
- El Karoui, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *The Annals of Statistics* **36**, 2717–2756.
- Elton, J. (1990). A multiplicative ergodic theorem for lipschitz maps. *Stochastic Processes and their Applications* **34**, 39–47.
- Fan, J., Li, Q. and Wang, Y. (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**, 247–265.
- Friston, K. (2011). Functional and effective connectivity: A review. *Brain Connectivity* **1**, 13–36.
- Huber, P. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics* **35**, 73–101.
- Huber, P. (1973). Robust regression: asymptotics, conjectures and monte carlo. *The Annals of Statistics* **1**, 799–821.
- Jarner, S. and Tweedie, R. (2001). Locally contracting iterated functions and stability of markov

- chains. *Journal of Applied Probability* **38**, 494–507.
- Kim, Y., Giacometti, R., Rachev, S., Fabozzi, F. and Mignacca, D. (2012). Measuring financial risk and portfolio optimization with a non-gaussian multivariate model. *Annals of Operations Research* **201**, 325–343.
- Kondrashov, D., Kravtsov, S., Robertson, A. and Ghil, M. (2005). A hierarchy of data-based enso models. *Journal of Climate* **18**, 4425–4444.
- Koopman, S. and Lucas, A. (2008). A non-gaussian panel time series model for estimating and decomposing default risk. *Journal of Business & Economic Statistics* **26**, 510–525.
- Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of Statistics* **37**, 4254.
- McMurry, T. and Politis, D. (2010). Banded and tapered estimates for autocovariance matrices and the linear process bootstrap. *Journal of Time Series Analysis* **31**, 471–482.
- McMurry, T. L. and Politis, D. N. (2015). High-dimensional autocovariance matrices and optimal linear prediction. *Electronic Journal of Statistics* **9**, 753–788.
- Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics* **34**, 1436–1462.
- Merlevède, F., Peligrad, M. and Rio, E. (2009). Bernstein inequality and moderate deviations under strong mixing conditions. In *High Dimensional Probability V: The Luminy Volume*, 273–292. Institute of Mathematical Statistics.
- Merlevède, F., Peligrad, M. and Rio, E. (2011). A Bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields* **151**, 435–474.
- Minsker, S. (2016). Sub-gaussian estimators of the mean of a random matrix with heavy-tailed entries. *arXiv preprint arXiv:1605.07129*.
- Nagaev, S. (1979). Large deviations of sums of independent random variables. *The Annals of Probability* **7**, 745–789.
- Peng, J., Wang, P., Zhou, N. and Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association* **104**, 735–746.
- Priestley, M. (1988). *Nonlinear and Nonstationary Time Series Analysis*. Academic Press, London.
- Ravikumar, P., Wainwright, M. J., Raskutti, G. and Yu, B. (2011). High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence. *Electronic Journal of Statistics* **5**, 935–980.
- Rosenblatt, M. (1971). *Markov Processes: Structure and Asymptotic Behavior*. Springer, Berlin.
- Rosenthal, H. (1970). On the subspaces of L^p ($p > 2$) spanned by sequences of independent random variables. *Israel Journal of Mathematics* **8**, 273–303.
- Rothman, A., Levina, E. and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association* **104**, 177–186.
- Rothman, A. J., Bickel, P. J., Levina, E. and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* **2**, 494–515.
- Shao, X. and Wu, W. (2007). Asymptotic spectral theory for nonlinear time series. *The Annals of Statistics* **35**, 1773–1801.
- Tong, H. (1990). *Non-linear Time Series: A Dynamical System Approach*. Oxford University Press, Oxford.

- Wiener, N. (1958). *Nonlinear Problems in Random Theory*. MIT Press and Wiley.
- Wu, W. (2005). Nonlinear system theory: Another look at dependence. *Proceedings of National Academy of Sciences of the United States of America* **102**, 14150–14154.
- Wu, W. and Min, W. (2005). On linear processes with dependent innovations. *Stochastic Processes and their Applications* **115**, 939–958.
- Wu, W. and Pourahmadi, M. (2009). Banded sample autocovariance matrices of stationary processes. *Statistica Sinica* **19**, 1755–1768.
- Wu, W. and Shao, X. (2004). Limit theorems for iterated random functions. *Journal of Applied Probability* **41**, 425–436.
- Wu, W. and Wu, Y. (2016). Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electronic Journal of Statistics* **10**, 352–379.
- Xue, L. and Zou, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *The Annals of Statistics* **40**, 2541–2571.
- Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research* **11**, 2261–2286.
- Zhang, D. and Wu, W. (2017). Gaussian approximation for high dimensional time series. *The Annals of Statistics* **45**, 1895–1919.

Danna Zhang

Department of mathematics, University of California, San Diego, La Jolla, CA 92093, USA.

E-mail: daz076@ucsd.edu

(Received May 2018; accepted June 2019)