Journal of Econometrics 208 (2019) 5–22

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Robust covariance estimation for approximate factor models

Jianqing Fan [a,*], Weichen Wang [a], Yiqiao Zhong [a]

[a] Department of Operations Research and Financial Engineering, Princeton University, USA

ARTICLE INFO

Article history:
Available online 6 October 2018

JEL classification:
C13
C38

Keywords:
Robust covariance matrix
Approximate factor model
M-estimator

ABSTRACT

In this paper, we study robust covariance estimation under the approximate factor model with observed factors. We propose a novel framework to first estimate the initial joint covariance matrix of the observed data and the factors, and then use it to recover the covariance matrix of the observed data. We prove that once the initial matrix estimator is good enough to maintain the element-wise optimal rate, the whole procedure will generate an estimated covariance with desired properties. For data with bounded fourth moments, we propose to use adaptive Huber loss minimization to give the initial joint covariance estimation. This approach is applicable to a much wider class of distributions, beyond sub-Gaussian and elliptical distributions. We also present an asymptotic result for adaptive Huber's M-estimator with a diverging parameter. The conclusions are demonstrated by extensive simulations and real data analysis.

1. Introduction

The problem of estimating a covariance matrix and its inverse has been fundamental in many areas of statistics and econometrics, including principal component analysis (PCA) and undirected graphical models for instance. The intense research in high dimensional statistics has contributed a stream of papers related to covariance matrix estimation, including sparse principal component analysis (Johnstone and Lu, 2009; Amini and Wainwright, 2008; Vu and Lei, 2013; Birnbaum et al., 2013; Berthet and Rigollet, 2013; Ma, 2013; Cai et al., 2013), sparse covariance estimation (Bickel and Levina, 2008; Cai and Liu, 2011; Cai et al., 2010; Lam and Fan, 2009; Ravikumar et al., 2011) and factor model analysis (Stock and Watson, 2002; Bai, 2003; Fan et al., 2008, 2013, 2016; Onatski, 2012). A strong interest in precision matrix estimation (undirected graphical model) has also emerged in the statistics community following the pioneering works in Meinshausen and Bühlmann (2006) and Friedman et al. (2008). In the application aspect, many areas such as portfolio allocation (Fan et al., 2008), have benefited from this continuing research.

In the high dimensional setting, the number of variables p is comparable or greater than the sample size n. This dimensionality poses a challenge to the estimation of covariance matrices. It has been shown in Johnstone and Lu (2009) that the empirical covariance matrix behaves poorly, and sparsity of leading eigenvectors circumvents this issue. Following this work, a flourishing literature on sparse PCA has developed in-depth analysis and refined algorithms; see Vu and Lei (2013), Berthet and Rigollet (2013) and Ma (2013). Taking a different route, Bickel and Levina (2008) advocated thresholding as a regularization approach to estimate a sparse matrix, in the sense that most entries of the matrix are close to zero and this approach was used independently in Fan et al. (2008) for estimating covariance matrix with factor structure.

Another challenge in high-dimensional statistics is that measurements may not have light tails. For example, large scale datasets are often obtained by using bio-imaging technology (e.g., fMRI and microarrays) that often leads to heavy-tailed

* Correspondence to: Department of ORFE, Sherrerd Hall, Princeton University, Princeton, NJ 08544, USA.
E-mail addresses: jqfan@princeton.edu (J. Fan), weichenw@princeton.edu (W. Wang), yiqiaoz@princeton.edu (Y. Zhong).

https://doi.org/10.1016/j.jeconom.2018.09.003
0304-4076/

measurement errors (Dinov et al., 2005). Moreover, it is well known that financial returns exhibit heavy tails. These invalidate the fundamental assumptions in high-dimensional statistics that data have sub-Gaussian or sub-exponential tails, popularly imposed in most of the aforementioned papers. Significant relaxation of the assumption requires some new ideas and forms the subject of this paper.

Recently, motivated by Fama–French model (Fama and French, 1993) from financial econometrics, Fan et al. (2008) and Fan et al. (2013) considered the covariance structure of the *static approximate factor model*, which models the covariance matrix by a low-rank signal matrix and a sparse noise matrix. The same model will also be the focus of the current paper. The model assumes existence of several low-dimensional factors that drive a large panel data $\{y_{it}\}_{i\le p, t\le n}$, that is

$$y_{it} = b_i^T f_t + u_{it}\,, \qquad i \le p,\ t \le n, \tag{1.1}$$

where f_t's are the common factors, which are observed; and b_i's are their corresponding factor loadings, which are considered as unknown but fixed parameters in this work. The noises u_{it}'s, known as the idiosyncratic component, are uncorrelated with the factors $f_t \in \mathbf{R}^r$. Here r is relatively small compared with p and n. We will treat r as fixed and independent of p and n throughout this paper. When the factors are known, this model subsumes the well-known CAPM model (Sharpe, 1964; Lintner, 1965) and Fama–French model (Fama and French, 1993). When f_t is unobserved, the model tries to recover the underlying factors for the movements of the whole panel data. Here the "approximate" factor model indicates that the covariance Σ_u of $u_t = (u_{1t}, \ldots, u_{pt})$ is sparse, including the strict factor model in which Σ_u is diagonal as a special case. In addition, "static" is a specific case of the dynamic model which takes into account the time lag and allows more general infinite dimensional representations (Forni et al., 2000; Forni and Lippi, 2001).

The covariance matrix of the outcome $y_t = (y_{1t}, \ldots, y_{pt})'$ from model (1.1) can be written as

$$\Sigma = B\Sigma_f B^T + \Sigma_u\,, \tag{1.2}$$

where $B \in \mathbf{R}^{p\times r}$ is the loading matrix consisting of b_i^T in each row, Σ_f is the covariance of f_t and Σ_u is the sparse covariance matrix for u_t. Here we assume the process of (f_t, u_t) is stationary so that Σ_f, Σ_u do not change over time. When factors are unknown, Fan et al. (2013) proposed applying PCA to obtain an estimate of the low rank part and sparse part Σ_u. The crucial assumption is that the factors are *pervasive*, meaning that the factors have non-negligible effects on a large amount of dimensions of the outcomes. Wang and Fan (2017) gave more explanation from the perspective of random matrix theories and relaxed the pervasiveness assumption in applications such as risk management and estimation of the false discovery proportion. See Onatski (2012) for more discussions on strong and weak factors.

In this paper, we consider estimating Σ with known factors. Unknown factors pose more difficulties for robust estimation, which will be explored in future works. The main focus of the paper is on robustness instead of factor recovery. Under exponential tails of the factors and noises, Fan et al. (2011) proposed the idea of performing thresholding on the estimate of Σ_u, obtained from the sample covariance of the residuals of multiple regression (1.1). The legitimacy of this approach hinges on the assumption that the tails of the factor and error distributions decay exponentially, which is likely to be violated in practice, especially in the financial applications. Thus, the need to extend the applicability of this approach beyond well-behaved noise has driven further research such as Fan et al. (2018), in which it is assumed that y_t has an elliptical distribution (Fang et al., 1990).

This paper studies model (1.1) under a much more relaxed condition: the random variables f_t and u_{it} have finite fourth moments. The main observation that motivates our method is that, the joint covariance matrix of $(y_t^T, f_t^T)^T$ supplies sufficient information to estimate $B\Sigma_f B^T$ and Σ_u. To estimate the joint covariance matrix in a robust way, the classical idea that dates back to Huber (1964) proves to be vital and effective. The novelty here is that we let the parameter diverge in order to control the bias in high-dimensional setting. The Huber loss function with a diverging parameter, together with other similar functions, has been shown to produce concentration bounds for M-estimators, when the random variables have heavy tails; see for example Catoni (2012) and Fan et al. (2017). This point will be clarified in Sections 2 and 3. The M-estimators considered here have additional merits in asymptotic analysis, which is studied in Section 3.3.

This paper can be placed in the broader context of low rank plus sparse representation. In the past few years, robust principal component analysis has received much attention among statisticians, applied mathematicians and computer scientists. Their focus is on identifying the low rank component and sparse component from a corrupted matrix (Chandrasekaran et al., 2011; Candès et al., 2011; Xu et al., 2010). However, the matrices considered therein do not come from random samples, and as a result, neither estimation nor inference is involved. While Agarwal et al. (2012) considered the noisy decomposition, still the focus is more on identifying and separating the low rank part and sparse part. In spite of connections with the robust PCA literature, such as the incoherence condition (see Section 2), this paper and its predecessors are more engaged in disentangling "true signal" from noise, in order to improve estimation of covariance matrices. In this respect, they bear more similarity to the literature of covariance matrix estimation.

We make a few notational definitions before presenting the main results. For a general matrix M, the max-norm of M, or the entry-wise maximum, is denoted as $\|M\|_\infty = \max_{ij}|M_{ij}|$. The operator norm of M is $\|M\| = \lambda_{\max}^{1/2}(M^TM)$ whereas the Frobenius norm is $\|M\|_F = \sqrt{\sum_{ij} M_{ij}^2}$. If, furthermore, M is symmetric, we denote $\lambda_j(M)$ as the jth largest eigenvalue, $\lambda_{\max}(M)$ as the largest one, and $\lambda_{\min}(M)$ as the smallest one. In the paper, C is a generic constant that may differ from line to line in both assumptions and proofs.

The paper is organized as follows. In Section 2, we present the procedure of robust covariance estimation where we only assume finite fourth moments for both factors and noises without specific distributional assumptions. The theoretical

justification will be provided in Section 3. Simulations will be carried out in Section 4 to demonstrate the effectiveness of the proposed procedure. We also conduct real data analysis on portfolio risk of S&P stocks via Fama–French model in Section 5. Technical proofs will be delayed to the appendix.

2. Robust covariance estimation

Consider the factor model (1.1) again with observed factors. It can be written in the vector form as

$$y_t = Bf_t + u_t\,, \tag{2.1}$$

where $y_t = (y_{1t}, \ldots, y_{pt})^T, f_t \in \mathbf{R}^r$ are the factors for $t = 1, \ldots, T$, $B = (b_1, \ldots, b_p)^T$ is the fixed unknown loading matrix and $u_t = (u_{1t}, \ldots, u_{pt})^T$ is uncorrelated with the factors. We assume that (u_t^T, f_t^T) have zero mean and they are independent for $t = 1, 2, \ldots, T$. A motivating example from economic and financial studies is the classical Fama–French model, where y_{it}'s represent excess returns of stocks in the market and f_t's are interpreted as common factors driving the market. It is more natural to allow for weak temporal dependence such as α-mixing as in the work of Fan et al. (2016). Though possible, we assume independence in this paper for the sake of simplicity of analysis.

2.1. Assumptions

We now state the main assumptions of the model. Let Σ_f be the covariance of f_t, and Σ_u the covariance of u_t. A covariance decomposition shows that Σ, the covariance of y_t, comprises two parts,

$$\Sigma = B\Sigma_f B^T + \Sigma_u\,. \tag{2.2}$$

We assume that Σ_u is sparse and the sparsity level is measured through

$$m_q = \max_{i\le p} \sum_{j\le p} |(\Sigma_u)_{ij}|^q, \qquad \text{for some } q \in [0, 1]. \tag{2.3}$$

If $q = 0$, m_q is defined to be $\max_{i\le p}\sum_{j\le p} \mathbb{1}((\Sigma_u)_{ij} \neq 0)$, i.e. the exact sparsity. An intuitive justification of the sparsity measurement stems from modeling of the covariance structure: after taking out the common factors, the rest only has weak cross-sectional dependence. In addition, we assume that $\|\Sigma_u\|$, as well as $\|\Sigma_f\|$, is bounded away from 0 and ∞. In the case of degenerate Σ_f, we can always consider rescaling the factors and reduce the number of observed factors to meet the requirement of non-vanishing minimum eigenvalue of Σ_f. This leads to our first assumption.

Assumption 2.1. There exists a constant $C > 0$ such that $C^{-1} \le \|\Sigma_u\| \le C$ and $C^{-1} \le \|\Sigma_f\| \le C$, where Σ_f is a $r \times r$ matrix with r being a fixed number.

Here assuming a fixed r is just for simplicity of presentation. It can be allowed to grow with n and p. Then we would need to keep track of r in the theoretical analysis and impose certain growth condition on r.

Another important feature of the factor model, observed by Stock and Watson (2002), is that the factors are *pervasive* in the sense that the low rank part of (2.2) is the dominant component of Σ; more specifically, the top r eigenvalues grow linearly as p. This motivates the following assumption.

Assumption 2.2. (i) There exists a constant $c > 0$ such that $\lambda_r(\Sigma) > cp$.

(ii) The elements of B are uniformly bounded by a constant C.

First note, assumption (ii) implies that $\lambda_1(\Sigma) \le \lambda_1(B\Sigma_f B^T) + \|\Sigma_u\| \le \lambda_1(\Sigma_f)\lambda_1(B^TB) + \|\Sigma_u\| = O(p)$. So together with (i), the above assumption requires leading eigenvalues to grow with the same order as p. This assumption is satisfied by the approximate factor model, since by Weyl's inequality, $\lambda_i(\Sigma)/p = \lambda_i(B\Sigma_f B^T)/p + o(1)$ if the main term is bounded from below. Furthermore, for illustrative purposes, if we additionally assume (though not needed in this paper) that each entry of B is iid with a finite second moment, it is not hard to see $\lambda_i(B\Sigma_f B^T)/p = \lambda_i(\Sigma_f(B^TB/p))$ satisfies such a condition with probability tending to one. Consequently, it is natural to assume $\lambda_i(\Sigma)/p$ is lower bounded for $i \le r$. Note that B is considered to be deterministic throughout the paper.

Assumption (ii) is related to the matrix incoherence condition. In fact, when $\lambda_{\max}(\Sigma)$ grows linearly with p, the condition of bounded $\|B\|_\infty$ is equivalent to an incoherent structure of top eigenvectors of Σ, which is standard in the matrix completion literature (Candès and Recht, 2009) and the robust PCA literature (Chandrasekaran et al., 2011).

We now consider the moment assumption of random variables in model (1.1).

Assumption 2.3. (f_t, u_t) is iid with mean zero and bounded fourth moment. That is, there exists a constant $C > 0$ such that $\max_k Ef_{kt}^4 < C$ and $\max_i Eu_{it}^4 < C$.

The independence assumption can be relaxed to mixing conditions, but we do not pursue this direction in the current paper. Note that our main Theorem 3.1 is essentially deterministic. So under certain mixing condition such as that used by Fan et al. (2011), as long as we achieve a max-norm error bound (3.2) in Corollary 3.1, all conclusions in Theorem 3.1 follow immediately. More details are in Section 3.

We are going to establish our results based on the above assumption which only requires bounded fourth moments.

2.2. Robust estimation procedure

The basic idea we propose is to, in the first step, estimate the covariance matrix of the joint vector (y_t, f_t) instead of simply the covariance of y_t, although the latter is our target. The covariance of the concatenated $p + r$ dimensional vector $z_t^T = (y_t^T, f_t^T)$ contains sufficient information to recover the low-rank and sparse structure. Observe that the covariance matrix $\Sigma_z := \mathrm{Cov}(z_t)$ can be expressed as

$$\Sigma_z = \begin{pmatrix} B\Sigma_f B^T + \Sigma_u & B\Sigma_f \\ \Sigma_f B^T & \Sigma_f \end{pmatrix} =: \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Any method which yields an estimate of Σ_z as an initial estimator or estimates of $\widehat{\Sigma}_{11}$, $\widehat{\Sigma}_{12}$, $\widehat{\Sigma}_{21}$, $\widehat{\Sigma}_{22}$ could be used to infer the unknown B, Σ_f and Σ_u. Specifically, using the estimator $\widehat{\Sigma}_z$, we can readily obtain an estimator of $B\Sigma_f B^T$ through the identity

$$B\Sigma_f B^T = \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

Subsequently, we can subtract the estimator of $B\Sigma_f B^T$ from $\widehat{\Sigma}_{11}$ to obtain $\widehat{\Sigma}_u$. With the sparsity structure of Σ_u assumed in Section 2.1, the well-studied thresholding (Bickel and Levina, 2008; Rothman et al., 2009; Cai and Liu, 2011) can be employed. Applying thresholding to $\widehat{\Sigma}_u$, we obtain a thresholded matrix $\widehat{\Sigma}_u^{\mathcal{T}}$ with guaranteed error in terms of the max-norm and the operator norm. The final step is to add $\widehat{\Sigma}_u^{\mathcal{T}}$ with the estimator of $B\Sigma_f B^T$ (from $\widehat{\Sigma}_z$ in the first step) to produce the final estimator $\widehat{\Sigma}^{\mathcal{T}}$ of Σ.

Due to the fact that we only assume bounded fourth moments for factors and errors, we estimate the covariance matrix Σ_z through robust methodology. For the sake of simplicity, we assume the vector z_t has zero mean, so the covariance matrix of z_t takes the form $Ez_t z_t^T$. We shall use the M-estimator proposed in Catoni (2012) and Fan et al. (2017), where the authors proved the concentration property in the estimation of population mean of a random variable with a finite second moment. Here the variables of interest are the entries of $z_t z_t^T$, and naturally we need bounded fourth moments of z_t.

In essence, minimizing a suitable loss function, say Huber loss, yields an estimator of the population mean with a deviation of order $n^{-1/2}$. The Huber loss reads

$$l_\alpha(x) = \begin{cases} 2\alpha|x| - \alpha^2, & |x| > \alpha, \\ x^2, & |x| \le \alpha. \end{cases} \tag{2.4}$$

Choosing $\alpha = \sqrt{(nv^2)/\log(\epsilon^{-1})}$, $\epsilon \in (0, 1)$ where v is an upper bound of the standard deviation of the iid random variables X_i of interest, Fan et al. (2017) showed that the minimizer $\widehat{\mu} = \mathrm{argmin}_\mu \sum_{i=1}^n l_\alpha(X_i - \mu)$ satisfies

$$P\left(|\widehat{\mu} - \mu| \le 4v\sqrt{\frac{\log(\epsilon^{-1})}{n}}\right) \ge 1 - 2\epsilon, \tag{2.5}$$

when $n \ge 8\log(\epsilon^{-1})$ where $\mu = EX_i$. This finite sample result holds for any distributions with bounded second moments, including asymmetric distributions generated by Z^2. This assumption of bounded second moments for mean estimation translates into a fourth moments assumption for our covariance estimation, because covariances are products of two random variables. When applying (2.5), we will take X_i to be the square of a random variable or products of two random variables. The diverging parameter α is chosen to reduce the bias of the M-estimator for asymmetric distributions. When applying this method to estimate Σ_z element-wisely, we expect $\widehat{\Sigma}_{11}$, $\widehat{\Sigma}_{12}$, $\widehat{\Sigma}_{21}$, $\widehat{\Sigma}_{22}$ to achieve a max-norm error of $O_P(\sqrt{\log p/n})$, where the logarithmic term is incurred when we bound the errors uniformly. The formal result will be given in Section 3.

In an earlier work, Catoni (2012) proposed solving the equation $\sum_{i=1}^n h[\alpha^{-1}(\mu - \widehat{\mu})] = 0$, where the strictly increasing $h(x)$ satisfies $-\log(1 - x + x^2/2) \le h(x) \le \log(1 + x + x^2/2)$. For $\epsilon \in (0, 1)$ and $n > 2\log(\epsilon^{-1})$, Catoni (2012) proved that

$$P\left(|\widehat{\mu} - \mu| \le v\sqrt{\frac{2\log(\epsilon^{-1})}{n - 2\log(\epsilon^{-1})}}\right) \ge 1 - 2\epsilon,$$

when $n \ge 4\log(\epsilon^{-1})$ and $\alpha = \sqrt{nv^2(1 + \frac{2\log(\epsilon^{-1})}{n-2\log(\epsilon^{-1})})/\{2\log(\epsilon^{-1})\}}$, where v is an upper bound of the standard deviation. This M-estimator can be also used for covariance estimation, though it usually has a larger bias as shown in Fan et al. (2017).

The whole procedure can be presented in the following steps:

Step 1 For each entry of the covariance matrix Σ_z, obtain a robust estimator by solving a convex minimization problem (through, for example, Newton–Raphson method):

$$(\widehat{\Sigma}_z^R)_{ij} = \mathrm{argmin}_x \sum_{t=1}^n l_\alpha(z_{it}z_{jt} - x), \tag{2.6}$$

where α is chosen as discussed above and $\widehat{\Sigma}_z = \widehat{\Sigma}_z^R = \begin{pmatrix} \widehat{\Sigma}_{11} & \widehat{\Sigma}_{12} \\ \widehat{\Sigma}_{21} & \widehat{\Sigma}_{22} \end{pmatrix}$.

Step 2 Derive an estimator of Σ_u through the algebraic manipulation

$$\widehat{\Sigma}_u = \widehat{\Sigma}_{11} - \widehat{\Sigma}_{12}\widehat{\Sigma}_{22}^{-1}\widehat{\Sigma}_{21},$$

and then apply adaptive thresholding of Cai and Liu (2011). That is,

$$(\widehat{\Sigma}_u^{\mathcal{T}})_{ij} = \begin{cases} (\widehat{\Sigma}_u)_{ij}, & i = j \\ s_{ij}((\widehat{\Sigma}_u)_{ij})\mathbb{1}(|(\widehat{\Sigma}_u)_{ij}| \geq \tau_{ij}), & i \neq j \end{cases}$$

where $s_{ij}(\cdot)$ is the generalized shrinkage function (Antoniadis and Fan, 2001; Rothman et al., 2009) and $\tau_{ij} = \tau((\widehat{\Sigma}_u)_{ii}(\widehat{\Sigma}_u)_{jj})^{1/2}$ is an entry-dependent threshold.

Step 3 Produce the final estimator for Σ:

$$\widehat{\Sigma}^{\mathcal{T}} = \widehat{\Sigma}_{12}\widehat{\Sigma}_{22}^{-1}\widehat{\Sigma}_{21} + \widehat{\Sigma}_u^{\mathcal{T}}.$$

Note in the above steps, the choice of the parameters v (in the definition of α) and τ_{ij} is not yet specified and will be discussed in Section 3.

There are $p(p+1)/2$ adaptive Huber estimators (2.6) that we need to compute in Step 1. Since all these Huber minimization problems share a similar structure, it is possible to speed up the computation by choosing the initial values smartly in practice, though the optimization is already fast in our simulations.

Before delving into the analysis of the procedure, we first deviate to look at a technical issue. Recall that $\widehat{\Sigma}_{22}$ is an estimator of Σ_f, by Weyl's inequality,

$$|\lambda_i(\widehat{\Sigma}_{22}) - \lambda_i(\Sigma_f)| \leq \|\widehat{\Sigma}_{22} - \Sigma_f\|.$$

Since both matrices are of low dimensionality, as long as we are able to estimate each entry of Σ_f with enough accuracy (see Lemma 3.1), $\|\widehat{\Sigma}_{22} - \Sigma_f\|$ vanishes with high probability as n diverges. Therefore, with high probability, $\widehat{\Sigma}_{22}$ is invertible, and there is no major issue implementing the procedure. In cases where positive semidefinite matrix is required, we can refine the matrix by projecting it to its nearest positive semidefinite version in terms of the max-norm. This projection can be done for both $\widehat{\Sigma}_u$ and $\widehat{\Sigma}_z$. For example, for $\widehat{\Sigma}_u$, we solve the following optimization problem:

$$\widetilde{\Sigma}_u = \operatorname*{argmin}_{\Sigma_u \succeq 0} \|\widehat{\Sigma}_u - \Sigma_u\|_\infty, \tag{2.7}$$

and use $\widetilde{\Sigma}_u$ as our estimate. Observe that

$$\|\widetilde{\Sigma}_u - \Sigma_u\|_\infty \leq \|\widetilde{\Sigma}_u - \widehat{\Sigma}_u\|_\infty + \|\widehat{\Sigma}_u - \Sigma_u\|_\infty \leq 2\|\widehat{\Sigma}_u - \Sigma_u\|_\infty.$$

Thus, except for a slightly worse constant, $\widetilde{\Sigma}_u$ inherits all the desired properties of $\widehat{\Sigma}_u$ (namely good convergence rates), as we will see in Section 3 that those properties would follow as soon as a max-norm bound holds. Hence we are able to replace $\widehat{\Sigma}_u$ with $\widetilde{\Sigma}_u$ without modifying our estimation procedure. Moreover, (2.7) can be cast into the semidefinite programming problem below,

$$\min_{t, \Sigma_u \succeq 0} t \quad \text{s.t.} \quad |\widehat{\Sigma}_u - \Sigma_u|_{ij} \leq t, \tag{2.8}$$

which can be solved by a semidefinite programming solver, e.g. Grant et al. (2008).

3. Theoretical analysis

In this section, we will show the theoretical properties of our robust estimator under bounded fourth moments. We will also show that when the data are known to be generated from more restricted families (e.g. sub-Gaussian), commonly used estimators, such as the sample covariance estimator, suffice as the initial estimator in Step 1.

3.1. General theoretical properties

From the above discussion on M-estimators and their concentration results, it is immediate to have the following lemma.

Lemma 3.1. *Suppose that a d-dimensional random vector X is centered and has finite fourth moment, i.e.* $EX = 0$, $\max_i EX_i^4 < +\infty$ *for* $i = 1, 2, \ldots, p$. *Let* $\sigma_{ij} = E(X_iX_j)$ *and* $\widehat{\sigma}_{ij}$ *be Huber's estimator with parameter* $\alpha = \sqrt{nv^2/\log(p^2/\delta)}$, *then there exists a universal constant C such that for any* $\delta \in (0, 1)$ *and* $n \geq C\log(p/\delta)$, *with probability* $1 - \delta$,

$$\max_{ij} \left|\widehat{\sigma}_{ij} - \sigma_{ij}\right| \leq Cv\sqrt{\frac{\log p + \log(1/\delta)}{n}}, \tag{3.1}$$

where v is a pre-determined parameter satisfying $v^2 \geq \max_{i,j\leq p} Var(X_iX_j)$.

In practice, we do not know any of the fourth moments in advance. To pick up a good v, one possibility is Lepski's adaptation method (Lepskii, 1992) where a sequence of geometrically increasing v is tried and the estimated v is picked up as the middle of the smallest confidence interval intersecting all the larger ones. See Catoni (2012) for details. Alternatively, we may simply use the empirical variance to give a rough bound of v, in a way similar to Fan et al. (2018).

Recall that z_t is a $p+r$ dimensional vector concatenating y_t and f_t. From Assumption 2.3, there is a constant C_0 as a uniform bound for Ez_{it}^4. This leads to the following result.

Corollary 3.1. *Suppose that $\widehat{\Sigma}_z$ is an estimator of covariance matrix Σ_z, whose entries are Huber's estimators with parameter $\alpha = \sqrt{nv^2/\log((p+r)^2/\delta)}$. Then there exists a universal constant C such that for any $\delta \in (0,1)$ and $n \geq C\log(p/\delta)$, with probability $1-\delta$,*

$$\|\widehat{\Sigma}_z - \Sigma_z\|_\infty \leq Cv\sqrt{\frac{\log p + \log(1/\delta)}{n}}, \tag{3.2}$$

where v is a pre-determined parameter satisfying $v^2 \geq C_0$.

After Step 1 of the proposed procedure, we obtain an estimator $\widehat{\Sigma}_z$ that achieves the optimal rate of element-wise convergence. With $\widehat{\Sigma}_z$, we proceed to establish convergence rates for both $\widehat{\Sigma}_u^{\mathcal{T}}$ and $\widehat{\Sigma}^{\mathcal{T}}$. The key theorem that links the estimation error under element-wise max-norm with other metrics is stated below.

Theorem 3.1. *Under Assumptions 2.1–2.3, if we have estimator $\widehat{\Sigma}_z$ satisfying*

$$\|\widehat{\Sigma}_z - \Sigma_z\|_\infty = O_P(\sqrt{\log p/n}), \tag{3.3}$$

then the three-step procedure in Section 2.2 with $\tau \asymp \sqrt{\log p/n}$ generates $\widehat{\Sigma}_u^{\mathcal{T}}$ and $\widehat{\Sigma}^{\mathcal{T}}$ satisfying

$$\|\widehat{\Sigma}_u^{\mathcal{T}} - \Sigma_u\|_2 = \|(\widehat{\Sigma}_u^{\mathcal{T}})^{-1} - \Sigma_u^{-1}\|_2 = O_P\Big(m_p\Big(\frac{\log p}{n}\Big)^{(1-q)/2}\Big), \tag{3.4}$$

and furthermore

$$\|\widehat{\Sigma}^{\mathcal{T}} - \Sigma\|_\infty = O_P\Big(\sqrt{\frac{\log p}{n}}\Big), \tag{3.5}$$

$$\|\widehat{\Sigma}^{\mathcal{T}} - \Sigma\|_\Sigma = O_P\Big(\frac{\sqrt{p}\log p}{n} + m_p\Big(\frac{\log p}{n}\Big)^{(1-q)/2}\Big), \tag{3.6}$$

$$\|(\widehat{\Sigma}^{\mathcal{T}})^{-1} - \Sigma^{-1}\| = O_P\Big(m_p\Big(\frac{\log p}{n}\Big)^{(1-q)/2}\Big), \tag{3.7}$$

where $\|A\|_\Sigma = p^{-1/2}\|\Sigma^{-1/2}A\Sigma^{-1/2}\|_F$ is the relative Frobenius norm defined in Fan et al. (2008), if n is large enough so that $m_p(\log p/n)^{(1-q)/2}$ is bounded.

Theorem 3.1 provides a nice interface connecting the max-norm guarantee with the desired convergence rates. Therefore, any robust method that attains the element-wise optimal rate as in Corollary 3.1 can be used in Step 1 instead of the current M-estimator approach.

3.2. Estimators under more restricted distributional assumptions

We analyzed theoretical properties of the robust procedure in the previous subsection under the assumption of bounded fourth moments. Theorem 3.1 shows that any estimator that achieves the optimal max-norm convergence rate could serve as an initial pilot estimator for Σ_z to be used in Step 2 and Step 3 of our procedure. Thus the procedure depends on the distributional assumption (Assumption 2.3) only through Step 1 where a proper estimator $\widehat{\Sigma}_z$ is proposed. Sometimes, we do have more information on the shapes of the distributions of factors and noises. For example, if the distribution of $z_t = (f_t^T, u_t^T)^T$ has a sub-Gaussian tail, the sample covariance matrix $\widehat{\Sigma}_z^S = n^{-1}\sum_{t=1}^n z_t z_t^T$ attains the optimal element-wise maximal rate for estimating Σ_z.

In an earlier work, Fan et al. (2011) proposed to simply regress observations y_t on f_t in order to obtain

$$\widehat{B} = Y^T F(F^T F)^{-1}, \tag{3.8}$$

where $Y = (y_1, \ldots, y_n)^T$ and $F = (f_1, \ldots, f_n)^T$. Then they thresholded the matrix $\widehat{\Sigma}_u = \widehat{\Sigma} - \widehat{B}\widehat{\Sigma}_f\widehat{B}^T$ where $\widehat{\Sigma} = n^{-1}YY^T$ and $\widehat{\Sigma}_f = n^{-1}F^TF$. This regression-based method is equivalent to applying $\widehat{\Sigma}_z^S$ directly in Step 1 and also equivalent to solving a least-square minimization problem, and thus suffers from robustness issue when the data come from heavy-tailed distributions. All the convergence rates achieved in Theorem 3.1 are identical with Fan et al. (2011) where exponentially decayed tails are assumed.

As we explained, if z_t is sub-Gaussian distributed, $\widehat{\Sigma}_z^S$ instead of $\widehat{\Sigma}_z^R$ can be used. If f_t and u_t exhibit heavy tails, another widely used assumption is multivariate t-distribution, which is included in the elliptical distribution family. The elliptical

distribution is defined as follows. Let $\mu \in \mathbf{R}^p$ and $\Sigma \in \mathbf{R}^{p\times p}$ with $\text{rank}(\Sigma) = q \le p$. A p-dimensional random vector y has an elliptical distribution, denoted by $y \sim ED_p(\mu, \Sigma, \zeta)$, if it has a stochastic representation (Fang et al., 1990)

$$y \overset{d}{=} \mu + \zeta AU\,, \tag{3.9}$$

where U is a uniform random vector on the unit sphere in $\mathbf{R}^q$, $\zeta \ge 0$ is a scalar random variable independent of U, $A \in \mathbf{R}^{p\times q}$ is a deterministic matrix satisfying $AA' = \Sigma$. To make the representation (3.9) identifiable, we require $\mathbb{E}\zeta^2 = q$ so that $\text{Cov}(y) = \Sigma$. Here we also assume continuous elliptical distributions with $\mathbb{P}(\zeta = 0) = 0$.

If f_t and u_t are uncorrelated and jointly elliptical, i.e., $z_t = (f_t^T, u_t^T)^T \sim ED_p(0, \text{diag}(\Sigma_f, \Sigma_u), \zeta)$, then a well-known good estimator for the correlation matrix R of z_t is the marginal Kendall's tau. Kendall's tau correlation coefficient is defined as

$$\hat\tau_{jk} := \frac{2}{n(n-1)}\sum_{i<i'} \text{sgn}((z_{ij} - z_{i'j})(z_{ik} - z_{i'k}))\,, \tag{3.10}$$

whose population counterpart is

$$\tau_{jk} := \mathbb{P}((z_{1j} - z_{2j})(z_{1k} - z_{2k}) > 0) - \mathbb{P}((z_{1j} - z_{2j})(z_{1k} - z_{2k}) < 0)\,. \tag{3.11}$$

For the elliptical family, the key identity $r_{jk} = \sin(\pi\tau_{jk}/2)$ relates Pearson correlation to Kendall's correlation (Fang et al., 1990). Using $\hat r_{jk} = \sin(\pi\hat\tau_{jk}/2)$, Han and Liu (2014) showed that $\widehat R$ is an accurate estimate of R, achieving $\|\widehat R - R\|_\infty = O_P(\sqrt{\log p/n})$. Let $\Sigma_z = DRD$ where R is the correlation matrix and $D = \text{diag}(\sigma_1, \ldots, \sigma_p)$ is a diagonal matrix consisting of standard deviations for each dimension. We construct $\widehat\Sigma_z^K$ by separately estimating D and R. As before, if the fourth moment exists, we estimate D by only considering $i = j$ in Step 1, namely by using the adaptive Huber method.

Therefore, if z_t is elliptically distributed, $\widehat\Sigma_z^K$ can be used as the initial pilot estimator for Σ_z in Step 1. Note that, unlike $\widehat\Sigma_z^K$, there is no closed-form expression for $\widehat\Sigma_z^R$. However, for general heavy-tailed distributions, there is no simple way to connect the Pearson correlation with Kendall's correlation. Thus we should favor $\widehat\Sigma_z^R$ instead. We will compare the three estimators $\widehat\Sigma_z^S$, $\widehat\Sigma_z^K$ and $\widehat\Sigma_z^R$ thoroughly through simulations in Section 4.

3.3. Asymptotics of robust mean estimators

In this section we look further into robust mean estimators. Though the result we shall present is asymptotic and not essential for our main Theorem 3.1, it is interesting in its own right and deserves some treatment.

Perhaps the best known result of Huber's mean estimator is the asymptotic minimax theory. Huber (1964) considered the so-called ϵ-contamination model:

$$\mathcal{P}_\epsilon = \{F \mid F(x) = (1-\epsilon)G(x-\theta) + \epsilon H(x),\ H \in \mathcal{F}, \theta \in \mathbf{R}\},$$

where G is a known distribution, ϵ is fixed and $\mathcal{F}$ is the family of symmetric distributions. Let T_n be the minimizer of $\sum_{i=1}^n \rho_H(x_i - \mu)$, where $\rho_H(x) = x^2/2$ for $|x| < \alpha$, and $\rho_H(x) = \alpha|x| - \alpha^2/2$ for $|x| \ge \alpha$, where α is fixed. In the special case where G is Gaussian, Huber's result showed that with an appropriate choice of α, Huber's estimator minimizes the maximal asymptotic variance among all translation invariant estimators, the maximum being taken over $\mathcal{P}_\epsilon$.

One problem with ϵ-contamination model is that it makes sense only when we assume symmetry of H, if θ is the quantity we are interested in. In contrast, Catoni (2012) and Fan et al. (2017) studied a different family, in which distributions have finite second moments. Bickel (1976) called them "local" and "global" models respectively, and offered a detailed discussion.

This paper, along with the preceding two papers (Catoni, 2012; Fan et al., 2017), studies robustness in the sense of the second model. The technical novelty primarily lies in the nice concentration property, which is fundamental to high dimensional statistics. This requires the parameter α of ρ_H to grow with n, versus being kept fixed, such that the condition in Corollary 3.1 is satisfied. It turns out that, in addition to the concentration property, we can establish results regarding its asymptotic behaviors in an exact manner.

Let $\rho_n(x) = x^2/2$ for $|x| < \alpha_n$ and $\rho_n(x) = \alpha_n|x| - \alpha_n^2/2$ for $|x| \ge \alpha_n$; its derivative $\psi_n = \rho_n'$. Let us write $\lambda_n(t) = E\psi_n(X - t)$. Denote t_n as a solution of $\lambda_n(t) = 0$, which is unique when n is sufficiently large, and T_n a solution of $\sum_{i=1}^n \psi_n(x_i - t) = 0$. We have the following theorem.

Theorem 3.2. *Suppose that $x_1, \ldots, x_n$ is drawn from some distribution F with mean μ and finite variance σ^2. Suppose $\{\alpha_n\}$ is any sequence with $\lim_{n\to\infty}\alpha_n = \infty$. Then, as $n \to \infty$,*

$$\sqrt{n}(T_n - t_n) \xrightarrow{d} N(0, \sigma^2),$$

and moreover

$$\frac{t_n - \mu}{E\psi_n(X - \mu)} \to 1.$$

Theorem 3.2 gives a decomposition of error $T_n - \mu$ into two components: variance and bias. The rate of bias $E\psi_n(X-\mu)$ depends on the distribution F and $\{\alpha_n\}$. When the distribution is either symmetric or $\liminf_n \alpha_n/\sqrt{n} > 0$, the second component $t_n - \mu$ is $o(1/\sqrt{n})$, a negligible quantity compared with the asymptotic variance. Note that unlike Huber's original approach, our robust estimator does not require symmetric restriction. This theorem also lends credibility to the bias–variance tradeoff we observed in the simulation (see Section 4.1).

It is worth comparing the above Huber loss minimization with another candidate for robust mean estimation called "median-of-means" given by Hsu and Sabato (2014). The method, as its name suggests, first divides samples into k subgroups and calculates means for each subgroup, then takes the median of those means as the final estimator. The first step basically symmetrizes the distribution by the central limit theorem and the second step is to robustify the procedure. According to Hsu and Sabato (2014), if we choose $k = 4.5\log(p/\delta)$ and element-wisely estimate Σ_z, similar to (2.5), with probability $1-\delta$, we have

$$\|\widehat{\Sigma}_z - \Sigma_z\|_\infty \le 3\sqrt{3}v\sqrt{\frac{\log p + \log(1/\delta)}{n}}.$$

Although "median-of-means" has the desired concentration property, unlike our estimator here, its asymptotic behavior differs from the empirical mean estimator, and as a consequence, it is not asymptotically efficient when the distribution F is Gaussian. Therefore, regarding efficiency, we prefer our proposed procedure in Section 2.2.

4. Simulations

We now present simulation results to demonstrate the improvement of the proposed robust method over the least-square based method (Fan et al., 2008, 2011) and Kendall's tau based method (Han and Liu, 2014; Fan et al., 2018) when factors and errors are (i) elliptically distributed; and (ii) generally heavy-tailed.

However, one must be cautious of the choice of the tuning parameter α, since it plays an important role in the quality of the robust estimates. Out of this concern, we shall discuss the intricacy of choosing parameter α before presenting the performance of robust estimates of covariance matrices.

4.1. Robust estimates of variances and covariances

For random variables $X_1, \ldots, X_p$ with zero mean that may potentially exhibit heavy-tailed behavior, the sample mean of $v_{ij} = E(X_iX_j)$ is not good enough for our estimation purpose. Though being unbiased, in the high dimensional setting, there is no guarantee that multiple sample means stay close to the true values simultaneously.

As shown in theoretical analysis, this problem is alleviated for robust estimators constructed through M-estimators, whose influence functions grow slowly at extreme values. The desired concentration property in (3.2) depends on the choice of parameter α, which decides the range outside which large values cease to become more influential. However, in practice, we have to make a good guess of $\mathrm{Var}(X_iX_j)$ as the theory suggests; even so, we may be too conservative in the choice of α.

To show this, we plot in Fig. 1 the histograms of our estimates of $v = \mathrm{Var}(X_i)$ in 1000 runs, where X_i is generated from a t-distribution with degree of freedom $\nu = 4.2$. The first three histograms show the estimates constructed from Huber's M-estimator, with parameter

$$\alpha = \beta\sqrt{\frac{n\,\mathrm{Var}(X_i^2)}{2}}, \tag{4.1}$$

where β is 0.2, 1, 5 respectively, and the last histogram is the usual sample estimate (or $\beta = \infty$). The quality of estimates ranges from large biases to large variances. We also plot in Fig. 2 the histograms of estimates of $v = \mathrm{Cov}(X_i, X_j)$, where (X_i, X_j), $i \ne j$ is generated from a multivariate t-distribution with $\nu = 4.2$ and an identity scale matrix. The only difference is that in (4.1), the variance of X_i^2 is replaced by the variance of X_iX_j.

From Fig. 1, we observe a bias–variance tradeoff phenomenon as α varies. This is also consistent with the theory in Section 3.3. When α is small, the robust method underestimates the variance, yielding a large bias due to the asymmetry of the distribution of X_i^2. As α increases, a larger variance is traded for a smaller bias, until $\alpha = \infty$, in which case the robust estimator simply becomes the sample mean.

For the covariance estimation, Fig. 2 exhibits a different phenomenon. Since the distribution of X_iX_j is symmetric for $i \ne j$, there is no bias incurred when α is small. Since the variance is smaller when α is smaller, we have a net gain in terms of the quality of estimates. In the extreme case where α is zero, we are actually estimating the median. Fortunately, under distributional symmetry, the mean and the median are the same.

The simple simulations help us to understand how to choose α in practice: if the distribution is close to a symmetric one, one can choose α aggressively, i.e. making α smaller; otherwise, a conservative α is preferred.

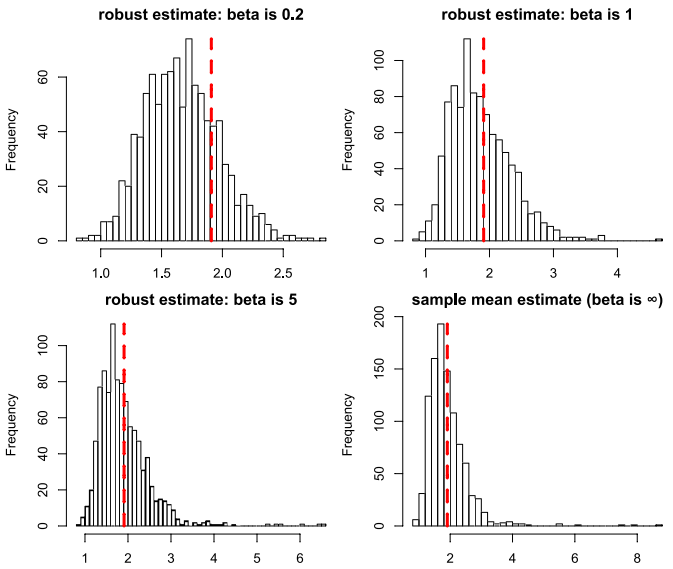

Fig. 1. The histograms show the estimates of $\mathrm{Var}(X_i)$ with different parameters α, parametrized by β via (4.1), in 1000 runs. $X_i \sim t_{4.2}$ so that the true variance $\mathrm{Var}(X_i) = 1.909$. The sample size $n = 100$.

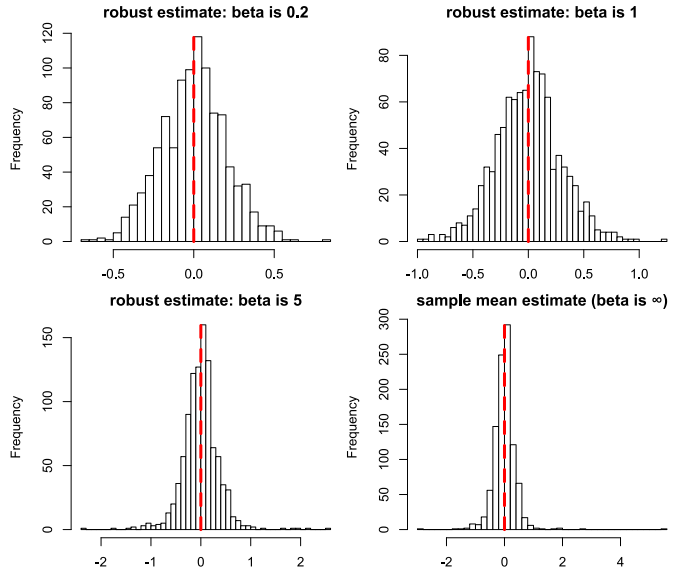

Fig. 2. The histograms show the estimates of $\mathrm{Cov}(X_i, X_j)$ with different parameters α in 1000 runs. The true covariance $\mathrm{Cov}(X_i, X_j) = 0$. $n = 100$ and the degree of freedom is 4.2.

4.2. Covariance matrix estimation

We implemented the robust estimation procedure with three initial pilot estimators $\widehat{\Sigma}_z^S$, $\widehat{\Sigma}_z^K$ and $\widehat{\Sigma}_z^R$. We simulated n samples of $z_t = (f_t^T, u_t^T)^T$ from a multivariate t-distribution with covariance matrix $\mathrm{diag}\{I_r, 5I_p\}$ and various degrees of freedom. Each row of B is independently generated from a standard normal distribution, and once it is generated, we treat it as fixed. The population covariance matrix of $y_t = Bf_t + u_t$ is $\Sigma = BB^T + 5I_p$. For p running from 200 to 900 and $n = p/2$, we calculated errors of the robust procedure in different norms. As suggested by the experiments in the previous section, we chose a larger parameter α to estimate the diagonal elements of Σ_z, and a smaller one to estimate its off-diagonal elements. We used the thresholding parameter $\tau = 2\sqrt{\log p/n}$.

The estimation errors are gauged in the following norms: $\|\widehat{\Sigma}_u^{\mathcal{T}} - \Sigma_u\|$, $\|(\widehat{\Sigma}^{\mathcal{T}})^{-1} - \Sigma^{-1}\|$ and $\|\widehat{\Sigma}^{\mathcal{T}} - \Sigma\|_\Sigma$ as shown in Theorem 3.1. We considered two different settings: (1) z_t is generated from multivariate t-distribution with very heavy

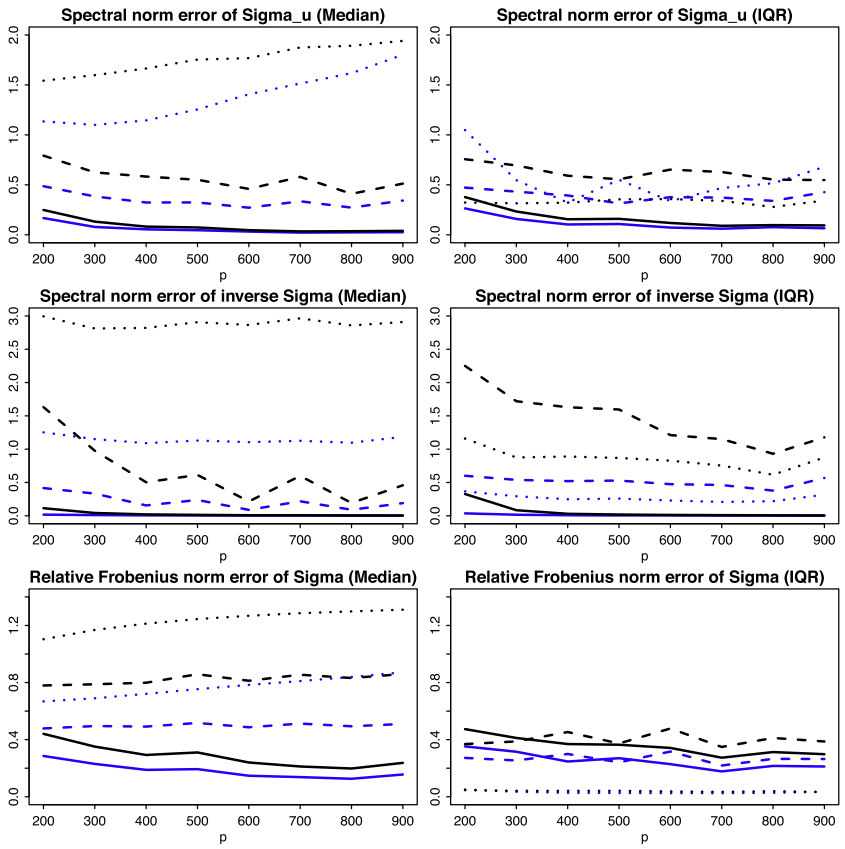

Fig. 3. Errors of robust estimates against varying p. Blue line represents ratio of errors with $\widehat{\Sigma}_z^R$ over errors with $\widehat{\Sigma}_z^S$, while black line represents ratio of errors with $\widehat{\Sigma}_z^K$ over errors with $\widehat{\Sigma}_z^S$. z_t is generated by multivariate t-distribution with $df = 3$ (solid), 5 (dashed) and ∞ (dotted). The median errors and their IQR over 100 simulations are reported. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

($\nu = 3$), medium heavy ($\nu = 5$), and light ($\nu = \infty$ or Gaussian) tail; (2) z_t is element-wise iid one-dimensional t-distribution with degree of freedom $\nu = 3$,5 and ∞. They are separately plotted in Figs. 3 and 4. The estimation errors of applying sample covariance matrix $\widehat{\Sigma}_z^S$ are used as the baseline for comparison. For example, if $\|\widehat{\Sigma}^{\mathcal{T}} - \Sigma\|_\Sigma$ is used to measure performance, the blue curve represents ratio $\|(\widehat{\Sigma}^{\mathcal{T}})^R - \Sigma\|_\Sigma / \|(\widehat{\Sigma}^{\mathcal{T}})^S - \Sigma\|_\Sigma$ while the black curve represents ratio $\|(\widehat{\Sigma}^{\mathcal{T}})^K - \Sigma\|_\Sigma / \|(\widehat{\Sigma}^{\mathcal{T}})^S - \Sigma\|_\Sigma$ where $(\widehat{\Sigma}^{\mathcal{T}})^R, (\widehat{\Sigma}^{\mathcal{T}})^K, (\widehat{\Sigma}^{\mathcal{T}})^S$ are respectively estimators given by the robust procedure with initial pilot estimators $\widehat{\Sigma}_z^R, \widehat{\Sigma}_z^K, \widehat{\Sigma}_z^S$ for Σ_z. Therefore if the ratio curve moves below 1, the method is better than the naive sample estimator given in Fan et al. (2011) and vice versa. The more it gets below 1, the more robust the procedure is against heavy-tailed randomness.

The first setting (Fig. 3) represents a heavy-tailed elliptical distribution, where we expect the two robust methods work better than the sample covariance based method, especially in the case of extremely heavy tails (solid lines for $\nu = 3$). As expected, both black curves and blue curves under the three measures behave visibly better (smaller than 1). On the other hand, if data are indeed Gaussian (dotted line for $\nu = \infty$), the method with sample covariance performs better under most measures (greater than 1). Nevertheless, our robust method still performs comparably with the sample covariance method, as the median error ratio stays around 1 whereas Kendall's tau method can be much worse than the sample covariance method. A plausible explanation is that the variance reduced compensates for the bias incurred in our procedure. In addition, the IQR (interquartile range) plots tell us the proposed robust method is indeed more stable than Kendall's tau.

The second setting (Fig. 4) provides an example of non-elliptical distributed heavy-tailed data. We can see that the performance of the robust method dominates the other two methods, which verifies the approach in this paper especially when data come from a general heavy-tailed distribution. While our method is able to deal with more general distributions, Kendall's tau method does not apply to distributions outside the elliptical family, which excludes the element-wise iid t distribution in this setting. This explains why under various measures, our robust method is better than Kendall's tau method by a clear margin. Note that even in the first setting where the data are indeed elliptical, with proper tuning, the proposed robust methods can still outperform Kendall's tau.

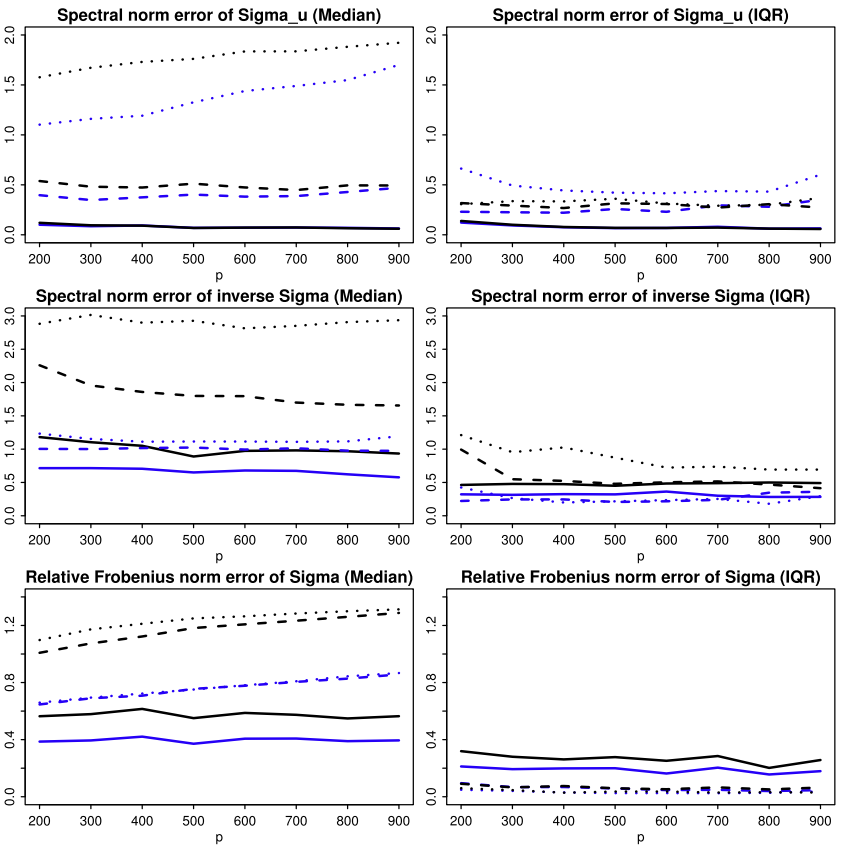

Fig. 4. Errors of robust estimates against varying p. Blue line represents ratio of errors with $\widehat{\Sigma}_z^R$ over errors with $\widehat{\Sigma}_z^S$, while black line represents ratio of errors with $\widehat{\Sigma}_z^K$ over errors with $\widehat{\Sigma}_z^S$. z_t is generated by element-wise iid t-distribution with $df = 3$ (solid), 5 (dashed) and ∞ (dotted). The median errors and their IQR over 100 simulations are reported. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

5. Real data analysis

In this section, we look into financial historical data during 2005–2013, and assess to what extent our factor model characterizes the data.

The dataset we used in our analysis consists of daily returns of 393 stocks, all of which are large market capitalization constituents of S&P 500 index, collected without missing values from 2005 to 2013. This dataset has also been used in Fan et al. (2016), where they investigated how covariates (e.g. size, volume) could be utilized to help estimate factors and factor loadings, whereas the focus of the current paper is to develop robust methods in the presence of heavy-tailed data.

In addition, we collected factors data in the same period, where the factors are calculated according to Fama–French three-factor model (Fama and French, 1993). After centering, the panel matrix we will use for analysis, is a 393 by 2265 matrix Y, in addition to a factor matrix F of size 2265 by 3. Here 2265 is the number of daily returns and 393 is the number of stocks.

5.1. Tail-heaviness

First, we look at how the daily returns are distributed. Especially, we are interested in the tails. In Fig. 5, we made Q–Qplots that compare the distribution of all y_{it} with either Gaussian distribution or t-distributions with varying degree of freedom, ranging from df $= 2$ to df $= 6$. We also fit a line in each plot, showing how much the return data deviate from the base distribution. It is clear that the data tail is heavier than that of a Gaussian distribution, and that t-distribution with df $= 4$ is almost in alignment with the return data. Similarly, we made the Q–Q plots for the factors in Fig. 6. The plots also show that t-distribution is better in terms of fitting the data; however, the tails are even heavier, and t-distribution with df $= 2$ seems to best fit the data.

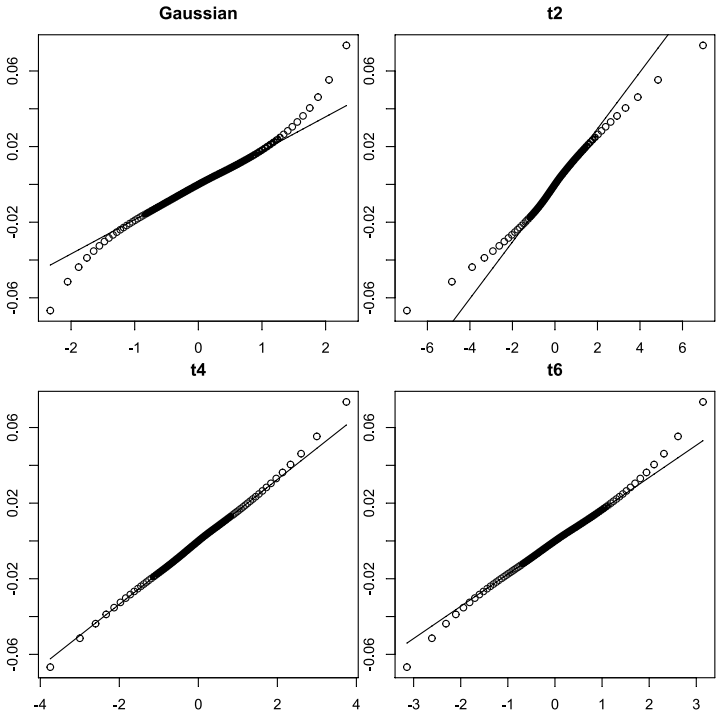

Fig. 5. Q–Q plot of excess returns y_{it} for all i and t against Gaussian distribution and t-distribution with degree of freedom 2, 4 and 6. For each plot, a line is fitted by connecting points at first and third quartiles.

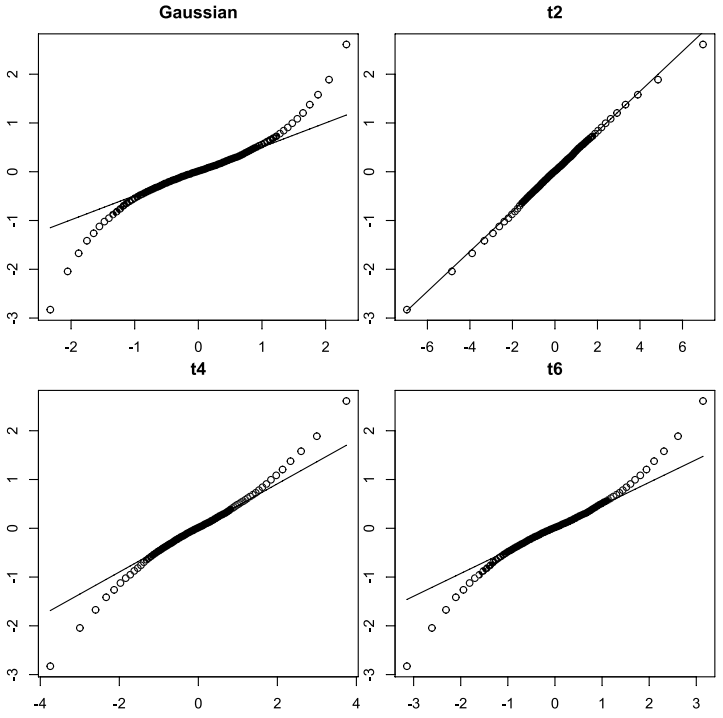

Fig. 6. Q–Q plot of factor f_{it} against Gaussian distribution and t-distribution with degree of freedom 2, 4 and 6. For each plot, a line is fitted by connecting points at first and third quartiles.

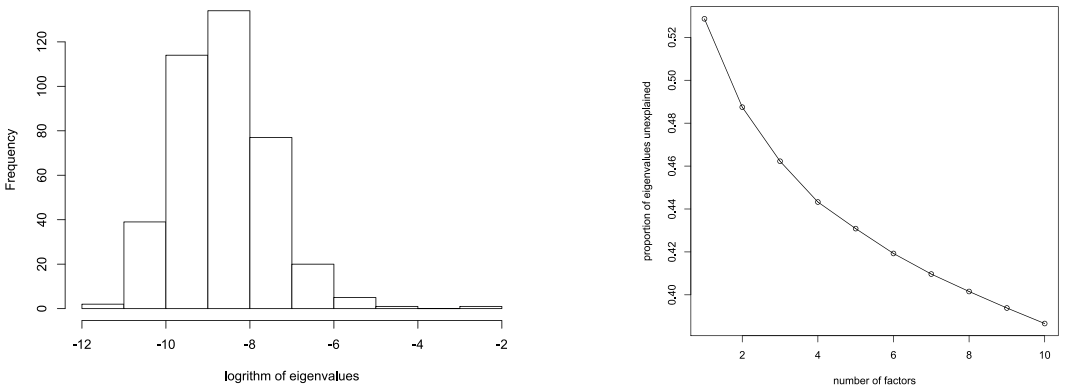

Fig. 7. Left panel: Histogram of eigenvalues of sample covariance matrix YY^T/n. The histogram is plotted on the logarithmic scale, i.e. each bin counts the number of log λ_i in a given range. Right panel: Proportion of residue eigenvalues $\sum_{i=K+1}^{p}\lambda_i/\sum_{i=1}^{p}\lambda_i$, against varying K, where λ_i is the ith largest eigenvalue of sample covariance matrix YY^T/n.

5.2. Spiked covariance structure

We now consider how the covariance matrix of returns looks like, since a spiked covariance structure would justify the pervasiveness assumption. To find the spectral structure, we calculated eigenvalues of the sample covariance matrix YY^T/n, and made a histogram based on logarithmic scale (see the left panel in Fig. 7). In the histogram, the counts in the rightmost four bins are 5, 1, 0 and 1, representing only a few large eigenvalues, which is a strong signal of a spiked structure. We also plotted the proportion of residue eigenvalues $\sum_{i=K+1}^{p}\lambda_i/\sum_{i=1}^{p}\lambda_i$, against K in the right panel of Fig. 7. The top 3 eigenvalues account for a major part of the variances, which supports the pervasive assumption.

The spiked covariance structure has been studied in Paul (2007), Johnstone and Lu (2009) and many other papers, but under their regime, the top eigenvalues or "spiked" eigenvalues do not grow with the dimension p. In this paper, the spiked eigenvalues have stronger signals, and thus are easier to be separated from the rest of eigenvalues. In this respect, the connotation of "spiked covariance structure" is closer to that in Wang and Fan (2017). As empirical evidence, this phenomenon also buttresses the motivation of study in Wang and Fan (2017).

5.3. Portfolio risk estimation

We consider portfolio risk estimation. To be specific, for a portfolio with weight vector $w \in \mathbf{R}^p$ on all the market assets, its risk is measured by quantity $w^T\Sigma w$ where Σ is the true covariance of excess returns of all the assets. Note that Σ is time varying. Here we consider a class of weights with gross exposure $c \geq 1$, that is $\sum_i w_i = 1$ and $\sum_i |w_i| = c$. We consider four scenarios $c = 1, 1.4, 1.8, 2.2$. Note that $(c-1)/2$ represents the level of exposure to short selling; in particular, $c = 1$ represents the case of no short selling.

To assess how well our robust estimator performs compared with the sample covariance, we calculated the covariance estimators $\widehat{\Sigma}_t^R$ and $\widehat{\Sigma}_t^S$, using the daily data of preceding 12 months, where $\widehat{\Sigma}_t^R$ is our robust covariance estimator and $\widehat{\Sigma}_t^S$ is the sample covariance, for every trading day from 2006 to 2013. We indexed those dates by t where t runs from 1 to 2013 (from 2006–01–01 to 2013–12–31, it happens to contain 2013 trading days, so here 2013 is the total number of trading days instead of a year indicator). Let γ_{t+1} be the excess return of the following trading day after t. For a weight vector w, the error we used to gauge the two approaches is

$$R^R(w) = \frac{1}{2013}\sum_{t=1}^{2013}\left|w^T\widehat{\Sigma}_t^R w - (w^T\gamma_{t+1})^2\right|, \quad R^S(w) = \frac{1}{2013}\sum_{t=1}^{2013}\left|w^T\widehat{\Sigma}_t^S w - (w^T\gamma_{t+1})^2\right|.$$

Note the bias–variance decomposition

$$E|w^T\widehat{\Sigma}_t w - (w^T\gamma_{t+1})^2|^2 = E|(w^T\gamma_{t+1})^2 - w^T\Sigma_t w|^2 + E|w^T\widehat{\Sigma}_t w - w^T\Sigma_t w|^2,$$

where $\Sigma_t = E\gamma_{t+1}\gamma_{t+1}^T$. The first term measures the size of the stochastic error that cannot be reduced while the second term is the estimation error for the risk of portfolio w.

To generate multiple random weights w with gross exposure c, we adopted the strategy used in Fan et al. (2015), which aims to generate a uniform distribution on the simplex $\{w : \sum_i w_i = 1, \sum_i |w_i| = c\}$: (1) for each index $i \leq p$ let $\eta_i = 1$ (long) with probability $(c+1)/2c$ and $\eta_i = -1$ (short) with probability $(c-1)/2c$; (2) generate iid ξ_i by exponential distribution; (3) for $\eta_i = 1$, let $w_i = \frac{c+1}{2}\cdot\xi_i/\sum_{\eta_i=1}\xi_i$ and for $\eta_i = -1$, let $w_i = -\frac{c-1}{2}\cdot\xi_i/\sum_{\eta_i=-1}\xi_i$. We made a set of scatter plots in Fig. 8, in which the x-axis represents $R^R(w)$ and the y-axis $R^S(w)$. In addition, we highlighted in the first plot

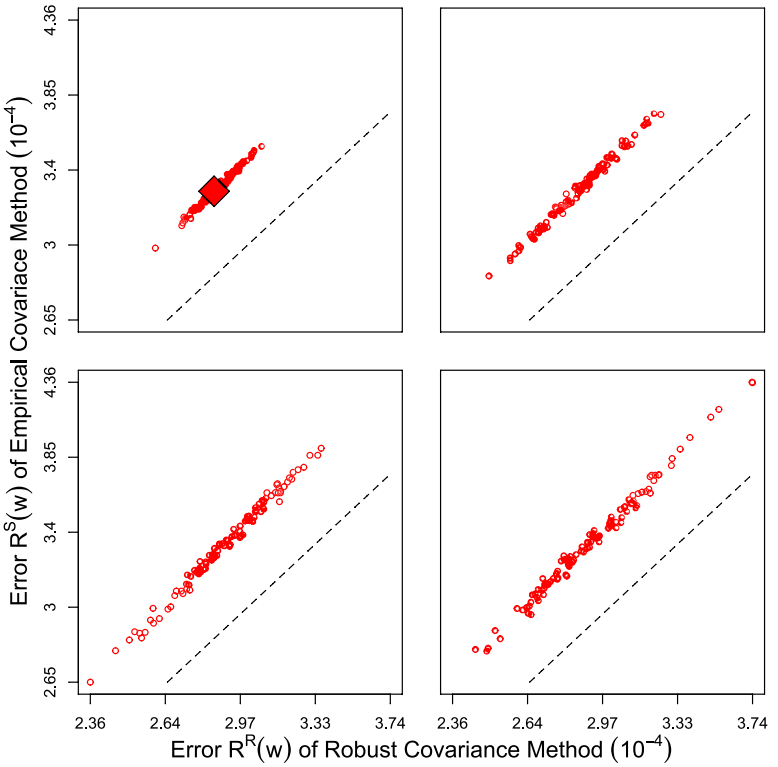

Fig. 8. $(R^R(w), R^S(w))$ for multiple randomly generated w. The four plots compare the errors of the two methods under different settings (upper left: no short selling; upper right: exposure $c = 1.4$; lower left: exposure $c = 1.8$; lower right: exposure $c = 2.2$). The red diamond in the first plot corresponds to uniform weights. The dashed line is the 45 degree line representing equal performance. Our robust method gives smaller errors.

the point with uniform weights (i.e. $w_i = 1/p$), which serves as a benchmark for comparison. The dashed line shows where the two approaches have the same performance. Clearly, for all w the robust approach has smaller risk errors, and therefore has better empirical performance in estimating portfolio risks.

Acknowledgments

This research was partially supported by NSF grants DMS-1206464 and DMS-1406266 and NIH grants R01-GM072611-11 and NIH R01GM100474- 04.

Appendix

Proof of Theorem 3.1. Since we have robust estimator $\widehat{\Sigma}_z$ such that $\|\widehat{\Sigma}_z - \Sigma_z\|_\infty = O_P(\sqrt{\log p/n})$, we clearly know $\widehat{\Sigma}_{11}, \widehat{\Sigma}_{12}, \widehat{\Sigma}_{21}, \widehat{\Sigma}_{22}$ achieve the same rate. Using this, let us first prove $\|\widehat{\Sigma}_u - \Sigma_u\|_\infty = O_P(\sqrt{\log p/n})$. Obviously,

$$\|\widehat{\Sigma}_{12}\widehat{\Sigma}_{22}^{-1}\widehat{\Sigma}_{21}^T - B\Sigma_f B^T\|_\infty = \|\widehat{\Sigma}_{12}\widehat{\Sigma}_{22}^{-1}\widehat{\Sigma}_{21}^T - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}^T\|_\infty = O_P(\sqrt{\log p/n}), \tag{A.1}$$

because the multiplication is along the fixed dimension r and each element is estimated with the rate of convergence $O_P(\sqrt{\log p/n})$. Also $\|\widehat{\Sigma}_{11} - \Sigma\|_\infty = O_P(\sqrt{\log p/n})$, therefore $\widehat{\Sigma}_u = \widehat{\Sigma}_{11} - \widehat{\Sigma}_{12}\widehat{\Sigma}_{22}^{-1}\widehat{\Sigma}_{21}^T$ is good enough to estimate $\Sigma_u = \Sigma - B\Sigma_f B^T$ with error $O_P(\sqrt{\log p/n})$ in max-norm.

Once the max error of sparse matrix Σ_u is controlled, it is not hard to show the adaptive procedure in Step 2 gives $\widehat{\Sigma}_u^{\mathcal{T}}$ such that the spectral error $\|\widehat{\Sigma}_u^{\mathcal{T}} - \Sigma_u\| = O_P(m_p w_n^{1-q})$ (Fan et al., 2011; Cai and Liu, 2011; Rothman et al., 2009) where we define $w_n = \sqrt{\log p/n}$. Furthermore, $\|(\widehat{\Sigma}_u^{\mathcal{T}})^{-1} - \Sigma_u^{-1}\| \le \|(\widehat{\Sigma}_u^{\mathcal{T}})^{-1}\|\|\widehat{\Sigma}_u^{\mathcal{T}} - \Sigma_u\|\|\Sigma_u^{-1}\|$. So $\|(\widehat{\Sigma}_u^{\mathcal{T}})^{-1} - \Sigma_u{}^{-1}\|$ is also $O_P(m_p w_n^{1-q})$ due to the lower boundedness of $\|\Sigma_u\|$. So (3.4) is valid.

Proving (3.5) is trivial. $\|\widehat{\Sigma}_u^{\mathcal{T}} - \Sigma_u\|_\infty \le \|\widehat{\Sigma}_u^{\mathcal{T}} - \widehat{\Sigma}_u\|_\infty + \|\widehat{\Sigma}_u - \Sigma_u\|_\infty = O_P(\tau + w_n) = O_P(w_n)$ when τ is chosen as the same order as w_n and thus

$$\|\widehat{\Sigma}^{\mathcal{T}} - \Sigma\|_\infty \le \|\widehat{\Sigma}_{12}\widehat{\Sigma}_{22}^{-1}\widehat{\Sigma}_{21}^T - B\Sigma_f B^T\|_\infty + \|\widehat{\Sigma}_u^{\mathcal{T}} - \Sigma_u\|_\infty = O_P(w_n).$$

Next let us take a look at the relative Frobenius convergence (3.6) for $\|\widehat{\Sigma}^{\mathcal{T}} - \Sigma\|_\Sigma$.

$$\begin{aligned}\|\widehat{\Sigma}^{\mathcal{T}} - \Sigma\|_\Sigma &\le \|\widehat{\Sigma}_{12}\widehat{\Sigma}_{22}^{-1}\widehat{\Sigma}_{21}^T - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}^T\|_\Sigma + \|\widehat{\Sigma}_u^{\mathcal{T}} - \Sigma_u\|_\Sigma \\ &\le \|(\widehat{\Sigma}_{12} - \Sigma_{12})\widehat{\Sigma}_{22}^{-1}(\widehat{\Sigma}_{21} - \Sigma_{21})^T\|_\Sigma + 2\|(\widehat{\Sigma}_{12} - \Sigma_{12})\widehat{\Sigma}_{22}^{-1}\Sigma_{21}^T\|_\Sigma \\ &\quad + \|\Sigma_{12}(\widehat{\Sigma}_{22}^{-1} - \Sigma_{22}^{-1})\Sigma_{21}^T\|_\Sigma + \|\widehat{\Sigma}_u^{\mathcal{T}} - \Sigma_u\|_\Sigma \\ &=: \Delta_1 + 2\Delta_2 + \Delta_3 + \Delta_4 .\end{aligned} \tag{A.2}$$

We bound the four terms one by one. The last term is the most straightforward,

$$\Delta_4 \le p^{-1/2}\|\Sigma_u^{\mathcal{T}} - \Sigma_u\|_F\|\Sigma^{-1}\| = O_P(\|\Sigma_u^{\mathcal{T}} - \Sigma_u\|) = O_P(m_p w_n^{1-q}).$$

Bound for Δ_1 uses the fact that $\|\widehat{\Sigma}_{22}^{-1}\|$ and $\|\Sigma^{-1}\|$ are $O_P(1)$ and $\|\widehat{\Sigma}_{12} - \Sigma_{12}\|_F = O_P(\sqrt{p\log p/n})$. So

$$\Delta_1 \le p^{-1/2}\|\widehat{\Sigma}_{12} - \Sigma_{12}\|_F^2\|\widehat{\Sigma}_{22}^{-1}\|\|\Sigma^{-1}\| = O_P\Big(\frac{\sqrt{p}\log p}{n}\Big);$$

Bound for Δ_3 needs additional conclusion that $\|\Sigma_{21}^T\Sigma^{-1}\Sigma_{12}\| \le \|B^T\Sigma^{-1}B\|\|\Sigma_{22}\|^2 \le 2\|\Sigma_{22}\| = O(1)$, where $B = \Sigma_{12}\Sigma_{22}^{-1}$ and the last inequality is shown in Fan et al. (2008). So

$$\begin{aligned}\Delta_3 &= p^{-1/2}\mathrm{tr}^{1/2}\Big((\widehat{\Sigma}_{22}^{-1} - \Sigma_{22}^{-1})\Sigma_{21}^T\Sigma^{-1}\Sigma_{12}(\widehat{\Sigma}_{22}^{-1} - \Sigma_{22}^{-1})\Sigma_{21}^T\Sigma^{-1}\Sigma_{12}\Big) \\ &\le p^{-1/2}\|(\widehat{\Sigma}_{22}^{-1} - \Sigma_{22}^{-1})\Sigma_{21}^T\Sigma^{-1}\Sigma_{12}\|_F \le p^{-1/2}\|\widehat{\Sigma}_{22}^{-1} - \Sigma_{22}^{-1}\|_F\|\Sigma_{21}^T\Sigma^{-1}\Sigma_{12}\| \\ &= O_P(\sqrt{\log p/(np)}).\end{aligned}$$

Lastly, by similar trick, we have

$$\begin{aligned}\Delta_2 &= p^{-1/2}\mathrm{tr}^{1/2}\Big((\widehat{\Sigma}_{12} - \Sigma_{12})\widehat{\Sigma}_{22}^{-1}\Sigma_{21}^T\Sigma^{-1}\Sigma_{21}\widehat{\Sigma}_{22}^{-1}(\widehat{\Sigma}_{12} - \Sigma_{12})\Sigma^{-1}\Big) \\ &\le p^{-1/2}\|\widehat{\Sigma}_{12} - \Sigma_{12}\|_F\|\widehat{\Sigma}_{22}^{-1}\|\|\Sigma^{-1}\|^{1/2}\|\Sigma_{21}^T\Sigma^{-1}\Sigma_{12}\|^{1/2} = O_P(\sqrt{\log p/n}).\end{aligned}$$

Combining results above, by (A.2), we conclude that $\|\widehat{\Sigma}^{\mathcal{T}} - \Sigma\|_\Sigma = O_P(\sqrt{p}\log p/n + m_p(\log p/n)^{(1-q)/2})$.

Finally we show the rate of convergence for $\|(\widehat{\Sigma}^{\mathcal{T}})^{-1} - \Sigma^{-1}\|$. By Woodbury formula,

$$\Sigma^{-1} = \Sigma_u^{-1} - \Sigma_u^{-1}\Sigma_{12}[\Sigma_{22} + \Sigma_{12}^T\Sigma_u^{-1}\Sigma_{21}]^{-1}\Sigma_{21}^T\Sigma_u^{-1}.$$

Thus, let $A = \Sigma_{22} + \Sigma_{12}^T\Sigma_u^{-1}\Sigma_{21}$, $\widehat{A} = \widehat{\Sigma}_{22} + \widehat{\Sigma}_{12}^T(\widehat{\Sigma}_u^{\mathcal{T}})^{-1}\widehat{\Sigma}_{21}$ and $D = \Sigma_u^{-1}\Sigma_{12}$, $\widehat{D} = (\widehat{\Sigma}_u^{\mathcal{T}})^{-1}\widehat{\Sigma}_{12}$, we have the following bound similar to (A.2):

$$\begin{aligned}\|(\widehat{\Sigma}^{\mathcal{T}})^{-1} - \Sigma^{-1}\| &\le \|\widehat{D}\widehat{A}^{-1}\widehat{D}^T - DA^{-1}D^T\| + \|(\widehat{\Sigma}_u^{\mathcal{T}})^{-1} - \Sigma_u^{-1}\| \\ &\le \|(\widehat{D} - D)\widehat{A}^{-1}(\widehat{D} - D)^T\| + 2\|(\widehat{D} - D)\widehat{A}^{-1}D^T\| \\ &\quad + \|D(\widehat{A}^{-1} - A^{-1})D^T\| + \|(\widehat{\Sigma}_u^{\mathcal{T}})^{-1} - \Sigma_u^{-1}\| \\ &=: \widetilde{\Delta}_1 + 2\widetilde{\Delta}_2 + \widetilde{\Delta}_3 + \widetilde{\Delta}_4 .\end{aligned} \tag{A.3}$$

From (3.4), $\widetilde{\Delta}_4 = O_P(m_p\omega_n^{1-q})$. For the remaining terms, we need to find the rates for $\|\widehat{D} - D\|$, $\|\widehat{A}^{-1}\|$, $\|D\|$ and $\|\widehat{A}^{-1} - A^{-1}\|$ separately. Note that $\|\Sigma_{12}\| = \|B\Sigma_{22}\| \le \|B\|\|\Sigma_{22}\| = O_P(\sqrt{p})$ by Assumption 2.2(ii). So $\|D\| = O_P(\sqrt{p})$ and

$$\|\widehat{D} - D\| \le \|(\widehat{\Sigma}_u^{\mathcal{T}})^{-1}\|\|\widehat{\Sigma}_{12} - \Sigma_{12}\| + \|\Sigma_{12}\|\|(\widehat{\Sigma}_u^{\mathcal{T}})^{-1} - \Sigma_u^{-1}\| = O_P(\sqrt{p}m_p\omega_n^{1-q}).$$

In addition, it is not hard to show $\|\widehat{A} - A\| = O_P(pm_p\omega_n^{1-q})$. In addition, we claim $\|A^{-1}\| = O_P(p^{-1})$ since $\lambda_{\min}(A) \ge \lambda_{\min}(\Sigma_{12}^T\Sigma_u^{-1}\Sigma_{21}) \ge \lambda_{\min}(\Sigma_u^{-1})\lambda_{\min}(\Sigma_f)\lambda_r(B\Sigma_f B^T)$ and by Weyl's inequality, $\lambda_r(B\Sigma_f B^T) \ge \lambda_r(\Sigma) - \|\Sigma\| \ge cp$ by Assumption 2.2(i). Therefore, $\|\widehat{A}^{-1} - A^{-1}\| \le \|A^{-1}\|\|\widehat{A}^{-1}\|\|\widehat{A} - A\|$ implies $\|\widehat{A}^{-1} - A^{-1}\| = O_P(p^{-1}m_p\omega_n^{1-q})$, and furthermore $\|\widehat{A}^{-1}\| = O_P(p^{-1})$. Finally we incorporate the above rates together and conclude

$$\begin{aligned}\widetilde{\Delta}_1 &= O_P(p^{-1}\|\widehat{D} - D\|^2) = O_P(m_p^2\omega_n^{2(1-q)}), \\ \widetilde{\Delta}_2 &= O_P(p^{-1/2}\|\widehat{D} - D\|) = O_P(m_p\omega_n^{1-q}), \\ \widetilde{\Delta}_3 &= O_P(p\|\widehat{A}^{-1} - A^{-1}\|) = O_P(m_p\omega_n^{1-q}).\end{aligned}$$

So combining rates for $\widetilde{\Delta}_i$, $i = 1, 2, 3, 4$, we show (3.7) is true. The proof is now complete. □

Proof of Theorem 3.2. Without loss of generality we can assume $\mu = 0$. By dominated converge theorem we know that for all t, $\lim_n \lambda_n(t) = -t$, that $\lambda_n(t)$ is differentiable, that $\lambda_n'(t) = -E\psi_n'(X - t)$, and that $\lim_n \lambda_n'(t) = -1$. With Taylor's expansion, we have

$$\lambda_n(t) = \lambda_n(0) + \lambda_n'(0)t + \Delta_n(t), \tag{A.4}$$

where $|\Delta_n(t)| \le |t| \sup\{|\lambda_n'(s) - \lambda_n'(0)| : 0 \le s \le t\}$. Observe that

$$\begin{aligned}\big|\lambda_n'(s) - \lambda_n'(0)\big| &= \big|P(|X - s| \le \alpha_n) - P(|X| \le \alpha_n)\big| \\ &\le P(|X - s| > \alpha_n) + P(|X| > \alpha_n).\end{aligned}$$

By Markov's inequality,

$$\sup\{|\lambda_n'(s) - \lambda_n'(0)| : 0 \le s \le t\} \le \frac{1}{\alpha_n}\big(2E|X| + |t|\big).$$

For any $\epsilon \in (0, 1)$, there exists $N > 0$, such that for all $n > N$,

$$|\lambda_n(0)| \le 2, \quad \frac{1+\epsilon/2}{1+\epsilon} \le -\lambda_n'(0) \le \frac{1-\epsilon/2}{1-\epsilon}, \quad \frac{1}{\alpha_n}(2E|X| + 4) \le \frac{\epsilon}{4(1+\epsilon)}.$$

Plugging $t = (1+\epsilon)\lambda_n(0)$ into (A.4),

$$\lambda_n((1+\epsilon)\lambda_n(0)) = \lambda_n(0) + \lambda_n'(0)(1+\epsilon)\lambda_n(0) + \Delta_n(t),$$

where $|\Delta_n(t)| \le (1+\epsilon)|\lambda_n(0)|\frac{\epsilon}{4(1+\epsilon)} = \epsilon|\lambda_n(0)|/4$. Equivalently,

$$\lambda_n((1+\epsilon)\lambda_n(0)) = \lambda_n(0)(1 + \lambda_n'(0)(1+\epsilon) + \beta_n),$$

where $|\beta_n| \le \epsilon/4$. Similarly,

$$\lambda_n((1-\epsilon)\lambda_n(0)) = \lambda_n(0)(1 + \lambda_n'(0)(1-\epsilon) + \beta_n'),$$

where $|\beta_n'| \le \epsilon/4$. Also we have $1 + \lambda_n'(0)(1+\epsilon) + \beta_n < 0$ and $1 + \lambda_n'(0)(1-\epsilon) + \beta_n' > 0$. Multiplying both sides of the equations, we deduce that

$$\lambda_n((1+\epsilon)\lambda_n(0)) \cdot \lambda_n((1-\epsilon)\lambda_n(0)) \le 0.$$

If $\lambda_n(0) = 0$, equation $\lambda_n(t) = 0$ has one zero $t = 0$; and in fact it is the unique one for sufficiently large n, since $\lambda_n(t)$ is nonincreasing and $\lambda_n'(0) \ne 0$ for n large enough. If $\lambda_n(0) \ne 0$, at least one zero lies in the interval with endpoints $(1+\epsilon)\lambda_n(0)$ and $(1-\epsilon)\lambda_n(0)$. Since $\lambda_n(0) \to 0$, for any zero t_n' in this interval we have $t_n' \to 0$, which implies $\lambda_n'(t_n') \to -1$. It follows that such zero is unique for sufficiently large n. This leads to $t_n/\lambda_n(0) \to 1$, thus proving the second claim in the theorem.

The proof of the first claim is similar in spirit to that of Huber (1964). Let us denote

$$T_n^- = \sup\{t : \sum_{i=1}^n \psi_n(x_i - t) > 0\},$$

$$T_n^+ = \inf\{t : \sum_{i=1}^n \psi_n(x_i - t) < 0\}.$$

By monotonicity, $T_n \in [T_n^-, T_n^+]$. Since

$$P(T_n^- < t) = P\Big(\sum_{i=1}^n \psi_n(x_i - t) \le 0\Big),$$

it follows that for any fixed $z \in \mathbf{R}$,

$$\begin{aligned}P(\sqrt{n}\,(T_n^- - t_n) < z) &= P(T_n^- < t_n + z/\sqrt{n}) \\ &= P(\sum_{i=1}^n \psi_n(x_i - u_n) \le 0) \\ &= P\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\psi_n(x_i - u_n) - \lambda_n(u_n)}{\sigma_n(u_n)} \le -\frac{\sqrt{n}\,\lambda_n(u_n)}{\sigma_n(u_n)}\Big),\end{aligned}$$

where we denote $u_n = t_n + z/\sqrt{n}$ and $\sigma_n(u) = E\psi_n(X - u)^2 - \lambda_n(u)^2$.

By dominate convergence theorem, $\lambda_n'(t_n) \to -1$ and $\sigma_n(u_n)^2 \to \sigma^2$. By Taylor expansion of $\lambda_n(u_n)$ at t_n,

$$\lambda_n(u_n) = \lambda_n(t_n) + z/\sqrt{n}\,\lambda_n'(t_n) + \Delta_z^n,$$

where $|\Delta_z^n| \le n^{-1/2}|z| \sup\{\lambda_n'(t_n + s) - \lambda_n'(t_n)|: 0 \le s \le z/\sqrt{n}\}$. A similar argument shows that

$$\sup\{\lambda_n'(t_n + s) - \lambda_n'(t_n)|: 0 \le s \le z/\sqrt{n}\} \le \frac{1}{\alpha_n}(2E(X) + 2|t_n| + |z|/\sqrt{n}) = o(1).$$

This leads to $\lambda_n(u_n) = z/\sqrt{n}\,(\lambda_n'(t_n) + o(1)) = z/\sqrt{n}\,(-1 + o(1))$, and thus $\sqrt{n}\,\lambda_n(u_n) \to -z$.

Let us write

$$\xi_i = \frac{\psi_n(x_i - u_n) - \lambda_n(u_n)}{\sigma_n(u_n)}$$

for the centered variance ξ_i with unit variance. If we can show

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \xrightarrow{d} N(0, 1), \tag{A.5}$$

then by continuity of Φ, standard normal distribution function, we have

$$P\Big(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \le -\frac{\sqrt{n}\,\lambda_n(u_n)}{\sigma_n(u_n)}\Big) \to \Phi\big(\frac{z}{\sigma}\big),$$

which gives $P(\sqrt{n}\,(T_n^- - t_n) < z) \to \Phi(z/\sigma)$. It is similar to show that $P(\sqrt{n}\,(T_n^+ - t_n) < z) \to \Phi(z/\sigma)$. At this point, we are able to conclude that the first claim in the theorem holds, i.e. $\sqrt{n}\,(T_n - t_n) \xrightarrow{d} N(0, \sigma^2)$.

To prove (A.5), it suffices to check Lindeberg's condition:

$$E(\xi_i^2 \mathbf{1}\{|\xi_i| > \sqrt{n}\,\epsilon\}) \to 0$$

for any $\epsilon > 0$. Notice that $\lambda_n(u_n) \to 0$ and $\sigma_n(u_n) \to \sigma$, we only need to show

$$E(\psi_n^2(X - u_n)\mathbf{1}\{|\psi_n(X - u_n)| > \sqrt{n}\,\epsilon\}) \to 0.$$

This is true due to

$$\psi_n^2(X - u_n) \le |X - u_n|^2 \le 2|X|^2 + 2u_n^2$$

and dominated convergence theorem. □

References

Agarwal, A., Negahban, S., Wainwright, M.J., 2012. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. Ann. Statist. 40 (2), 1171–1197.

Amini, A.A., Wainwright, M.J., 2008. High-dimensional analysis of semidefinite relaxations for sparse principal components. In: Information Theory, 2008. ISIT 2008. IEEE International Symposium on. IEEE, pp. 2454–2458.

Antoniadis, A., Fan, J., 2001. Regularization of wavelet approximations. J. Amer. Statist. Assoc. 96 (455).

Bai, J., 2003. Inferential theory for factor models of large dimensions. Econometrica 71, 135–171.

Berthet, Q., Rigollet, P., 2013. Optimal detection of sparse principal components in high dimension. Ann. Statist. 41 (4), 1780–1815.

Bickel, P.J., 1976. Another look at robustness: A review of reviews and some new developments. Scand. J. Stat. 145–168.

Bickel, P.J., Levina, E., 2008. Covariance regularization by thresholding. Ann. Statist. 2577–2604.

Birnbaum, A., Johnstone, I.M., Nadler, B., Paul, D., 2013. Minimax bounds for sparse PCA with noisy high-dimensional data. Ann. Statist. 41 (3), 1055.

Cai, T., Liu, W., 2011. Adaptive thresholding for sparse covariance matrix estimation. J. Amer. Statist. Assoc. 106 (494), 672–684.

Cai, T., Ma, Z., Wu, Y., 2013. Optimal estimation and rank detection for sparse spiked covariance matrices. Probab. Theory Related Fields 161 (3–4), 781–815.

Cai, T.T., Zhang, C.-H., Zhou, H.H., 2010. Optimal rates of convergence for covariance matrix estimation. Ann. Statist. 38 (4), 2118–2144.

Candès, E.J., Li, X., Ma, Y., Wright, J., 2011. Robust principal component analysis? J. ACM 58 (3), 11.

Candès, E.J., Recht, B., 2009. Exact matrix completion via convex optimization. Found. Comput. Math. 9 (6), 717–772.

Catoni, O., 2012. Challenging the empirical mean and empirical variance: A deviation study. In: Annales de l'Institut Henri Poincaré, Probabilités et Statistiques, vol. 48. Institut Henri Poincaré, pp. 1148–1185, 4.

Chandrasekaran, V., Sanghavi, S., Parrilo, P.A., Willsky, A.S., 2011. Rank-sparsity incoherence for matrix decomposition. SIAM J. Optim. 21 (2), 572–596.

Dinov, I.D., Boscardin, J.W., Mega, M.S., Sowell, E.L., Toga, A.W., 2005. A wavelet-based statistical analysis of fMRI data. Neuroinformatics 3 (4), 319–342.

Fama, E.F., French, K.R., 1993. Common risk factors in the returns on stocks and bonds. J. Financ. Econ. 33 (1), 3–56.

Fan, J., Fan, Y., Lv, J., 2008. High dimensional covariance matrix estimation using a factor model. J. Econometrics 147 (1), 186–197.

Fan, J., Li, Q., Wang, Y., 2017. Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. J. R. Stat. Soc. Ser. B Stat. Methodol. 79 (1), 247–265.

Fan, J., Liao, Y., Mincheva, M., 2011. High dimensional covariance matrix estimation in approximate factor models. Ann. Statist. 39 (6), 3320.

Fan, J., Liao, Y., Mincheva, M., 2013. Large covariance estimation by thresholding principal orthogonal complements. J. R. Stat. Soc.: Ser. B 75, 1–44.

Fan, J., Liao, Y., Shi, X., 2015. Risks of large portfolios. J. Econometrics 186, 367–387.

Fan, J., Liao, Y., Wang, W., 2016. Projected principal component analysis in factor models. Ann. Statist. 44 (1), 219.

Fan, J., Liu, H., Wang, W., 2018. Large covariance estimation through elliptical factor models. Ann. Statist. 46 (4), 1383–1414.

Fang, K.-T., Kotz, S., Ng, K.W., 1990. Symmetric Multivariate and Related Distributions. Chapman and Hall.

Forni, M., Hallin, M., Lippi, M., Reichlin, L., 2000. The generalized dynamic-factor model: Identification and estimation. Rev. Econ. Stat. 82 (4), 540–554.

Forni, M., Lippi, M., 2001. The generalized dynamic factor model: Representation theory. Econometric Theory 17 (06), 1113–1141.

Friedman, J., Hastie, T., Tibshirani, R., 2008. Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9 (3), 432–441.

Grant, M., Boyd, S., Ye, Y., 2008. CVX: Matlab software for disciplined convex programming.

Han, F., Liu, H., 2014. Scale-invariant sparse PCA on high-dimensional meta-elliptical data. J. Amer. Statist. Assoc. 109 (505), 275–287.

Hsu, D., Sabato, S., 2014. Heavy-tailed regression with a generalized median-of-means. In: Proceedings of the 31st International Conference on Machine Learning, ICML-14, pp. 37–45.

Huber, P.J., 1964. Robust estimation of a location parameter. Ann. Math. Stat. 35 (1), 73–101.

Johnstone, I.M., Lu, A.Y., 2009. On consistency and sparsity for principal components analysis in high dimensions. J. Amer. Statist. Assoc. 104 (486), 682–693.

Lam, C., Fan, J., 2009. Sparsistency and rates of convergence in large covariance matrix estimation. Ann. Statist. 37 (6B), 4254.

Lepskii, O., 1992. Asymptotically minimax adaptive estimation. I: Upper bounds. Optimally adaptive estimates. Theory Probab. Appl. 36 (4), 682–697.

Lintner, J., 1965. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. Rev. Econ. Stat. 13–37.
Ma, Z., 2013. Sparse principal component analysis and iterative thresholding. Ann. Statist. 41 (2), 772–801.
Meinshausen, N., Bühlmann, P., 2006. High-dimensional graphs and variable selection with the lasso. Ann. Statist. 1436–1462.
Onatski, A., 2012. Asymptotics of the principal components estimator of large factor models with weakly influential factors. J. Econometrics 168 (2), 244–258.
Paul, D., 2007. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. Statist. Sinica 17 (4), 1617–1642.
Ravikumar, P., Wainwright, M.J., Raskutti, G., Yu, B., et al., 2011. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. Electron. J. Stat. 5, 935–980.
Rothman, A.J., Levina, E., Zhu, J., 2009. Generalized thresholding of large covariance matrices. J. Amer. Statist. Assoc. 104 (485), 177–186.
Sharpe, W.F., 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. J. Financ. 19 (3), 425–442.
Stock, J., Watson, M., 2002. Forecasting using principal components from a large number of predictors. J. Acoust. Soc. Am. 97, 1167–1179.
Vu, V.Q., Lei, J., 2013. Minimax sparse principal subspace estimation in high dimensions. Ann. Statist. 41 (6), 2905–2947.
Wang, W., Fan, J., 2017. Asymptotics of empirical eigen-structure for high dimensional spiked covariance. Ann. Statist. 45 (3), 1342–1374.
Xu, H., Caramanis, C., Sanghavi, S., 2010. Robust PCA via outlier pursuit. In: Advances in Neural Information Processing Systems, pp. 2496–2504.