The Econometrics Journal

Econometrics Journal (2016), volume **19**, pp. C1–C32.
doi: 10.1111/ectj.12061

An overview of the estimation of large covariance and precision matrices

JIANQING FAN[†], YUAN LIAO[‡] AND HAN LIU[†]

[†]*Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08540, USA*.
E-mail: jqfan@princeton.edu, hanliu@princeton.edu

[‡]*Department of Mathematics, University of Maryland, College Park, MD 20742, USA*.
E-mail: yuanliao@umd.edu

First version received: April 2015; final version accepted: February 2016

Summary The estimation of large covariance and precision matrices is fundamental in modern multivariate analysis. However, problems arise from the statistical analysis of large panel economic and financial data. The covariance matrix reveals marginal correlations between variables, while the precision matrix encodes conditional correlations between pairs of variables given the remaining variables. In this paper, we provide a selective review of several recent developments on the estimation of large covariance and precision matrices. We focus on two general approaches: a rank-based method and a factor-model-based method. Theories and applications of both approaches are presented. These methods are expected to be widely applicable to the analysis of economic and financial data.

Keywords: *Approximate factor model, Elliptical distribution, Graphical model, Heavy-tailed, High-dimensionality, Low-rank matrix, Principal components, Rank-based methods, Sparse matrix, Thresholding.*

1. INTRODUCTION

The estimation of large covariance and precision (inverse covariance) matrices underlies fundamental problems in modern multivariate analysis, which has applications in many fields, ranging from economics and finance to biology, social networks and health sciences (Fan et al., 2014). When the dimension of the covariance matrix is large, the estimation problem is generally challenging. It is well known that the sample covariance based on the observed data is singular when the dimension is larger than the sample size. In addition, the aggregation of a massive amount of estimation errors can lead to considerable adverse impacts on the estimation accuracy. Therefore, the estimation of large covariance and precision matrices has seen a rapid growth in research attention over the past decade.

In recent years, researchers have proposed various regularization techniques to consistently estimate large covariance and precision matrices. To estimate large covariance matrices, one of the key assumptions made in the literature is that the target matrix of interest is sparse (i.e., many entries are either zero or nearly so); see Bickel and Levina (2008), Lam and Fan (2009), El Karoui (2010) and Rigollet and Tsybakov (2012). To estimate large precision matrices, it is

often the case that the precision matrix is sparse. A commonly used method for estimating the sparse precision matrix is to employ an ℓ_1-penalized maximum likelihood; see, e.g., Banerjee et al. (2008), Yuan and Lin (2007), Friedman et al. (2008) and Rothman et al. (2008). To further reduce the estimation bias, Lam and Fan (2009) and Shen et al. (2012) proposed non-convex penalties for sparse precision matrix estimation and studied their theoretical properties. For more theory on penalized likelihood methods, see Fan and Li (2001), Fan and Peng (2004), Zou (2006), Zhao and Yu (2006), Bickel et al. (2009) and Wainwright (2009).

The literature has been further expanded into robust estimation based on regularized rank-based approaches; see, e.g., Liu et al. (2012) and Xue and Zou (2012). The rank-based method is particularly appealing when the distribution of the data-generating process is non-Gaussian and heavy-tailed. The literature includes, for instance, Han and Liu (2013), Wegkamp and Zhao (2013) and Mitra and Zhang (2014), etc. The heavy-tailed data are often modelled by the elliptical distribution family, which has been widely used for financial data analysis. See Owen and Rabinovitch (1983), Hamada and Valdez (2008) and Frahm and Jaekel (2008).

In addition, in many applications, the sparsity property is not directly applicable. For example, financial returns depend on common equity market risks, housing prices depend on the general level of economic health, gene expressions can be stimulated by cytokines, among others. Because of the presence of common factors, it is unrealistic to assume that many outcomes are uncorrelated. A natural extension is the conditional sparsity (i.e., conditional on the common factors, the covariance matrix of the remaining components of the outcome variables is sparse). For this, we use a factor model.

The factor model is one of the most useful tools for understanding the common dependence among multivariate outputs, which has broad applications in the statistics and econometrics literature. For instance, it is commonly used to measure the vector of economic outputs or the excess returns of financial assets over time, and has been found to produce good out-of-sample forecasts for macroeconomic variables; see, e.g., Boivin and Ng (2005) and Stock and Watson (2002). In high dimensions, the unknown factors and loadings are typically estimated by the principal components method, and the separations between the common factors and idiosyncratic components are characterized via pervasiveness assumptions; see, e.g., Stock and Watson (2002), Bai (2003), Bai and Ng (2002), Fan et al. (2008, 2013), Breitung and Tenhofen (2011), Onatski (2012) and Lam and Yao (2012), among others. In the statistical literature, the separations between the common factors and idiosyncratic components are carried out by the low-rank plus sparsity decomposition; see, e.g., Candès and Recht (2009), Koltchinskii et al. (2011), Fan et al. (2011), Negahban and Wainwright (2011), Cai et al. (2013) and Ma (2013).

In this paper, we provide a selective review of several recent developments in the estimation of large covariance and precision matrices. We focus on two general approaches: a rank-based method and a factor-model-based method. Theories and applications of both approaches are presented. Note that this paper is not an exhaustive survey, and many other regularization methods are also commonly used in the literature, such as the shrinkage method; see, e.g., Ledoit and Wolf (2003, 2004). We refer to Fan and Liu (2013) and Pourahmadi (2013), and references therein, for reviews of other commonly used methods. Furthermore, another active area in financial econometrics is high-frequency based covariance matrix estimation, where both the sampling frequency and the dimension of the covariance matrix are large. While we do not discuss this approach in detail, the reader is referred to recent developments by Wang and Zou (2010), who directly estimated sparse volatility matrices, and by Aıt-Sahalia and Xiu (2015),

who considered a factor-based approach along the lines of Fan et al. (2013) for high-frequency data.

This paper is organized as follows. In Section 2, we present methods of estimating sparse covariance matrices. In Section 3, we review methods of estimating sparse precision matrices. In Section 4, we discuss robust sparse covariance and precision matrix estimations using rank-based estimators. In Sections 5 and 6, we present the factor-model-based method in the cases of observable and unobservable factors, respectively. In Section 7, we introduce the structured factor model. Finally, in Section 8, we provide a further discussion.

Let $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ denote the minimum and maximum eigenvalues of $\mathbf{A}$, respectively. Let $\psi_{\max}(\mathbf{A})$ be the largest singular value of $\mathbf{A}$. We use $\|\mathbf{A}\|_2$ and $\|\mathbf{A}\|_{\mathrm{F}}$ to denote the operator norm and Frobenius norm of a matrix $\mathbf{A}$, respectively defined as $\lambda_{\max}^{1/2}(\mathbf{A}'\mathbf{A})$ and $\mathrm{tr}^{1/2}(\mathbf{A}'\mathbf{A})$. Throughout this paper, we use p and T to denote the dimension of the covariance matrix of interest, and the sample size, respectively. Let $\mathbf{v} = (v_1, \ldots, v_p)' \in \mathbb{R}^p$ be a real valued vector; we define the vector norms: $\|\mathbf{v}\|_1 = \sum_{j=1}^p |v_j|$, $\|\mathbf{v}\|_2^2 = \sum_{j=1}^p v_j^2$ and $\|\mathbf{v}\|_\infty = \max_{1\le j\le p} |v_j|$. Let $\mathcal{S}$ be a subspace of $\mathbb{R}^p$; we use $\mathbf{v}_{\mathcal{S}}$ to denote the projection of $\mathbf{v}$ on to $\mathcal{S}$: $\mathbf{v}_{\mathcal{S}} = \mathrm{argmin}_{\mathbf{u}\in\mathcal{S}} \|\mathbf{u} - \mathbf{v}\|_2^2$. We also define the orthogonal complement of $\mathcal{S}$ as $\mathcal{S}^\perp = \{\mathbf{u} \in \mathbb{R}^p \mid \mathbf{u}'\mathbf{v} = 0, \text{for any } \mathbf{v} \in \mathcal{S}\}$. Let $\mathbf{A} \in \mathbb{R}^{p\times p}$ and $I, J \subset \{1, \ldots, N\}$ be two sets. Denote by $\mathbf{A}_{I,J}$ the submatrix of $\mathbf{A}$ with rows and columns indexed by I and J. Letting $\mathbf{A}_{*j} = (\mathbf{A}_{1j}, \ldots, \mathbf{A}_{pj})'$ and $\mathbf{A}_{k*} = (\mathbf{A}_{k1}, \ldots, \mathbf{A}_{kp})'$ denote the jth column and kth row of $\mathbf{A}$ in vector forms, we define the matrix norms: $\|\mathbf{A}\|_1 = \max_j \|\mathbf{A}_{*j}\|_1$, $\|\mathbf{A}\|_\infty = \max_k \|\mathbf{A}_{k*}\|_1$ and $\|\mathbf{A}\|_{\max} = \max_j \|\mathbf{A}_{*j}\|_\infty$. We also define matrix element-wise (pseudo-) norms: $\|\mathbf{A}\|_{1,\mathrm{off}} = \sum_{j\neq k} |\mathbf{A}_{jk}|$ and $\|\mathbf{A}\|_{\infty,\mathrm{off}} = \max_{j\neq k} |\mathbf{A}_{jk}|$. We write $a_n \asymp b_n$ if there are positive constants c_1 and c_2 independent of n such that $c_1 b_n \le a_n \le c_2 b_n$.

2. ESTIMATING SPARSE COVARIANCE MATRICES

Let Y_{it} be the observed data for the ith ($i = 1, \ldots, p$) individual at time $t = 1, \ldots, T$ (or the tth observation for the ith variable). We are interested in estimating the $p \times p$ covariance matrix $\Sigma = (\sigma_{ij})_{p\times p}$ of $\boldsymbol{Y}_t = (Y_{1t}, \ldots, Y_{pt})'$, assumed to be independent of t. The sample covariance matrix is defined as

$$\mathbf{S} = \frac{1}{T-1}\sum_{t=1}^{T}(\boldsymbol{Y}_t - \bar{\boldsymbol{Y}})(\boldsymbol{Y}_t - \bar{\boldsymbol{Y}})', \quad \bar{\boldsymbol{Y}} = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{Y}_t.$$

When $p > T$, however, it is well known that $\mathbf{S}$ is singular. It also accumulates many estimation errors due to the large number of free parameters to estimate.

Sparsity is one of the most essential assumptions for high-dimensional covariance matrix estimation, which assumes that a majority of the off-diagonal elements are nearly zero, and effectively reduces the number of free parameters to estimate. Specifically, it assumes that there is $q \ge 0$, so that the following defined quantity

$$m_p = \begin{cases} \max_{i\le p} \sum_{j=1}^p 1\{\sigma_{ij} \neq 0\}, & \text{if } q = 0, \\ \max_{i\le p} \sum_{j=1}^p |\sigma_{ij}|^q, & \text{if } 0 < q < 1, \end{cases} \tag{2.1}$$

is either bounded or grows slowly as $p \to \infty$. Here, $1\{\cdot\}$ denotes the indicator function. Such an assumption is reasonable in many applications. For instance, in a longitudinal study where

variables have a natural order, variables are likely weakly correlated when they are far apart (Wu and Pourahmadi, 2003). Under the sparsity assumption, many regularization-based estimation methods have been proposed. This section selectively overviews several state-of-the-art statistical methods for estimating large sparse covariance matrices.

2.1. Thresholding estimation

One of the most convenient methods to estimate sparse covariance matrices is thresholding, which sets small estimated elements to zero (Bickel and Levina, 2008). Let s_{ij} be the (i, j)th element of the sample covariance matrix $\mathbf{S}$. For a pre-specified thresholding value ω_T, define

$$\widehat{\mathbf{\Sigma}} = (\widehat{\sigma}_{ij})_{p\times p}, \quad \widehat{\sigma}_{ij} = \begin{cases} s_{ij}, & \text{if } i = j, \\ s_{ij}1\{|s_{ij}| > \omega_T\}, & \text{if } i \neq j. \end{cases} \tag{2.2}$$

The thresholding value should dominate the maximum estimation error $\max_{i\neq j} |s_{ij} - \sigma_{ij}|$. When the data are either Gaussian or sub-Gaussian, it can be taken as

$$\omega_T = C\sqrt{\frac{\log p}{T}}, \quad \text{for some } C > 0,$$

so that the probability of the exception event $\{\max_{i\neq j} |s_{ij} - \sigma_{ij}| > \omega_T\}$ tends to zero very fast.[1]

The advantage of thresholding is that it avoids estimating small elements so that noise does not accumulate. The decision about whether an element should be estimated is much easier than the attempt to estimate it accurately. Indeed, under some regularity conditions, Bickel and Levina (2008) showed that, if $m_p\omega_T^{1-q} \to 0$ as $p, T \to \infty$, we have

$$\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_2 = O_P(m_p\omega_T^{1-q}) \qquad \text{and} \qquad \|\widehat{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}^{-1}\|_2 = O_P(m_p\omega_T^{1-q}), \tag{2.3}$$

where m_p and q are as defined in (2.1). In the case that all the 'small' elements of $\mathbf{\Sigma}$ are exactly zero so that we take $q = 0$, the above convergence rate becomes $O_P(\sqrt{\log p/T})$ if m_p is bounded. Because each element in the covariance matrix can be estimated with an error of order $O_P(T^{-1/2})$, it only costs us a $\log(p)$ factor to learn the unknown locations of the non-zero elements. The optimality of such a thresholding procedure has been studied by Cai and Zhou (2012).

2.2. Adaptive thresholding and entry-dependent thresholding

The simple thresholding (2.2) does not take the varying scales of the marginal standard deviations into account. One way to account for this is to threshold on the t-type statistics. For example, using the simple thresholding, we can define the adaptive thresholding estimator (Cai and Liu, 2011):

$$\widehat{\mathbf{\Sigma}} = (\widehat{\sigma}_{ij})_{p\times p}, \quad \widehat{\sigma}_{ij} = \begin{cases} s_{ij}, & \text{if } i = j, \\ s_{ij}1\{|s_{ij}|/\mathrm{SE}(s_{ij}) > \omega_T\}, & \text{if } i \neq j, \end{cases} \tag{2.4}$$

where $\mathrm{SE}(s_{ij})$ is the estimated standard error of s_{ij}. As elaborated by Cai and Liu (2011), the above-defined thresholding estimator is adaptive to the tail distributions of the data, and allows for possibly heavy-tailed distributions.

[1] A random variable X is called sub-Gaussian if there are $c, C > 0$ so that $P(|X| > t) \leq C\exp(-ct^2)$ for all $t > 0$.

A simpler method to take the scale into account is to directly apply thresholding on the correlation matrix. Let $\mathbf{R} = \text{diag}(\mathbf{S})^{-1/2}\mathbf{S}\,\text{diag}(\mathbf{S})^{-1/2} = (r_{ij})_{p\times p}$ be the sample correlation matrix. We then apply the simple thresholding on the off-diagonal elements of $\mathbf{R}$, and obtain the thresholded correlation matrix $\mathbf{R}^{\mathcal{T}}$. So, the (i, j)th element of $\mathbf{R}^{\mathcal{T}}$ is $r_{ij}1\{|r_{ij}| > \omega_T\}$ when $i \neq j$, and one if $i = j$. Then the estimated covariance matrix is defined as

$$\widehat{\mathbf{\Sigma}}^{*} = \text{diag}(\mathbf{S})^{1/2}\mathbf{R}^{\mathcal{T}}\text{diag}(\mathbf{S})^{1/2}.$$

In particular, when $\omega_T = 0$, it is exactly the sample covariance matrix, as no thresholding is employed, whereas when $\omega_T = 1$, it is a diagonal matrix with marginal sample variances on its diagonal. This form is more appropriate than the simple thresholding because it is thresholded on the standardized scale. Moreover, $\widehat{\mathbf{\Sigma}}^{*}$ is equivalent to applying the entry dependent thresholding

$$\omega_{T,ij} = \sqrt{s_{ii}s_{jj}}\omega_T$$

to the original sample covariance $\mathbf{S}$.

2.3. Generalized thresholding

The introduced thresholding estimators (2.2) and (2.4) are based on a simple thresholding rule, known as hard-thresholding. In regression and wavelet shrinkage contexts – see, e.g., Donoho et al. (1995) – hard-thresholding performs worse than some more flexible regularization methods, such as soft-thresholding and the smoothly clipped absolute deviation (SCAD; Fan and Li, 2001), which combine thresholding with shrinkages.

The generalized thresholding rules of Antoniadis and Fan (2001) can be applied to the estimation of large covariance matrices. A generalized thresholding rule depends on a thresholding parameter ω_T and a shrinkage function $h(\cdot;\omega_T) : \mathbb{R} \to \mathbb{R}$, which satisfies

$$(a)\quad |h(z,\omega_T)| \leq |z|; \qquad (b)\quad h(z;\omega_T) = 0 \text{ for } |z| \leq \omega_T; \qquad (c)\quad |h(z;\omega_T) - z| \leq \omega_T.$$

Here, condition (a) establishes shrinkage. While the shrinkage may slightly increase biases, it also reduces the variance and results in a more stable estimate of the covariance matrix. Condition (b) is the thresholding effect, and condition (c) limits the amount of shrinkage. There are a number of useful thresholding functions that are commonly used in the literature. For instance, soft-thresholding takes $h(z;\omega_T) = \text{sgn}(z)(|z| - \omega_T)_+$, where $(x)_+ = \max\{x, 0\}$. Another example is the SCAD thresholding of Fan and Li (2001). For some $a > 2$, it is defined as

$$h(z;\omega_T) = \begin{cases} \text{sgn}(|z| - \omega_T)_+, & |z| \leq 2\omega_T, \\ \{(a-1)z - \text{sgn}(z)a\omega_T\}/(a-2), & 2\omega_T < |z| \leq a\omega_T, \\ z, & |z| > a\omega_T. \end{cases}$$

This is a compromise between hard- and soft-thresholding, whose amount of shrinkage decreases as $|z|$ increases and hence results in a nearly unbiased estimation. Furthermore, the so-called minimax concave penalty (MCP) thresholding, proposed by Zhang (2010), is another example, defined, for some $a > 1$, as

$$h(z;\omega_T) = \begin{cases} (a/(a-1))\text{sgn}(|z| - \omega_T)_+, & |z| \leq a\omega_T, \\ z, & |z| > a\omega_T. \end{cases}$$

We can then define a generalized thresholding covariance estimator:

$$\widehat{\mathbf{\Sigma}} = (\widehat{\sigma}_{ij})_{p\times p}, \quad \widehat{\sigma}_{ij} = \begin{cases} s_{ij}, & \text{if } i = j, \\ h(s_{ij}; \omega_T), & \text{if } i \neq j. \end{cases} \tag{2.5}$$

Note that this admits the hard-thresholding estimator (2.2) as a special case by taking $h(z; \omega_T) = z1\{|z| > \omega_T\}$. Both the adaptive thresholding and entry-dependent thresholding can also be incorporated, by respectively setting $h(s_{ij}, \mathrm{SE}(s_{ij})\omega_T)$ and $h(s_{ij}, \sqrt{s_{ii}s_{jj}}\omega_T)$ on the (i, j)th element of the estimated covariance matrix when $i \neq j$. In addition, Rothman et al. (2009) have shown that the use of generalized thresholding rules does not affect the rate of convergence in (2.3), but it increases the family of shrinkages.

2.4. *Positive definiteness*

If the covariance matrix is sparse, it then follows from (2.3) that the thresholding estimator $\widehat{\mathbf{\Sigma}}$ is asymptotically positive definite. However, it is often more desirable to require the positive definiteness under finite samples. We discuss two approaches to achieving the finite sample positive definiteness.

2.4.1. Choosing the thresholding constant. For simplicity, we focus on the constant thresholding value $\omega_{T,ij} = \omega_T$; the case of entry-dependent thresholding can be dealt with similarly. The finite sample positive definiteness depends on the choice of the thresholding value ω_T, which also depends on a prescribed constant C through $\omega_T = C\sqrt{\log p/T}$. We write $\widehat{\mathbf{\Sigma}}(C) = \widehat{\mathbf{\Sigma}}$ to emphasize its dependence on C. When C is sufficiently large, the estimator becomes diagonal, and its minimum eigenvalue is strictly positive. We can then decrease the choice of C until it reaches

$$C_{\min} = \inf\{C > 0 : \lambda_{\min}(\widehat{\mathbf{\Sigma}}(M)) > 0, \quad \forall M > C\}.$$

Thus, $C_{\min}$ is well defined and for all $C > C_{\min}$, and $\widehat{\mathbf{\Sigma}}(C)$ is positive definite under finite sample. We can obtain $C_{\min}$ by solving $\lambda_{\min}(\widehat{\mathbf{\Sigma}}(C)) = 0, C \neq 0$. Figure 1 (taken from Fan et al., 2013) plots the minimum eigenvalue of $\widehat{\mathbf{\Sigma}}(C)$ as a function of C for a random sample from a Gaussian distribution with $p > T$, using three different thresholding rules. It is clearly seen from the figure that there is a range of C in which the covariance estimator is both positive definite and non-diagonal. When the minimum eigenvalue reaches its maximum value, the covariance estimator becomes diagonal. In practice, we can choose C in the range $(C_{\min} + \epsilon, M)$ for a small ϵ and large enough M by, e.g., cross-validations. This method was suggested by Fan et al. (2013) in a more complicated setting. Moreover, we also see from Figure 1 that the hard-thresholding rule yields the narrowest range for the choice C to give both positive definiteness and the non-diagonality.

2.4.2. Nearest positive definite matrices. An alternative approach to achieving the finite sample positive definiteness is through solving a constrained optimization problem. Qi and Sun (2006) introduced an algorithm for computing the nearest correlation matrix. Recall that $\mathbf{R}^{\mathcal{T}}$ is the thresholded correlation matrix, defined in Section 2.2; we find its nearest positive definite correlation matrix $\widehat{\mathbf{R}}$ by solving

$$\widehat{\mathbf{R}} = \underset{\mathbf{A}}{\operatorname{argmin}} \|\mathbf{R}^{\mathcal{T}} - \mathbf{A}\|_{\mathrm{F}}^2, \quad \text{s.t. } \mathbf{A} \geq 0, \operatorname{diag}(\mathbf{A}) = \mathbf{I}_p.$$

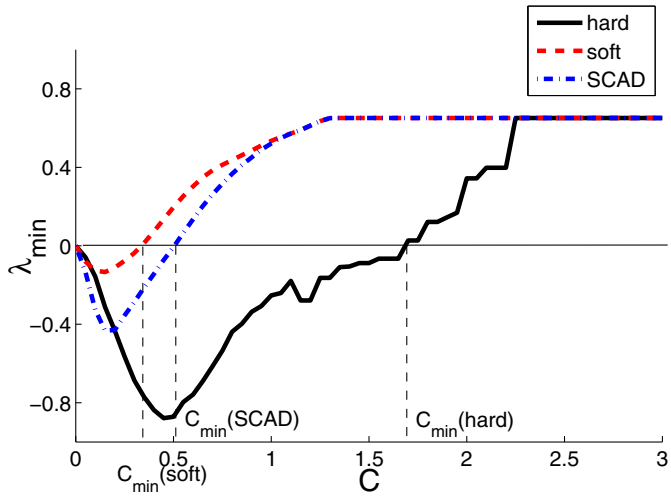

Figure 1. Minimum eigenvalue of $\widehat{\boldsymbol{\Sigma}}(C)$ as a function of C.

We can then transform back to the covariance matrix as: $\widehat{\boldsymbol{\Sigma}} = \text{diag}(\mathbf{S})^{1/2}\widehat{\mathbf{R}}\text{diag}(\mathbf{S})^{1/2}$. This procedure is often called nearest correlation matrix projection, and can be solved effectively using the R-package `nearPD`.

The nearest correlation matrix projection, however, does not necessarily result in a sparse solution. Liu et al. (2014) introduced a covariance estimation method called EC2 (estimation of covariance with eigenvalue constraints). To motivate this method, note that the thresholding method (2.5) can be equivalently cast as the solution to a penalized least-squares problem:

$$\widehat{\boldsymbol{\Sigma}} = \underset{\Sigma=(\sigma_{ij})}{\text{argmin}} \left\{ \frac{1}{2}\|\mathbf{S} - \boldsymbol{\Sigma}\|_{\text{F}}^2 + \sum_{i \neq j} P_{\omega_T}(\sigma_{ij}) \right\}.$$

Here, $P_{\omega_T}(\cdot)$ is a penalty function, which corresponds to the shrinkage function $h(\cdot, \omega_T)$. For instance, when

$$P_{\omega_T}(t) = \omega_T^2 - (|t| - \omega_T)^2 1\{|t| < \omega_T\},$$

the solution is the hard-thresholding estimator (2.2) (Antoniadis, 1997). See Antoniadis and Fan (2001) for the corresponding penalty functions of several popular shrinkage functions. The sparsity of the resulting estimator is hence due to the penalizations. We can modify the above penalized least-squares problem by adding an extra constraint to obtain positive definiteness:

$$\widetilde{\boldsymbol{\Sigma}} = \underset{\lambda_{\min}(\Sigma) \geq \tau}{\text{argmin}} \left\{ \frac{1}{2}\|\mathbf{S} - \boldsymbol{\Sigma}\|_{\text{F}}^2 + \sum_{i \neq j} P_{\omega_T}(\sigma_{ij}) \right\}. \tag{2.6}$$

Here, $\tau > 0$ is a pre-specified tuning parameter that controls the smallest eigenvalue of the estimated covariance matrix $\widetilde{\boldsymbol{\Sigma}}$. As a result, both sparsity and positive definiteness are guaranteed. Liu et al. (2014) showed that the problem (2.6) is convex when the penalty function is convex, and developed an efficient algorithm to solve it. More details on the algorithm and theory of this estimator are given in Section 4.

3. ESTIMATING SPARSE PRECISION MATRICES

Estimating a large inverse covariance matrix $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$ is another fundamental problem in modern multivariate analysis. Unlike the covariance matrix $\mathbf{\Sigma}$, which only captures the marginal correlations among $\boldsymbol{Y}_t = (Y_{1t}, \ldots, Y_{pt})'$, the inverse covariance matrix $\mathbf{\Theta}$ captures the conditional correlations among these variables and is closely related to undirected graphs under a Gaussian model.

More specifically, we define an undirected graph $G = (V, E)$, where V contains nodes corresponding to the p variables in $\boldsymbol{Y}_t$ and the edge $(j, k) \in E$ if and only if $\mathbf{\Theta}_{jk} \neq 0$. Under a Gaussian model $\boldsymbol{Y}_t \sim N(\mathbf{0}, \mathbf{\Sigma})$, the graph G describes the conditional independence relationships among $\boldsymbol{Y}_t = (Y_{1t}, \ldots, Y_{pt})'$. Let $\boldsymbol{Y}_{t,\setminus\{j,k\}} = \{Y_{\ell t} : \ \ell \neq j, k\}$. Y_{jt} is independent of Y_{kt} given $\boldsymbol{Y}_{t,\setminus\{j,k\}}$ for all $(j, k) \notin E$.

To illustrate the difference between the marginal and conditional uncorrelatedness, we consider a Gaussian model $\boldsymbol{Y}_t \sim N(\mathbf{0}, \mathbf{\Sigma})$ with

$$\mathbf{\Sigma} = \begin{pmatrix} 1.05 & -0.23 & 0.05 & -0.02 & 0.05 \\ -0.23 & 1.45 & -0.25 & 0.10 & -0.25 \\ 0.05 & -0.25 & 1.10 & -0.24 & 0.10 \\ -0.02 & 0.10 & -0.24 & 1.10 & -0.24 \\ 0.05 & -0.25 & 0.10 & -0.24 & 1.10 \end{pmatrix}$$

and

$$\mathbf{\Theta} = \begin{pmatrix} 1 & 0.2 & 0 & 0 & 0 \\ 0.2 & 1 & 0.2 & 0 & 0.2 \\ 0 & 0.2 & 1 & 0.2 & 0 \\ 0 & 0 & 0.2 & 1 & 0.2 \\ 0 & 0.2 & 0 & 0.2 & 1 \end{pmatrix}.$$

We see that the inverse covariance matrix $\mathbf{\Theta}$ has many zero entries. Thus, the undirected graph G defined by $\mathbf{\Theta}$ is sparse. However, the covariance matrix $\mathbf{\Sigma}$ is dense, which implies that every pair of variables is marginally correlated. Thus, the covariance matrix and inverse covariance matrix encode very different information. For example, even though Y_{1t} and Y_{5t} are conditionally uncorrelated given the other variables, they are marginally correlated. In addition to the graphical model problem, sparse precision matrix estimation has many other applications. Examples include high-dimensional discriminant analysis (Cai et al., 2011), portfolio allocation (Fan et al., 2008, 2012), principal component analysis and complex data visualization (Tokuda et al., 2011).

The estimation of the precision matrix $\mathbf{\Theta}$ requires very different techniques from estimating the covariance matrix. In the following subsections, we introduce several large precision estimation methods under the assumption that $\mathbf{\Theta}$ is sparse.

3.1. Penalized likelihood method

One of the most commonly used approaches to estimating sparse precision matrices is through the penalized maximum likelihood (Fan and Li, 2001). When $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_T$ are independently and identically distributed as $N(\mathbf{0}, \mathbf{\Sigma})$, the negative Gaussian log-likelihood function is given by $\ell(\mathbf{\Theta}) = \text{tr}(\mathbf{S}\mathbf{\Theta}) - \log|\mathbf{\Theta}|$. When the data are either non-Gaussian or weakly dependent,

$\ell(\boldsymbol{\Theta})$ becomes the quasi negative log-likelihood. Nevertheless, we then consider the following penalized likelihood method:

$$\widehat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta}=(\theta_{ij})_{p\times p}}{\operatorname{argmin}} \{\operatorname{tr}(\mathbf{S}\boldsymbol{\Theta}) - \log|\boldsymbol{\Theta}| + \sum_{i\neq j} P_{\omega_T}(|\theta_{ij}|)\}.$$

Here, the penalty function $P_{\omega_T}(|\theta_{ij}|)$, defined the same way as in Section 2.4.2, encourages the sparsity of $\widehat{\boldsymbol{\Theta}}$. One of the commonly used convex penalties is the ℓ_1 penalty $P_{\omega_T}(t) = \omega_T|t|$, and the problem is well studied in the literature; see, e.g., Yuan and Lin (2007), Friedman et al. (2008) and Banerjee et al. (2008). Other related works are found in, e.g., Meinshausen and Bühlmann (2006) and Wille et al. (2004).

In general, we recommend the use of folded concave penalties such as SCAD and MCP, as explained in Section 2.3. The main idea is that, in contrast to convex penalties, the shrinkage bias of folded concave penalties decreases as the magnitude of the parameter increases. Therefore, they are nearly unbiased, and have good performances for the sparse recovery (Lam and Fan, 2009). Computationally, while utilization of the ℓ_1-based penalty has many advantages for numerical optimizations, optimizations involving folded concave penalties can also be carried out using local linear approximations. More specifically, the penalized likelihood can be computed by an iterated reweighed Lasso. Given the estimate $\widehat{\boldsymbol{\Theta}}^{(k)} = (\widehat{\theta}_{ij}^{(k)})$ at the kth iteration, by Taylor expansion, we approximate

$$P_{\omega_T}(|\theta_{ij}|) \approx P_{\omega_T}(|\widehat{\theta}_{ij}^{(k)}|) + P'_{\omega_T}(|\widehat{\theta}_{ij}^{(k)}|)(|\theta_{ij}| - |\widehat{\theta}_{ij}^{(k)}|) \equiv Q_{\omega_T}(|\theta_{ij}|).$$

The linear approximation Q_{ω_T} is the convex majorant of the folded concave function at $|\widehat{\theta}_{ij}^{(k)}|$, namely, it satisfies

$$P_{\omega_T}(|\theta_{ij}|) \leq Q_{\omega_T}(|\theta_{ij}|) \quad \text{and} \quad P_{\omega_T}(|\widehat{\theta}_{ij}^{(k)}|) = Q_{\omega_T}(|\widehat{\theta}_{ij}^{(k)}|).$$

Replacing the penalty $P_{\omega_T}(|\theta_{ij}|)$ by its convex majorant $Q_{\omega_T}(|\theta_{ij}|)$, we update the estimate by (Fan et al., 2009)

$$\widehat{\boldsymbol{\Theta}}^{(k+1)} = \arg\min_{\boldsymbol{\Theta}=(\theta_{ij})} \{\operatorname{tr}(\mathbf{S}\boldsymbol{\Theta}) - \log|\boldsymbol{\Theta}| + \sum_{i\neq j} P'_{\omega_T}(|\widehat{\theta}_{ij}^{(k)}|)|\theta_{ij}|\} + c, \tag{3.1}$$

where c is a constant that does not depend on $\boldsymbol{\Theta}$. The problem (3.1) is convex and can be solved by the graphical Lasso algorithm of Friedman et al. (2008). Such an algorithm is called the majorization–minimization algorithm (Lange et al., 2000). Because the penalty function is majorized from above, it can easily be shown that the original objective function is decreasing in the iterations. Indeed, let $f(\boldsymbol{\Theta}) = \operatorname{tr}(\mathbf{S}\boldsymbol{\Theta}) - \log|\boldsymbol{\Theta}| + \sum_{i\neq j} P_{\omega_T}(|\theta_{ij}|)$ be the target value and let $g(\boldsymbol{\Theta})$ be its majorization function with $P_{\omega_T}(|\theta_{ij}|)$ replaced by $Q_{\omega_T}(|\theta_{ij}|)$. Then,

$$f(\widehat{\boldsymbol{\Theta}}^{(k+1)}) \leq g(\widehat{\boldsymbol{\Theta}}^{(k+1)}) \leq g(\widehat{\boldsymbol{\Theta}}^{(k)}) = f(\widehat{\boldsymbol{\Theta}}^{(k)}),$$

where the first inequality follows from the majorization, the second inequality comes from the minimization, and the last equality follows the majorization at the point $\widehat{\boldsymbol{\Theta}}^{(k)}$.

Theoretical properties of $\widehat{\boldsymbol{\Theta}}$ have been thoroughly studied by Rothman et al. (2008) and Lam and Fan (2009).

3.2. Column-by-column estimation methods

Another approach to estimating the precision matrix $\mathbf{\Theta}$ is through column-by-column regressions. More specifically, under the Gaussian model $\boldsymbol{Y}_t \sim N(\mathbf{0}, \mathbf{\Sigma})$ with $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$, unlike the penalized likelihood based method, this approach requires all the columns of $\mathbf{\Theta}$ to be sparse. Under this assumption, Meinshausen and Bühlmann (2006) exploit a sequence of Lasso regressions to separately estimate the columns of the precision matrix $\mathbf{\Theta}$. In follow-up works, Yuan (2010), Cai et al. (2011) and Liu and Luo (2012) proposed the graphical Dantzig selector, the CLIME estimator and the SCIO estimator, respectively. More recently, Sun and Zhang (2013) and Liu and Wang (2012) have proposed the scaled-Lasso and TIGER methods, which are tuning-insensitive and achieve rate optimality under different norms. Compared to the penalized likelihood methods, the column-by-column estimation methods are computationally simpler and more amenable to theoretical analysis. In this section, we mainly introduce the methods proposed by Meinshausen and Bühlmann (2006) and Yuan (2010) due to their simplicity.

The main idea of the column-by-column precision matrix estimation is to exploit the relationship between the conditional distribution of multivariate Gaussian and linear regressions. More specifically, let $\boldsymbol{Y} \sim N(\mathbf{0}, \mathbf{\Sigma})$, and the conditional distribution of Y_j given $\boldsymbol{Y}_{\setminus j}$ is given by

$$Y_j \mid \boldsymbol{Y}_{\setminus j} \sim N\big(\boldsymbol{\alpha}_j' \boldsymbol{Y}_{\setminus j}\,,\, \sigma_j^2\big),$$

where $\boldsymbol{\alpha}_j = (\mathbf{\Sigma}_{\setminus j,\setminus j})^{-1}\mathbf{\Sigma}_{\setminus j,j} \in \mathbb{R}^{p-1}$ and $\sigma_j^2 = \mathbf{\Sigma}_{jj} - \mathbf{\Sigma}_{\setminus j,j}(\mathbf{\Sigma}_{\setminus j,\setminus j})^{-1}\mathbf{\Sigma}_{\setminus j,j}$. Equivalently, we can represent this conditional distribution using a linear regression model

$$Y_j = \boldsymbol{\alpha}_j' \boldsymbol{Y}_{\setminus j} + \epsilon_j, \tag{3.2}$$

where $\epsilon_j \sim N(0, \sigma_j^2)$ is independent of $\boldsymbol{Y}_{\setminus j}$. Using the block matrix inversion formula, we have

$$\mathbf{\Theta}_{jj} = \sigma_j^{-2}, \qquad \mathbf{\Theta}_{\setminus j,j} = -\sigma_j^{-2}\boldsymbol{\alpha}_j. \tag{3.3}$$

Therefore, we can recover $\mathbf{\Theta}$ in a column-by-column manner by regressing Y_j on $\boldsymbol{Y}_{\setminus j}$ for $j = 1, 2, \ldots, p$. For example, let $\mathbf{Y} \in \mathbb{R}^{T\times p}$ be the data matrix. We denote by $\boldsymbol{\alpha}_j := (\alpha_{j1}, \ldots, \alpha_{j(p-1)})' \in \mathbb{R}^{p-1}$. Because the columns of $\mathbf{\Theta}$ are sparse, $\boldsymbol{\alpha}_j$ is also a sparse vector.

Inspired by the linear regression model in (3.2) and the fact that $\boldsymbol{\alpha}_j$ is sparse, Meinshausen and Bühlmann (2006) propose to estimate each $\boldsymbol{\alpha}_j$ by solving the Lasso regression

$$\widehat{\boldsymbol{\alpha}}_j = \underset{\boldsymbol{\alpha}_j \in \mathbb{R}^{p-1}}{\operatorname{argmin}} \frac{1}{2T}\|Y_{*j} - Y_{*\setminus j}\boldsymbol{\alpha}_j\|_2^2 + \lambda_j \|\boldsymbol{\alpha}_j\|_1,$$

where λ_j is a tuning parameter. Once $\widehat{\boldsymbol{\alpha}}_j$ is obtained, we obtain the neighbourhood edges by reading out the non-zero coefficients of $\boldsymbol{\alpha}_j$. The final graph estimate $\widehat{G}$ is obtained by combining the neighbourhoods for all the N nodes (e.g., if both nodes treat each other as neighbours, we put an edge connecting these two nodes). To estimate $\mathbf{\Theta}$, we also estimate σ_j^2 using the fitted sum of squared residuals

$$\widehat{\sigma}_j^2 = \frac{1}{T}\|Y_{*j} - Y_{*\setminus j}\alpha_j\|_2^2,$$

then plug it into (3.3).

Instead of the Lasso regression, Yuan (2010) proposes to estimate $\boldsymbol{\alpha}_j$ by solving

$$\widehat{\boldsymbol{\alpha}}_j = \underset{\boldsymbol{\alpha}_j \in \mathbb{R}^{p-1}}{\operatorname{argmin}} \|\boldsymbol{\alpha}_j\|_1 \qquad \text{subject to} \quad \|\mathbf{S}_{\setminus j,j} - \mathbf{S}_{\setminus j,\setminus j}\alpha_j\|_\infty \le \gamma_j,$$

where $\mathbf{S} := T^{-1}\mathbf{Y}'\mathbf{Y}$ is the sample covariance matrix and γ_j is a tuning parameter. This corresponds to replacing the Lasso regression by the Dantzig selector (Candès and Tao, 2007). The constraint corresponds to a sample version of $\boldsymbol{\Sigma}_{\backslash j,j} - \boldsymbol{\Sigma}_{\backslash j,\backslash j}\boldsymbol{\alpha}_j = 0$, with γ_j indicating the tolerance of the estimation error. Once $\widehat{\boldsymbol{\alpha}}_j$ is given, we can estimate σ_j^2 by $\widehat{\sigma}_j^2 = [1 - 2\widehat{\boldsymbol{\alpha}}_j'\mathbf{S}_{\backslash j,j} + \widehat{\boldsymbol{\alpha}}_j' S_{\backslash j,\backslash j}\widehat{\alpha}_j]^{-1}$. We then obtain an estimator $\widehat{\boldsymbol{\Theta}}$ of $\boldsymbol{\Theta}$ by plugging $\widehat{\boldsymbol{\alpha}}_j$ and $\widehat{\sigma}_j^2$ into (3.3). Yuan (2010) analyses the L_1-norm error $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_1$ and shows its minimax optimality over certain model space.

More recently, Sun and Zhang (2013) have proposed to estimate $\boldsymbol{\alpha}_j$ and σ_j by solving a scaled-Lasso problem:

$$\widehat{\mathbf{b}}_j, \widehat{\sigma}_j = \underset{\mathbf{b}=(b_1,\ldots,b_p)',\sigma}{\operatorname{argmin}} \left\{ \frac{\mathbf{b}'\mathbf{S}\mathbf{b}}{2\sigma} + \frac{\sigma}{2} + \lambda \sum_{k=1}^{p} \mathbf{S}_{kk}|b_k| \quad \text{subject to} \quad b_j = -1 \right\}.$$

Once $\widehat{\mathbf{b}}_j$ is obtained, we estimate $\widehat{\boldsymbol{\alpha}}_j = (\widehat{b}_1, \ldots, \widehat{b}_{j-1}, \widehat{b}_{j+1}, \ldots, \widehat{b}_p)'$. We then obtain the estimator of $\boldsymbol{\Theta}$ by plugging $\widehat{\boldsymbol{\alpha}}_j$ and $\widehat{\sigma}_j$ into (3.3). Sun and Zhang (2013) provided the spectral-norm rate of convergence of the obtained precision matrix estimator.

Similar to the idea of the graphical Dantzig selector, Cai et al. (2011) proposed the CLIME estimator, which stands for Constrained ℓ_1-Minimization for Inverse Matrix Estimation. This method directly estimates the jth column of $\boldsymbol{\Theta}$ by solving

$$\widehat{\boldsymbol{\Theta}}_{*j} = \underset{\boldsymbol{\Theta}_{*j}}{\operatorname{argmin}} \|\boldsymbol{\Theta}_{*j}\|_1 \quad \text{subject to } \|\mathbf{S}\boldsymbol{\Theta}_{*j} - \mathbf{e}_j\|_\infty \le \delta_j, \quad \text{for } j = 1, \ldots, p,$$

where $\mathbf{e}_j$ is the jth canonical vector (i.e., the vector with the jth element being 1, while the remaining elements are 0) and δ_j is a tuning parameter. Again, the constraint represents a sample version of $\boldsymbol{\Sigma}\boldsymbol{\Theta}_{*j} - \mathbf{e}_j = 0$. This optimization problem can be formulated into a linear program and has the potential to scale to large problems. Under regularity conditions, Cai et al. (2011) show that the estimator $\widehat{\boldsymbol{\Theta}}$ is asymptotically positive definite, and derive its rate of convergence.

4. ROBUST SPARSE PRECISION AND COVARIANCE ESTIMATION

The methods introduced in Sections 2 and 3 exploit the sample covariance matrix as input statistics. The theoretical justification of these methods relies on the sub-Gaussian assumption of the data. However, financial returns are often heavy-tailed. This section introduces a regularized rank-based framework for estimating large precision and covariance matrices under elliptical distributions. First, we introduce a rank-based precision matrix estimator, which naturally handles heavy-tailed features and conducts parameter estimation under the elliptical models. We also introduce a robust estimation method without assuming ellipticity on the distributions. Secondly, we introduce an adaptive rank-based covariance matrix estimator, which extends the generalized thresholding operator by adding an explicit eigenvalue constraint. We also provide interpretations of these rank-based estimators under the more general elliptical copula model, which illustrates a trade-off between model flexibility and interpretability.

Throughout this section, except for Section 4.1.3, we assume that the data follow an elliptical distribution (Fang et al., 1990), defined as below.

DEFINITION 4.1. (ELLIPTICAL DISTRIBUTION) *Given $\boldsymbol{\mu} \in \mathbb{R}^p$ and a symmetric positive semi-definite matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ with* $\text{rank}(\boldsymbol{\Sigma}) = r \leq p$, *a p-dimensional random vector* $\boldsymbol{Y} = (Y_1, \ldots, Y_p)'$ *follows an elliptical distribution with parameters $\boldsymbol{\mu}$, ξ and $\boldsymbol{\Sigma}$, denoted by $\boldsymbol{Y} \sim EC(\boldsymbol{\mu}, \xi, \boldsymbol{\Sigma})$, if $\boldsymbol{Y}$ has a stochastic representation*

$$\boldsymbol{Y} \overset{d}{=} \boldsymbol{\mu} + \xi \mathbf{A}\mathbf{u}, \tag{4.1}$$

where $\xi \geq 0$ is a continuous random variable independent of $\mathbf{u}$. Here $\mathbf{u} \in \mathbb{S}^{r-1}$ is uniformly distributed on the unit sphere in $\mathbb{R}^r$, and $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}'$.

For notational convenience, we use ξ instead of the distribution of ξ in the notation $EC(\boldsymbol{\mu}, \xi, \boldsymbol{\Sigma})$. Note that the model in (4.1) is not identifiable as we can rescale $\mathbf{A}$ and ξ without changing the distribution. In this section, we require $\mathbb{E}(\xi^2) < \infty$ and $\text{rank}(\boldsymbol{\Sigma}) = p$ to ensure the existence of the inverse of $\boldsymbol{\Sigma}$. In addition, we impose an identifiability condition $\mathbb{E}(\xi^2) = p$ to ensure that $\boldsymbol{\Sigma}$ is the covariance matrix of $\boldsymbol{Y}$. We still denote $\boldsymbol{\Theta} := \boldsymbol{\Sigma}^{-1}$.

4.1. Robust estimation of covariance matrix

This section introduces two methods for robust estimation of the covariance matrices. They will be used as an input that can be further regularized for sparse precision estimator and sparse covariance matrix estimation.

4.1.1. Robust estimation of correlation matrices. To robustly estimate $\mathbf{R}$, we adopt a Kendall's tau estimator proposed in Fang et al. (1990). We define the population Kendall's tau correlation between Y_j and Y_k as

$$\tau_{kj} = E\,\text{sign}((Y_j - \widetilde{Y}_j)(Y_k - \widetilde{Y}_k)),$$

where $\widetilde{Y}_j$ and $\widetilde{Y}_k$ are independent copies of Y_j and Y_k, respectively. For elliptical distributions, it is a well-known result that $\mathbf{R}_{kj}$ and τ_{kj} satisfy[2]

$$\mathbf{R} = [\mathbf{R}_{kj}] = \left(\sin\left(\frac{\pi}{2}\tau_{kj}\right)\right). \tag{4.2}$$

The sample version Kendall's tau statistic between Y_j and Y_k is

$$\widehat{\tau}_{kj} = \frac{2}{T(T-1)} \sum_{t<t'} \text{sign}((Y_{kt} - Y_{kt'})(Y_{jt} - Y_{jt'}))$$

for all $k \neq j$, and $\widehat{\tau}_{kj} = 1$ otherwise. We can plug $\widehat{\tau}_{kj}$ into (4.2) and obtain a rank-based correlation matrix estimator

$$\widehat{\mathbf{R}} = [\widehat{\mathbf{R}}_{kj}] = \left(\sin\left(\frac{\pi}{2}\widehat{\tau}_{kj}\right)\right). \tag{4.3}$$

4.1.2. Robust estimation of standard deviations. To estimate $\mathbf{D}$, we exploit an M-estimator proposed by Catoni (2012). Specifically, let $\psi(t) = \text{sign}(t) \cdot \log(1 + |t| + t^2/2)$ be a univariate

[2] More details can be found in Fang et al. (1990).

function where sign(0) = 0. Let $\widehat{\mu}_j$ and $\widehat{m}_j$ be the estimators of $\mathbb{E}Y_{jt}$ and $\mathbb{E}Y_{jt}^2$ by solving the following two estimating equations:

$$\sum_{t=1}^{T} \psi\Big((Y_{jt} - \mu_j)\sqrt{\frac{2}{TK_{\max}}}\Big) = 0, \tag{4.4}$$

$$\sum_{t=1}^{T} \psi\Big((Y_{jt}^2 - m_j)\sqrt{\frac{2}{TK_{\max}}}\Big) = 0, \tag{4.5}$$

where $K_{\max}$ is an upper bound of $\max_j \mathrm{Var}(Y_{jt})$ and $\max_j \mathrm{Var}(Y_{jt}^2)$. We assume that $K_{\max}$ is known. Catoni (2012) shows that the solutions to (4.4) and (4.5) must exist and can be efficiently solved using the Newton–Raphson algorithm (Stoer et al., 1993). Once $\widehat{m}_j$ and $\widehat{\mu}_j$ are obtained, we estimate the marginal standard deviation σ_j by

$$\widehat{\sigma}_j = \sqrt{\max\{\widehat{m}_j - \widehat{\mu}_j^2, K_{\min}\}}, \tag{4.6}$$

where $K_{\min}$ is a lower bound of $\min_j \sigma_j^2$ and is assumed to be known.

Compared to the sample covariance matrix, a remarkable property of $\widehat{\mathbf{R}}$ and $\widehat{\sigma}_j$ is that they concentrate to their population quantities exponentially fast even for heavy-tailed data. More specifically, Liu et al. (2012) show that

$$\|\widehat{\mathbf{R}} - \mathbf{R}\|_{\max} = O_P\Big(\sqrt{\frac{\log p}{T}}\Big) \qquad \text{and} \qquad \max_{1\le j\le p} |\widehat{\sigma}_j - \sigma_j| = O_P\Big(\sqrt{\frac{\log p}{T}}\Big). \tag{4.7}$$

In contrast, the sample correlation matrix and sample standard deviation do not have the above properties for heavy-tailed data. A simpler and less biased robust estimation is to use RA-mean in estimating the first two moments (Fan et al., 2016a). It simply takes $\psi(t) = \mathrm{sign}(t) \cdot \max(|t|, 1)$, the Huber's ψ-function. It also admits a similar concentration inequality with less bias. Indeed, the whole covariance matrix, not just the variance, can be estimated by this approach and it admits a similar concentration inequality, which we now introduce.

4.1.3. Robust estimation of covariance without ellipticity. This section gives a direct robust estimation of covariance matrix using RA-mean, which minimizes the robust approximation of the quadratic function (RA-quadratic). An advantage of this is that we avoid the assumption of elliptical distributions. Let $\{y_i\}_{i=1}^n$ be an i.i.d. sample from some unknown distribution with $E(y_i) = \mu$ and $\mathrm{Var}(y_i) = \sigma^2$. The RA-mean estimator $\widehat{\mu}_\alpha$ of μ is the solution of

$$\sum_{i=1}^{n} \psi(\alpha(y_i - \mu)) = 0,$$

for parameter $\alpha \to 0$, where $\psi(t) = \mathrm{sign}(t) \cdot \max(|t|, 1)$ corresponds to Huber's loss. Fan et al. (2016a) shows that for any $t \le \sqrt{2n}$ and $v > \sigma$,

$$P(\sqrt{n}|\widehat{\mu}_\alpha - \mu| \ge vt) \le 2\exp(-t^2/16), \qquad \text{for } \alpha = t/(4v\sqrt{n}). \tag{4.8}$$

Regarding $\sigma_{ij} = E[X_i X_j]$ as the expected value of the random variable $X_i X_j$, it can be estimated, based on a time series of length T, with accuracy

$$P(\sqrt{T}|\widehat{\sigma}_{ij} - \sigma_{ij}| \geq vt) \leq 2\exp(-t^2/16), \qquad \text{for } \alpha = t/(4v\sqrt{T}).$$

where $v \geq \max_{i,j\leq p}\sqrt{\mathrm{Var}(X_i X_j)}$ and $\widehat{\sigma}_{ij}$ is RA-mean estimator using data $\{X_{ik}X_{jk}\}_{k=1}^{T}$. Because there are only $O(p^2)$ elements, by taking $t^2 = 16a\log(p)$ for some $a > 2$ and the union bound, we have

$$P\Big\{\max_{i,j\leq p}|\widehat{\sigma}_{ij} - \sigma_{ij}| \geq 4v\sqrt{\frac{a\log p}{n}}\Big\} \leq 2p^{2-a}.$$

This covariance estimator that has been made robust requires a much weaker condition than the sample covariance and does not require ellipticity. It can be regularized further in the same way as the sample covariance matrix. Note that the above derivation assumes that $E[X_i] = 0$ for $i = 1, \ldots, p$. This can be removed by using the RA-mean estimator.

4.2. Robust sparse precision matrix estimation

To estimate $\boldsymbol{\Theta}$, our key observation is that the covariance matrix $\boldsymbol{\Sigma}$ can be decomposed as $\boldsymbol{\Sigma} = \mathbf{DRD}$, where $\mathbf{R}$ is the Pearson's correlation matrix, and $\mathbf{D} = \mathrm{diag}(\sigma_1, \ldots, \sigma_p)$ where σ_j is the standard deviation of Y_j. Because $\mathbf{D}$ is diagonal, we can represent the precision matrix as $\boldsymbol{\Theta} = \mathbf{D}^{-1}\boldsymbol{\Delta}\mathbf{D}^{-1}$, where $\boldsymbol{\Delta} = \mathbf{R}^{-1}$ is the inverse correlation matrix. This relationship motivates a three-step procedure for estimating the precision matrix $\boldsymbol{\Theta}$. First, obtain robust estimators $\widehat{\mathbf{R}}$ and $\widehat{\mathbf{D}}$ for $\mathbf{R}$ and $\mathbf{D}$. Then, apply an inverse correlation matrix estimation procedure on $\widehat{\mathbf{R}}$ to obtain $\widehat{\boldsymbol{\Delta}}$, an estimator for $\boldsymbol{\Delta}$. Finally, assemble $\widehat{\boldsymbol{\Delta}}$ and $\widehat{\mathbf{D}}$ to obtain a sparse precision matrix estimator $\widehat{\boldsymbol{\Theta}}$.

For light-tailed distributions (e.g., Gaussian or sub-Gaussian), we can directly use the sample correlation matrix and sample standard deviation to estimate $\mathbf{R}$ and $\mathbf{D}$. However, for heavy-tailed elliptical data, the sample correlation matrix and standard deviation estimators are inappropriate. Instead, we exploit a combination of the Kendall's tau estimator and Catoni's M-estimator, which are explained in detail in the following subsections. We would like to note that the RA-mean based robust covariance estimator in Section 4.1.3 can also be employed.

4.2.1. The EPIC method for inverse correlation matrix estimation. Once $\widehat{\mathbf{R}}$ and $\widehat{\mathbf{D}}$ are obtained, we first estimate the inverse correlation matrix $\boldsymbol{\Delta} = \mathbf{R}^{-1}$, then combine it with $\widehat{\mathbf{D}}$ to estimate the inverse covariance matrix. In this subsection, we describe the EPIC (i.e., Estimating Precision matrIx with Calibration) method, which is a robust precision matrix estimation method proposed by Zhao and Liu (2014). Its main idea is to estimate the jth column of $\boldsymbol{\Delta}$ by plugging $\widehat{\mathbf{R}}$ into the convex program,

$$(\widehat{\boldsymbol{\Delta}}_{*j}, \widehat{\tau}_j) = \underset{\boldsymbol{\Delta}_{*j},\tau_j}{\mathrm{argmin}} \ \|\boldsymbol{\Delta}_{*j}\|_1 + \frac{1}{2}\tau_j, \qquad \text{s.t. } \|\widehat{\mathbf{R}}\boldsymbol{\Delta}_{*j} - \mathbf{I}_{*j}\|_\infty \leq \lambda\tau_j, \ \|\Delta_{*j}\|_1 \leq \tau_j. \quad (4.9)$$

Here, τ_j serves as an auxiliary variable, which ensures that we can use the same regularization parameter λ for estimating different columns of $\boldsymbol{\Delta}$ (Gautier and Tsybakov, 2011). Both the objective function and constraints in (4.9) contain τ_j, which ensures that τ_j is bounded. Zhao and Liu (2014) show that the regularization parameter λ in (4.9) does not depend on the unknown quantity $\boldsymbol{\Delta}$. Thus, we can use the same λ to estimate different columns of $\boldsymbol{\Delta}$.

The optimization problem in (4.9) can be equivalently formulated into a linear program. For notational simplicity, we omit the index j in (4.9). We denote $\boldsymbol{\Delta}_{*j}$, $\mathbf{I}_{*j}$ and τ_j by $\boldsymbol{\gamma}$, $\mathbf{e}$ and τ, respectively. Let γ^+ and γ^- be the positive and negative parts of γ. By reparametrizing $\boldsymbol{\gamma} = \boldsymbol{\gamma}^+ - \boldsymbol{\gamma}^-$, we rewrite (4.9) as the following linear program

$$(\widehat{\boldsymbol{\gamma}}^+, \widehat{\boldsymbol{\gamma}}^-, \widehat{\tau}) = \underset{\boldsymbol{\gamma}^+, \boldsymbol{\gamma}^-, \tau}{\operatorname{argmin}} \mathbf{1}'\boldsymbol{\gamma}^+ + \mathbf{1}'\boldsymbol{\gamma}^- + c\tau \tag{4.10}$$

$$\text{s.t.} \begin{bmatrix} \widehat{\mathbf{R}} & -\widehat{\mathbf{R}} & -\boldsymbol{\lambda} \\ -\widehat{\mathbf{R}} & \widehat{\mathbf{R}} & -\boldsymbol{\lambda} \\ \mathbf{1}' & \mathbf{1}' & -1 \end{bmatrix} \begin{bmatrix} \boldsymbol{\gamma}^+ \\ \boldsymbol{\gamma}^- \\ \tau \end{bmatrix} \leq \begin{bmatrix} \mathbf{e} \\ -\mathbf{e} \\ 0 \end{bmatrix},$$

$$\boldsymbol{\gamma}^+ \geq \mathbf{0}, \boldsymbol{\gamma}^- \geq \mathbf{0}, \tau \geq 0,$$

where $\boldsymbol{\lambda} = \lambda\mathbf{1}$. The optimization problem in (4.10) can be solved by any linear program solver (e.g., the classical simplex method as suggested in Cai et al., 2011). In particular, it can be efficiently solved using the parametric simplex method (Vanderbei, 2008), which naturally exploits the underlying sparsity structure, and attains better empirical performance than a general-purpose solver.

4.2.2. Symmetric precision matrix estimation. Once we have obtained the inverse correlation matrix estimate $\widehat{\boldsymbol{\Delta}}$, we can estimate $\boldsymbol{\Theta}$ by

$$\widetilde{\boldsymbol{\Theta}} = \widehat{\mathbf{D}}^{-1}\widehat{\boldsymbol{\Delta}}\widehat{\mathbf{D}}^{-1}.$$

The EPIC method does not guarantee the symmetry of $\widetilde{\boldsymbol{\Theta}}$. To obtain a symmetric estimator, we take an additional projection step,

$$\widehat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} \|\boldsymbol{\Theta} - \widetilde{\boldsymbol{\Theta}}\|_* \qquad \text{s.t. } \boldsymbol{\Theta} = \boldsymbol{\Theta}', \tag{4.11}$$

where $\|\cdot\|_*$ can be the matrix ℓ_1-, Frobenius, or elementwise max norm. For both the Frobenius and elementwise max norms, (4.11) has a closed-form solution

$$\widehat{\boldsymbol{\Theta}} = \frac{1}{2}\big(\widetilde{\boldsymbol{\Theta}} + \widetilde{\boldsymbol{\Theta}}'\big).$$

When using the matrix ℓ_1-norm, the optimization problem in (4.11) does not have a closed-form solution. For this, we can exploit the smoothed proximal gradient algorithm to solve it. More details about this algorithm can be found in Zhao and Liu (2014).

Consider a class of sparse symmetric matrices

$$\mathcal{U}(s, M, \kappa_u) = \{\boldsymbol{\Delta} \in \mathbb{R}^{p\times p} | \boldsymbol{\Delta} \succ 0,\ \max_j \sum_k \mathrm{I}\{\boldsymbol{\Delta}_{kj} \neq 0\} \leq s, \|\boldsymbol{\Delta}\|_1 \leq M, \Lambda_{\max}(\boldsymbol{\Delta}) \leq \kappa_u\},$$

where κ_u is a constant, and (s, p, M) may scale with the sample size T. Under some mild conditions, Zhao and Liu (2014) show that if we take $\lambda = \kappa_1\sqrt{(\log p)/T}$ and choose the matrix ℓ_1-norm as $\|\cdot\|_*$ in (4.11), then for large enough T, we have

$$\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_2 = O_P\Big(M \cdot s\sqrt{\frac{\log p}{T}}\Big). \tag{4.12}$$

Moreover, if we choose the Forbenius norm as $\|\cdot\|_*$ in (4.11), then for large enough T,

$$\frac{1}{p}\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_{\mathrm{F}}^2 = O_P\Big(M^2 \frac{s \log p}{T}\Big). \tag{4.13}$$

4.3. Robust sparse covariance matrix estimation

In this subsection, we consider the problem of estimating the covariance matrix $\boldsymbol{\Sigma}$ under the elliptical model (4.1). Similar to Section 2, we impose a sparsity assumption on $\boldsymbol{\Sigma}$. To estimate $\boldsymbol{\Sigma}$, Liu et al. (2014) introduce a regularized rank-based estimation method called EC2, which can be viewed as a robust extension of the generalized thresholding operator (Rothman et al., 2009). The EC2 estimator can be formulated as the solution to a convex program, which ensures the positive definiteness of the estimated covariance matrix. Unlike most existing methods, the EC2 estimator explicitly constrains the smallest eigenvalue of the estimated covariance matrix.

4.3.1. The EC2 Estimator. Recall that $\boldsymbol{\Sigma} = \mathbf{DRD}$. Similar to the EPIC method, we calculate the EC2 estimator in three steps. In the first step, we obtain robust estimators $\widehat{\mathbf{R}}$ and $\widehat{\mathbf{D}}$ for $\mathbf{R}$ and $\mathbf{D}$. In the second step, we apply an optimization procedure on $\widehat{\mathbf{R}}$ to obtain $\widehat{\mathbf{R}}^{\mathrm{EC2}}$, a sparse estimator for $\mathbf{R}$. In the third step, we assemble $\widehat{\mathbf{R}}^{\mathrm{EC2}}$ and $\widehat{\mathbf{D}}$ to obtain the final sparse covariance matrix estimator $\widehat{\boldsymbol{\Sigma}} = \widehat{\mathbf{D}}\widehat{\mathbf{R}}^{\mathrm{EC2}}\widehat{\mathbf{D}}$. Specifically, we calculate $\widehat{\mathbf{R}}$ and $\widehat{\mathbf{D}}$ as in (4.3) and (4.6). In the following, we focus on explaining how to obtain $\widehat{\mathbf{R}}^{\mathrm{EC2}}$ based on $\widehat{\mathbf{R}}$.

Recall that $\widehat{\mathbf{R}}$ is the Kendall's tau matrix defined in (4.3). The $\widehat{\mathbf{R}}^{\mathrm{EC2}}$ is calculated as

$$\widehat{\mathbf{R}}^{\mathrm{EC2}} := \underset{\mathrm{diag}(\mathbf{R})=1}{\operatorname{argmin}} \frac{1}{2}\|\widehat{\mathbf{R}} - \mathbf{R}\|_{\mathrm{F}}^2 + \lambda\|\mathbf{R}\|_{1,\mathrm{off}} \qquad \text{s.t.} \quad \tau \le \Lambda_{\min}(\mathbf{R}), \tag{4.14}$$

where $\lambda > 0$ is a regularization parameter and $\tau > 0$ is a desired minimum eigenvalue lower bound of the estimator, which is assumed to be known. The EC2 method simultaneously conducts sparse estimation and guarantees the positive-definiteness of the solution. The equality constraint $\mathrm{diag}(\mathbf{R}) = 1$ ensures that $\widehat{\mathbf{R}}^{\mathrm{EC2}}$ is a correlation matrix. Once $\widehat{\mathbf{R}}^{\mathrm{EC2}}$ is obtained, we convert it to the final covariance matrix estimator $\widehat{\boldsymbol{\Sigma}}$ as described above. Liu et al. (2014) prove the convexity of the formulation in (4.14). Alternatively, one can apply thresholding on $\widehat{\mathbf{R}}$ to obtain a positive-definite estimator.

4.3.2. Asymptotic properties of the EC2 estimator. To establish the asymptotic properties of the EC2 estimator, for $0 \le q < 1$, we consider the following class of sparse correlation matrices:

$$\mathcal{M}(q, M_p, \delta) := \{\mathbf{R} : \max_{1\le j\le p} \sum_{k\ne j} |R_{jk}|^q \le M_p \quad \text{and} \quad R_{jj} = 1 \text{ for all } j, \Lambda_{\min}(\mathbf{R}) \ge \delta\}.$$

We also define a class of covariance matrices,

$$\mathcal{U}(\kappa, q, M_p, \delta) := \{\boldsymbol{\Sigma} : \max_j \boldsymbol{\Sigma}_{jj} \le \kappa \quad \text{and} \quad \mathbf{D}^{-1}\boldsymbol{\Sigma}\mathbf{D}^{-1} \in \mathcal{M}(q, M_p, \delta)\}, \tag{4.15}$$

where $\mathbf{D} = \mathrm{diag}(\sqrt{\boldsymbol{\Sigma}_{11}}, \ldots, \sqrt{\boldsymbol{\Sigma}_{pp}})$. The definition of this class is similar to the universal thresholding class defined by Bickel and Levina (2008).

Under the assumption that the data follow an elliptical distribution, Liu et al. (2014) show that, for large enough T, the EC2 estimator $\widehat{\mathbf{\Sigma}}$ satisfies

$$\sup_{\mathbf{\Sigma}\in\mathcal{U}(\kappa,q,M_p,\delta_{\min})} \mathbb{E}\|\widehat{\mathbf{\Sigma}}^{\mathrm{EC2}} - \mathbf{\Sigma}\|_2 \leq c_1 \cdot M_p\Big(\frac{\log p}{T}\Big)^{(1-q)/2}. \tag{4.16}$$

Cai and Zhou (2012) show that the rate in (4.16) attains the minimax lower bound over the class $\mathcal{U}(\kappa, q, M_d, \delta_{\min})$ under the Gaussian model. Thus, the EC2 estimator is asymptotically rate optimal under the flexible elliptical model with covariance matrix in $\mathcal{U}(\kappa, q, M_d, \delta_{\min})$.

5. FACTOR MODEL-BASED COVARIANCE ESTIMATION WITH OBSERVABLE FACTORS

Most of the aforementioned methods of estimating $\mathbf{\Sigma}$ assume that the covariance matrix is sparse. Though this assumption is reasonable for some applications, it is not always appropriate. For example, financial stocks share the same market risks and hence their returns are highly correlated; all the genes from the same pathway may be co-regulated by a small amount of regulatory factors, which makes the gene expression data highly correlated; when genes are stimulated by cytokines, their expressions are also highly correlated. The sparsity assumption is obviously inappropriate in these situations.

In many applications, the responses of cross-sectional units often depend on a few common factors $\boldsymbol{f}$:

$$Y_{it} = \mu_i + \mathbf{b}\boldsymbol{f}_t + u_{it}. \tag{5.1}$$

Here, μ_i is the mean, $\mathbf{b}_i$ is a vector of factor loadings, $\boldsymbol{f}_t$ is a $K \times 1$ vector of common factors and u_{it} is the error term, usually called the idiosyncratic component, uncorrelated with $\boldsymbol{f}_t$. Factor models have long been employed in financial studies, where Y_{it} often represents the excess returns of the ith asset (or stock) on time t. The literature includes, for instance, Fama and French (1992), Chamberlain and Rothschild (1983) and Campbell et al. (1996). It is also commonly used in macroeconomics for forecasting diffusion index (e.g., Stock and Watson, 2002). The literature on high-dimensional factor models is rapidly growing. Beside those mentioned above, the list also includes, for instance, Breitung and Tenhofen (2011), Boivin and Ng (2006), Doz et al. (2012), Forni et al. (2000), Forni and Lippi (2001), Kapetanios (2010), Lam and Yao (2012), Hallin and Liška (2007), Choi (2012), Tsai and Tsay (2010), Onatski (2012), Bai and Li (2012), Alessi et al. (2010) and Park et al. (2009), among many others.

This section introduces a method of estimating $\mathbf{\Sigma}$ using factor models. We focus on the case when the factors are observable. The observable factor models are of considerable interest as they are often the case in empirical analyses in finance. We allow $p, T \to \infty$ and we allow that p can grow much faster than T. In contrast, the number of factors K either needs to be bounded or grows slowly.

5.1. Conditional sparsity

The factor model (5.1) can be put in a matrix form as

$$\boldsymbol{Y}_t = \mu + \mathbf{B}\boldsymbol{f}_t + u_t, \tag{5.2}$$

where $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_p)'$ and $\boldsymbol{u}_t = (u_{1t}, \ldots, u_{pt})'$. We are interested in $\boldsymbol{\Sigma}$, the $p \times p$ covariance matrix of $\boldsymbol{Y}_t$, and its inverse $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$, which are assumed to be time-invariant. Under model (5.2) and the assumption that $\boldsymbol{f}_t$ and $\boldsymbol{u}_t$ are uncorrelated, $\boldsymbol{\Sigma}$ is given by

$$\boldsymbol{\Sigma} = \mathbf{B}\,\mathrm{Cov}(\boldsymbol{f}_t)\mathbf{B}' + \boldsymbol{\Sigma}_u, \tag{5.3}$$

where $\boldsymbol{\Sigma}_u = (\sigma_{u,ij})_{p\times p}$ is the covariance matrix of $\boldsymbol{u}_t$. Estimating the covariance matrix $\boldsymbol{\Sigma}_u$ of the idiosyncratic components $\{\boldsymbol{u}_t\}$ is also important for statistical inferences.

Fan et al. (2008) studied model (5.3) when $p \to \infty$ possibly faster than T. They assumed $\boldsymbol{\Sigma}_u$ to be a diagonal matrix, which corresponds to the classical strict factor model, and might be restrictive in practical applications. However, factor models are often only justified as being approximate, in which $Y_{1t}, \ldots, Y_{pt}$ are still mutually correlated given the factors, though the mutual correlations are weak. This gives rise to the approximate factor model studied by Chamberlain and Rothschild (1983). In the approximate factor model, $\boldsymbol{\Sigma}_u$ is a non-diagonal covariance matrix, and admits many small off-diagonal entries.

In the decomposition (5.3), we assume $\boldsymbol{\Sigma}_u$ to be sparse. This can be interpreted as the conditional sparse covariance model. Given the common factors $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_T$, the conditional (after taking out the linear projection on to the space spanned by the factors) covariance matrix of $\boldsymbol{Y}_t$ is sparse. Let

$$m_{u,p} = \begin{cases} \max_{i\le p} \sum_{j=1}^{p} 1\{\sigma_{u,ij} \neq 0\}, & \text{if } q = 0, \\ \max_{i\le p} \sum_{j=1}^{p} |\sigma_{u,ij}|^q, & \text{if } 0 < q < 1. \end{cases} \tag{5.4}$$

We require $m_{u,p}$ either to be bounded or to grow slowly as $p \to \infty$. The conditional sparsity assumption is slightly stronger than those of the approximate factor model in Chamberlain and Rothschild (1983), but is still a natural assumption. In contrast, note that in the presence of common factors, $\boldsymbol{\Sigma}$ itself is hardly a sparse matrix.

Empirically, the sparsity of $\boldsymbol{\Sigma}_u$ may arise from block diagonal structures, where the blocks may match industry sectors. Ang et al. (2008) modelled the idiosyncratic components by

$$u_{it} = \boldsymbol{\lambda}_i' \boldsymbol{v}_t + \epsilon_{it},$$

where $\boldsymbol{v}_t$ denotes ten unobserved industry-specific factors, and the jth component of $\boldsymbol{\lambda}_i$ is defined as $\lambda_{ij} = 1\{i \text{ belongs to industry } j\}$, $j = 1, \ldots, 10$; ϵ_{it} is independent across i. This then leads to a sparse covariance $\boldsymbol{\Sigma}_u$. Ang et al. (2008) provided empirical evidences to support this model.

5.2. Estimation

To facilitate the notation, we take $\mu_i = 0$ in (5.1). When the factors are observable, we can estimate $\mathbf{B}$ by the ordinary least-squares (OLS): $\widehat{\mathbf{B}} = (\widehat{\mathbf{b}}_1, \ldots, \widehat{\mathbf{b}}_N)'$, where

$$\widehat{\mathbf{b}}_i = \arg\min_{\mathbf{b}_i} \frac{1}{T} \sum_{t=1}^{T} (Y_{it} - \mathbf{b}_i' \boldsymbol{f}_t)^2, \quad i = 1, \ldots, N.$$

Then, $\widehat{\boldsymbol{u}}_t = \boldsymbol{Y}_t - \widehat{\mathbf{B}}\boldsymbol{f}_t$ is the residual vector at time t. We then construct the residual covariance matrix as

$$\mathbf{S}_u = \frac{1}{T} \sum_{t=1}^{T} \widehat{u}_t \widehat{u}_t' = (s_{u,ij}).$$

Because $\mathbf{\Sigma}_u$ is sparse, we now apply thresholding on $\mathbf{S}_u$ to regularize the estimator. Define

$$\widehat{\mathbf{\Sigma}}_u = (\widehat{\sigma}_{u,ij})_{p\times p}, \quad \widehat{\sigma}_{u,ij} = \begin{cases} s_{u,ii}, & i = j; \\ h(s_{u,ij}; \omega_{T,ij}), & i \neq j. \end{cases}$$

Here, $h(.; \omega_{T,ij})$ is a general thresholding rule as described in Section 2. Both the adaptive thresholding and entry dependent thresholding can also be incorporated, by respectively setting either $\omega_{T,ij} = \mathrm{SE}(s_{u,ij})\omega_T$ or $\omega_{T,ij} = \sqrt{s_{u,ii}s_{u,jj}}\omega_T$, with

$$\omega_T = CK\sqrt{\frac{\log p}{T}}$$

for some $C > 0$. As in the discussions in Section 2, $C > 0$ can be chosen via cross-validation in a proper range to guarantee the finite sample positive definiteness.

The covariance matrix $\mathrm{Cov}(\boldsymbol{f}_t)$ can be estimated by the sample covariance matrix

$$\widehat{\mathrm{Cov}}(\boldsymbol{f}_t) = \frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{f}_t - \bar{\boldsymbol{f}})(\boldsymbol{f}_t - \bar{\boldsymbol{f}})', \quad \bar{\boldsymbol{f}} = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}_t,$$

which does not require a regularization as the number of factors is assumed to be small. Therefore, we obtain a substitution estimator:

$$\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{B}}\widehat{\mathrm{Cov}}(\boldsymbol{f}_t)\widehat{\mathbf{B}}' + \widehat{\mathbf{\Sigma}}_u.$$

By the Sherman–Morrison–Woodbury formula, we estimate the precision matrix as

$$\widehat{\mathbf{\Sigma}}^{-1} = \widehat{\mathbf{\Sigma}}_u^{-1} - \widehat{\mathbf{\Sigma}}_u^{-1}\widehat{\mathbf{B}}[\widehat{\mathrm{Cov}}(\boldsymbol{f}_t)^{-1} + \widehat{\mathbf{B}}'\widehat{\mathbf{\Sigma}}_u^{-1}\widehat{\mathbf{B}}]^{-1}\widehat{\mathbf{B}}'\widehat{\mathbf{\Sigma}}_u^{-1}.$$

Under regularity conditions, Fan et al. (2011) showed that when $m_{u,p}\omega_T^{1-q} \to 0$,

$$\|\widehat{\mathbf{\Sigma}}_u - \mathbf{\Sigma}_u\|_2 = O_P(m_{u,p}\omega_T^{1-q}), \quad \|\widehat{\mathbf{\Sigma}}_u^{-1} - \mathbf{\Sigma}_u^{-1}\|_2 = O_P(m_{u,p}\omega_T^{1-q}).$$

However, it is difficult to obtain a satisfactory convergence rate for $\widehat{\mathbf{\Sigma}}$ under either the operator or the Frobenius norm. We illustrate this problem in the following example.

EXAMPLE 5.1. Consider the specific case $K = 1$ with the known loading $\boldsymbol{B} = \mathbf{1}_p$ and $\mathbf{\Sigma}_u = \mathbf{I}$. Then $\mathbf{\Sigma} = \mathrm{Var}(f_1)\mathbf{1}_p\mathbf{1}_p' + \mathbf{I}$, where $\mathbf{1}_p$ denotes the p-dimensional column vector of ones with $\|\mathbf{1}_p\mathbf{1}_p'\|_2 = p$, and we only need to estimate $\mathrm{Var}(f_1)$ using the sample variance. Then, it follows that

$$\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_2 = |\frac{1}{T}\sum_{t=1}^{T}(f_{1t} - \bar{f}_1)^2 - \mathrm{Var}(f_{1t})| \cdot \|\mathbf{1}_p\mathbf{1}_p'\|_2.$$

Therefore, it follows from the central limit theorem that $(\sqrt{T}/p)\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_2$ is asymptotically normal. Hence, $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_2$ diverges if $p \gg \sqrt{T}$, even for such a simplified toy model.

In the above toy example, the bad rate of convergence is mainly due to the large quantity $\|\mathbf{1}_p\mathbf{1}_p'\|_2$, which comes from the high-dimensional factor loadings. In general, the high-dimensional loading matrix accumulates many estimation errors.

However, Fan et al. (2011) showed that we can obtain a good convergence rate when estimating $\mathbf{\Sigma}^{-1}$:

$$\|\widehat{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}^{-1}\|_2 = O_P(m_{u,p}\omega_T^{1-q}).$$

Intuitively, the good performance of $\widehat{\mathbf{\Sigma}}^{-1}$ follows from the fact that the eigenvalues of $\mathbf{\Sigma}^{-1}$ are uniformly bounded, whereas the leading eigenvalues of $\mathbf{\Sigma}$ diverge fast.

6. FACTOR-MODEL-BASED COVARIANCE ESTIMATION WITH LATENT FACTORS

In many empirical studies using factor models, the common factors are often latent, that is, they are unobservable. In this case, the covariance matrix of $\boldsymbol{Y}_t$ has the same decomposition as before,

$$\mathbf{\Sigma} = \mathbf{B}\,\mathrm{Cov}(\boldsymbol{f}_t)\mathbf{B}' + \mathbf{\Sigma}_u, \tag{6.1}$$

but the latent factors also need to be estimated. Similar to the case of observable factors, the model can be assumed to be conditionally sparse, where $\mathbf{\Sigma}_u$ is a sparse matrix but not necessarily diagonal. In this section, we shall assume the number of factors to be bounded.

6.1. The pervasive condition

Note that unlike the classical factor analysis (e.g., Lawley and Maxwell, 1971), when $\mathbf{\Sigma}_u$ is non-diagonal, the decomposition (6.1) is not identifiable under fixed (p, T), as $\boldsymbol{Y}_t$ is the only observed data in the model. Here the identification means the separation of the low-rank part $\mathbf{B}\,\mathrm{Cov}(\boldsymbol{f}_t)\mathbf{B}'$ from $\mathbf{\Sigma}_u$ in the decomposition (6.1). Interestingly, however, the identification of $\mathbf{B}\,\mathrm{Cov}(\boldsymbol{f}_t)\mathbf{B}'$ can be achieved asymptotically, by letting $p \to \infty$ and requiring the eigenvalues of $\mathbf{\Sigma}_u$ either to be uniformly bounded or to grow slowly relative to p. As we will soon see, we assume without loss of generality that $\mathrm{Cov}(\boldsymbol{f}_t)$ is identity.

What makes the asymptotic identification possible is the following pervasive assumption, which is one of the key conditions assumed in the literature; see, e.g., Stock and Watson (2002) and Bai (2003).

ASSUMPTION 6.1. *The eigenvalues of the $K \times K$ matrix $p^{-1}\mathbf{B}'\mathbf{B} = (1/p)\sum_{i=1}^{p} \mathbf{b}_i\mathbf{b}_i'$ are uniformly bounded away from both zero and infinity, as $p \to \infty$.*

When this assumption is satisfied, the factors are said to be pervasive. It requires the factors to impact on most of the cross-sectional individuals. It then follows that the first K eigenvalues of $\mathbf{B}\mathbf{B}'$, which are p times those of $p^{-1}\mathbf{B}'\mathbf{B}$, are bounded from below by cp for some $c > 0$, and should grow fast with p. However,

$$\|\mathbf{\Sigma}_u\|_2 \le \max_{i\le p} \sum_{j=1}^{p} |\sigma_{u,ij}|^q |\sigma_{u,ii}\sigma_{u,jj}|^{(1-q)/2} \le m_{u,p} \max_{i\le p} \sigma_{u,ii}^{1-q}. \tag{6.2}$$

Hence, when $m_{u,p}$ grows at $o(p)$, the leading eigenvalues of the two components on the right-hand side of (6.1) are well separated as $p \to \infty$. This guarantees that the covariance decomposition is asymptotically identified. Intuitively, as the dimension increases, the information about the common factors accumulates, while the information about the idiosyncratic components does not. This eventually distinguishes the factor components $\mathbf{B}\boldsymbol{f}_t$ from $\boldsymbol{u}_t$.

Below we shall introduce a principal component analysis (PCA) based method to estimate the covariance matrix.

6.2. Principal Component and Factor Analysis

Before introducing the estimator of $\mathbf{\Sigma}$ in the case of latent factors, we first elucidate why PCA can be used for the factor analysis when the number of variables is large. First of all, note that even if $\mathbf{B}\,\mathrm{Cov}(\boldsymbol{f}_t)\mathbf{B}'$ is asymptotically identifiable, $\mathbf{B}$ and $\boldsymbol{f}_t$ are not separately identifiable, as the pair $(\mathbf{B}, \boldsymbol{f}_t)$ is equivalent to the pair $(\mathbf{BH}^{-1}, \mathbf{H}\boldsymbol{f}_t)$ for any $K \times K$ non-singular matrix $\mathbf{H}$. To resolve the ambiguity between $\mathbf{B}$ and $\boldsymbol{f}_t$, we impose the identifiability constraint that $\mathrm{Cov}(\boldsymbol{f}_t) = \mathbf{I}_K$ and that the columns of $\mathbf{B}$ are orthogonal. Under this canonical form, it then follows from (6.1) that

$$\mathbf{\Sigma} = \mathbf{BB}' + \Sigma_u.$$

We now attempt to identify $\boldsymbol{B}$ and hence $\mathbf{\Sigma}_u$ from $\mathbf{\Sigma}$. Let $\widetilde{\mathbf{b}}_1, \ldots, \widetilde{\mathbf{b}}_K$ be the columns of $\mathbf{B}$. Because the columns of $\mathbf{B}$ are orthogonal,

$$\mathbf{BB}'\widetilde{\mathbf{b}}_j = \widetilde{b}_j \|\widetilde{b}_j\|_2^2, \quad \text{for } j \le K.$$

Therefore, $\widetilde{\mathbf{b}}_1/\|\widetilde{\mathbf{b}}_1\|_2, \cdots, \widetilde{\mathbf{b}}_K/\|\widetilde{\mathbf{b}}_K\|_2$ are the eigenvectors of $\mathbf{BB}'$, corresponding to the largest K eigenvalues $\{\|\widetilde{\mathbf{b}}_j\|_2^2\}_{j=1}^K$; the rest $p-K$ eigenvalues of $\mathbf{BB}'$ are zeros. To guarantee the uniqueness (up to a sign change) of the leading eigenvectors, we also assume $\{\|\widetilde{\mathbf{b}}_j\|_2\}_{j=1}^K$ are distinct and sorted in a decreasing order. To see how large these eigenvalues are, note that the first K eigenvalues of $\mathbf{BB}'$ are the same as those of $\mathbf{B}'\mathbf{B}$. Hence, it follows from the pervasive assumption (Assumption 6.1) that

$$\|\widetilde{\mathbf{b}}_j\|_2^2 \ge cp, \quad j = 1, \ldots, K. \tag{6.3}$$

Next, let us associate the leading eigenvalues of $\mathbf{BB}'$ with those of $\mathbf{\Sigma}$. Let $\lambda_1, \ldots, \lambda_K$ denote the K largest eigenvalues of $\mathbf{\Sigma}$, and let $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K$ be their corresponding eigenvectors. Applying Wely's theorem and the $\sin(\theta)$-theorem of Davis (1963), Fan et al. (2013) showed

$$\|\boldsymbol{\xi}_j - \widetilde{\mathbf{b}}_j/\|\widetilde{\mathbf{b}}_j\|_2\|_2 = O(p^{-1}\|\mathbf{\Sigma}_u\|_2), \qquad \text{for all } j \le K.$$

and

$$|\lambda_j - \|\widetilde{\mathbf{b}}_j\|_2^2| \le \|\mathbf{\Sigma}_u\|_2, \text{ for } j \le K, \qquad |\lambda_j| \le \|\mathbf{\Sigma}_u\|_2, \text{ for } j > K.$$

These results demonstrate the following.

1. The leading eigenvectors of $\mathbf{\Sigma}$ are approximately equal to the normalized columns of $\mathbf{B}$, as $p \to \infty$ and $\widetilde{\mathbf{b}}_j \approx \sqrt{\lambda_j}\boldsymbol{\xi}_j$ (see point 4 below for further details). In other words, the factor analysis and the principal analysis are approximately the same.
2. The leading eigenvalues of $\mathbf{\Sigma}$ grow at the rate $O(p)$. This can be seen from applying the triangular inequality and (6.2), (6.3):

$$\lambda_j > \|\widetilde{\mathbf{b}}_j\|_2^2 - |\lambda_j - \|\widetilde{\mathbf{b}}_j\|^2| \ge cp - m_{u,p}\max_{i\le p}\sigma_{u,ii}^{1-q}, \quad \forall j = 1, \ldots, K.$$

3. The latent factor f_{jt} is approximately $\boldsymbol{\xi}_j' \boldsymbol{Y}_t/\sqrt{\lambda_j}$ for $j = 1, \ldots, K$. To see this, left-multiplying $\widetilde{\mathbf{b}}_j'/\|\widetilde{\mathbf{b}}_j\|_2^2$ to $\boldsymbol{Y}_t = \mathbf{B}\boldsymbol{f}_t + u_t$, and noting that the columns of $\mathbf{B}$ are orthogonal, we have

$$f_{jt} = \widetilde{\mathbf{b}}_j' \boldsymbol{Y}_t/\|\widetilde{\mathbf{b}}_j\|_2^2 - \widetilde{\mathbf{b}}_j' \boldsymbol{u}_t/\|\widetilde{\mathbf{b}}_j\|_2^2.$$

The second term on the right is the weighted average of noise $\boldsymbol{u}_t$ over all p individuals and hence typically negligible when p is large. The first term is

$$\frac{\widetilde{\mathbf{b}}_j' \boldsymbol{Y}_t}{\|\widetilde{\mathbf{b}}_j\|_2^2} = \frac{\widetilde{\mathbf{b}}_j'/\|\widetilde{\mathbf{b}}_j\|_2 \boldsymbol{Y}_t}{\|\widetilde{\mathbf{b}}_j\|_2} \approx \frac{\boldsymbol{\xi}_j' \boldsymbol{Y}_t}{\sqrt{\lambda_j}}.$$

Hence, as $p \to \infty$, $f_{jt} \approx \boldsymbol{\xi}_j' \boldsymbol{Y}_t/\sqrt{\lambda_j}$.

4. Note that $\mathbf{B}\mathbf{B}' = \sum_{j=1}^K \widetilde{\mathbf{b}}_j \widetilde{\mathbf{b}}_j'$. It can be formally proved that

$$\|\mathbf{B}\mathbf{B}' - \sum_{j=1}^K \lambda_j \xi_j \xi_j'\|_{\max} = O(p^{-1/2}),$$

which can be understood as the (asymptotic) identification of the decomposition $\boldsymbol{\Sigma} = \mathbf{B}\mathbf{B}' + \boldsymbol{\Sigma}_u$.

Therefore, we conclude that the first K eigenvalues of $\boldsymbol{\Sigma}$ are very spiked, whereas the remaining eigenvalues are either bounded or grow slowly. In addition, both the latent factors and loadings can be approximated using the eigenvalues and eigenvectors of $\boldsymbol{\Sigma}$ and $\boldsymbol{Y}_t$. Finally, the covariance decomposition is asymptotically identified as the dimension increases. This builds the connection between the PCA and high-dimensional factor models.

6.3. The POET estimator

Fan et al. (2013) proposed a non-parametric estimator of $\boldsymbol{\Sigma}$ when the factors are unobservable, called POET (Principal Orthogonal complEment Thresholding). To motivate their estimator, note that from the discussions of the previous subsection, heuristically we have

$$\sum_{j=1}^K \widetilde{\mathbf{b}}_j \widetilde{\mathbf{b}}_j' \approx \sum_{j=1}^K \lambda_j \boldsymbol{\xi}_j \boldsymbol{\xi}_j'.$$

In addition, $\boldsymbol{\Sigma}$ has the spectral decomposition $\boldsymbol{\Sigma} = \sum_{j=1}^p \lambda_j \boldsymbol{\xi}_j \boldsymbol{\xi}_j'$ and the factor decomposition $\boldsymbol{\Sigma} = \mathbf{B}\mathbf{B}' + \boldsymbol{\Sigma}_u$. Therefore,

$$\boldsymbol{\Sigma}_u \approx \sum_{j=K+1}^p \lambda_j \boldsymbol{\xi}_j \boldsymbol{\xi}_j'.$$

Under the conditional sparsity assumption, $\sum_{j=K+1}^p \lambda_j \boldsymbol{\xi}_j \boldsymbol{\xi}_j'$ is approximately a sparse matrix. One can then estimate $\boldsymbol{\Sigma}_u$ by thresholding the sample analogue of $\sum_{j=K+1}^p \lambda_j \boldsymbol{\xi}_j \boldsymbol{\xi}_j'$.

Specifically, the POET estimator is defined as follows. Let $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \cdots \geq \widehat{\lambda}_p$ be the ordered eigenvalues of the sample covariance matrix $\mathbf{S}$, and $\widehat{\boldsymbol{\xi}}_1, \ldots, \widehat{\boldsymbol{\xi}}_p$ be the corresponding eigenvectors. Then the sample covariance has the following spectral decomposition:

$$\mathbf{S} = \sum_{i=1}^{K} \widehat{\lambda}_i \widehat{\xi}_i \widehat{\xi}_i' + S_u.$$

Here, $\mathbf{S}_u = \sum_{k=K+1}^{p} \widehat{\lambda}_k \widehat{\xi}_k \widehat{\xi}_k' = (s_{u,ij})$, called the principal orthogonal complement. We apply the generalized thresholding rule on $\mathbf{S}_u$. Define

$$\widehat{\boldsymbol{\Sigma}}_u = (\widehat{\sigma}_{u,ij})_{p\times p}, \quad \widehat{\sigma}_{u,ij} = \begin{cases} s_{u,ii}, & i = j; \\ h(s_{u,ij}; \widetilde{\omega}_{T,ij}), & i \neq j. \end{cases}$$

For instance, the entry dependent thresholding sets $\widetilde{\omega}_{T,ij} = \sqrt{s_{u,ii} s_{u,jj}}\widetilde{\omega}_T$. Importantly, $\widetilde{\omega}_T$ is different from before when the factors are latent, and should be set to

$$\widetilde{\omega}_T = C\Big(\sqrt{\frac{\log p}{T}} + \frac{1}{\sqrt{p}}\Big).$$

It was then shown by Fan et al. (2013) that

$$\max_{i,j\leq p} |s_{u,ij} - \sigma_{u,ij}| = O_P(\widetilde{\omega}_T).$$

The extra term $1/\sqrt{p}$ in $\widetilde{\omega}_T$ is the price paid for not knowing the latent factors, and is negligible when p grows faster than T. Intuitively, when the dimension is sufficiently large, the latent factors can be extracted accurately enough as if they were observable.

The POET estimator of $\boldsymbol{\Sigma}$ is then defined as

$$\widehat{\boldsymbol{\Sigma}}_K = \sum_{i=1}^{K} \widehat{\lambda}_i \widehat{\boldsymbol{\xi}}_i \widehat{\boldsymbol{\xi}}_i' + \widehat{\boldsymbol{\Sigma}}_u. \tag{6.4}$$

This estimator is optimization-free and is very easy to compute.

Note that $\widehat{\boldsymbol{\Sigma}}_K$ requires knowledge of K, which is the number of factors and is practically unknown. There is a large body of literature on determining the number of factors and many consistent estimators have been proposed, such as Bai and Ng (2002), Alessi et al. (2010), Hallin and Liška (2007) and Ahn and Horenstein (2013). In addition, numerical studies in Fan et al. (2013) showed that the covariance estimator is robust to overestimating K. Therefore, in practice, we can also choose a relatively large number for K even if it is not a consistent estimator of the true number of factors. In the following, we suppress the subscript K, and simply write $\widehat{\boldsymbol{\Sigma}}$ as the POET estimator.

6.4. Asymptotic Results

Under the conditional sparsity assumption and some regularity conditions, Fan et al. (2013) showed that when $\widetilde{\omega}_T^{1-q} m_{u,p} \to 0$, we have

$$\|\widehat{\boldsymbol{\Sigma}}_u - \boldsymbol{\Sigma}_u\|_2 = O_P(\widetilde{\omega}_T^{1-q} m_{u,p}), \quad \|\widehat{\boldsymbol{\Sigma}}_u^{-1} - \boldsymbol{\Sigma}_u^{-1}\|_2 = O_P(\widetilde{\omega}_T^{1-q} m_{u,p}).$$

However, the problem of bad rate of convergence for $\mathbf{\Sigma}$ is still present, because the first K eigenvalues of $\mathbf{\Sigma}$ grow with p. We can further illustrate this point in the following example (taken from Fan et al., 2013).

EXAMPLE 6.1. Consider an ideal case where we know the spectrum except for the first eigenvector of $\mathbf{\Sigma}$, and assume that the largest eigenvalue $\lambda_1 \geq cp$ for some $c > 0$. Let $\widehat{\boldsymbol{\xi}}_1$ be the estimated first eigenvector and define the covariance estimator $\widehat{\mathbf{\Sigma}} = \lambda_1 \widehat{\boldsymbol{\xi}}_1 \widehat{\boldsymbol{\xi}}_1' + \sum_{j=2}^{p} \lambda_j \boldsymbol{\xi}_j \boldsymbol{\xi}_j'$. Assume that $\widehat{\boldsymbol{\xi}}_1$ is a good estimator in the sense that $\|\widehat{\boldsymbol{\xi}}_1 - \boldsymbol{\xi}_1\|^2 = O_p(T^{-1})$. However,

$$\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_2 = \|\lambda_1(\widehat{\boldsymbol{\xi}}_1 \widehat{\boldsymbol{\xi}}_1' - \boldsymbol{\xi}_1 \boldsymbol{\xi}_1')\|_2 = \lambda_1 O_p(\|\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}\|_2) = O_p(\lambda_1 T^{-1/2}),$$

which can diverge when $T = O(p^2)$.

Similar to the case of observable factors, we can estimate the precision matrix with a satisfactory rate under the operator norm. The intuition still follows from the fact that $\mathbf{\Sigma}^{-1}$ has bounded eigenvalues. Indeed, Fan et al. (2013) showed that $\widehat{\mathbf{\Sigma}}^{-1}$ has the same rate of convergence as that of $\widehat{\mathbf{\Sigma}}_u^{-1}$. A more refined theoretical analysis of the POET estimator is provided in Fan et al. (2015), which is rooted from recent progress of the random matrix theory. See Fan et al. (2015), and references therein. In addition, recent development of matrix theory can also be found in Johnstone (2001), Paul (2007), Jung and Marron (2009) and Vershynin (2012), among others.

7. STRUCTURED FACTOR MODELS

7.1. Motivations

In the usual asymptotic analysis for factor models, accurate estimations of the space spanned by the eigenvectors of $\mathbf{\Sigma}$ require a relatively large T. In particular, the factor loadings and factors themselves can be estimated no faster than $O_P(T^{-1/2})$. However, data sets of large sample size are not always available. Often we face the high-dimensional low sample size (HDLSS) scenario, as described in Jung and Marron (2009). This is particularly the case in financial studies of asset returns, as their dynamics can vary substantially over a longer time horizon. Therefore, to capture the current market condition, financial analysts wish to use a short time horizon to infer as well as possible the risk factors, as well as their associated loading matrix. To achieve this, we need additional data covariate information and modelling of the factor loadings.

Suppose that there is a d-dimensional vector of observed covariates associated with the ith variable, $\boldsymbol{X}_i = (X_{i1}, \cdots, X_{id})$, which is independent of u_{it}. For instance, in financial applications, $\boldsymbol{X}_i$ can be a vector of firm-specific characteristics (market capitalization, price-earning ratio, etc.); in health studies, $\boldsymbol{X}_i$ can be individual characteristics (e.g., age, weight, clinical and genetic information). To incorporate the information carried by the observed characteristics, Connor and Oliver (2007) and Connor et al. (2012) model explicitly the loading matrix as a function of covariates $\boldsymbol{X}$. This reduces significantly the number of parameters in the loading. Specifically, they proposed and studied the following semi-parametric factor model:

$$Y_{it} = \sum_{k=1}^{K} g_k(\boldsymbol{X}_i) f_{kt} + u_{it}, \quad i = 1, \cdots, p, t = 1, \cdots, T. \tag{7.1}$$

Here, $g_k(\boldsymbol{X}_i)$ is an unknown function of the characteristics and they assume further the additive modelling

$$g_k(\boldsymbol{X}_i) = g_{k1}(X_{i1}) + \cdots + g_{kd}(X_{id}). \tag{7.2}$$

In model (7.2), the factor loading of the kth factor on the ith variable is $b_{ik} = g_k(\boldsymbol{X}_i)$. Additional works on semi-parametric factor models include, e.g., Park et al. (2009) and Song et al. (2014).

Fan et al. (2016b) recognized that the above semi-parametric model (7.1) might be restrictive for applications, as we do not expect that the covariates capture completely the factor loadings. They extend the model to the following more flexible semi-parametric mixed effect model:

$$Y_{it} = \sum_{k=1}^{K} [g_k(\boldsymbol{X}_i) + \gamma_{ik}] f_{kt} + u_{it}, \quad i = 1, \cdots, p, t = 1, \cdots, T. \tag{7.3}$$

Here, γ_{ik} is an unobservable random component with mean zero, representing the component of the factor loading b_{ik} that cannot be explained by the covariates $\boldsymbol{X}_i$. They developed econometric techniques to test the model specifications (7.1) and (7.3). Their empirical results, using the returns of the components of the S&P500 index and four exogenous variables (size, value, momentum and volatility) as in Connor et al. (2012), provide stark evidence that model (7.1) cannot be validated empirically, whereas (7.3) is consistent with the empirical data.

7.2. Projected PCA

The basic idea of projected PCA is to smooth the observations $\{Y_{it}\}_{i=1}^p$ for each given day t against its associated covariates $\{\boldsymbol{X}_i\}_{i=1}^p$. More specifically, let $\{\widehat{Y}_{it}\}_{i=1}^p$ be the fitted value after running a regression of $\{Y_{it}\}_{i=1}^p$ against $\{\boldsymbol{X}_i\}_{i=1}^p$ for each given t. The regression model can be the usual linear regression or additive regression model (7.2). This results in a smooth or projected observation matrix $\widehat{\mathbf{Y}}_t$, which will also be denoted by $\mathbf{PY}_t$. The projected PCA is then to run PCA based on the projected data $\{\widehat{\mathbf{Y}}_t\}_{t=1}^T$.

To provide the rationale behind this idea, we now generalize model (7.3) further to illustrate the idea behind the projected PCA. Specifically, consider the factor model

$$\mathbf{Y} = \mathbf{BF}' + \mathbf{U},$$

where $\mathbf{Y}$ and $\mathbf{U}$ are $p \times T$ matrices of y_{it} and u_{it}. Suppose that there is a d-dimensional vector of observed covariates associated with the ith variable, $\boldsymbol{X}_i = (X_{i1}, \cdots, X_{id})$, which is independent of u_{it}. For a pre-determined J, let $\phi_1, \ldots, \phi_J$ be a set of basis functions. Let $\phi(\boldsymbol{X}_i)' = (\phi_1(X_{i1}), \ldots., \phi_J(X_{i1}), \ldots, \phi_J(X_{id}))$ and $\Phi(\mathbf{X}) = (\phi(\boldsymbol{X}_1), \ldots, \phi(\boldsymbol{X}_p))'$ be a $p \times (Jd)$ matrix of the sieve-transformed $\mathbf{X}$. Then the projection matrix on the space spanned by $\mathbf{X} = (\boldsymbol{X}_1, \ldots, X_p)$ can be taken as

$$\mathbf{P} = \Phi(\mathbf{X})(\Phi(\mathbf{X})'\Phi(\mathbf{X}))^{-1}\Phi(\mathbf{X})'.$$

This corresponds to modelling $g_k(\boldsymbol{X}_i)$ in (7.3) by the additive model (7.2) and approximating each term using the series expansion. The projected data $\mathbf{PY}$ is the fitted value of the additive model (7.2) with basis functions $\phi_1, \ldots, \phi_J$:

$$Y_{it} = \sum_{k=1}^{K} \Big(\sum_{j=1}^{J} \beta_{jk,t} \phi_j(X_{ik}) \Big) + \varepsilon_{it}, \quad i = 1, \cdots, p; t = 1, \cdots, T.$$

The design matrix does not vary with t, neither does the projection matrix $\mathbf{P}$.

We make the following key assumptions.

ASSUMPTION 7.1. *(a) Pervasiveness – with probability approaching one, all the eigenvalues of* $(1/p)(\mathbf{PB})'\mathbf{PB}$ *are bounded away from both zero and infinity as* $p \to \infty$*; (b) orthogonality –* $\mathbb{E}(u_{it}|X_{i1}, \ldots, X_{id}) = 0$*, for all* $i \leq p, t \leq T$.

The above assumptions require that the strengths of the loading matrix should be as strong after the projection, i.e. $\mathbf{B}$ should depend on $\mathbf{X}$ so that $\mathbf{PB} \neq 0$. Assumption 7.1(b) implies that if we apply $\mathbf{P}$ to both sides of $\mathbf{Y} = \mathbf{BF}' + \mathbf{U}$, then

$$\mathbf{PY} \approx \mathbf{PBF}',$$

where $\mathbf{PU} \approx 0$ due to the orthogonality condition. Hence, the projection removes the noise in the factor model (noise are smoothed out). In addition, for the purpose of normalizations, we assume $\mathrm{Cov}(\boldsymbol{f}_t) = \mathbf{I}_K$, and that $(\mathbf{PB})'\mathbf{PB}$ is a diagonal matrix.

We now describe the rationale of the projected PCA. For simplicity, we ignore the effect of $\mathbf{PU}$. Let us consider the $p \times p$ covariance matrix of the projected data $\mathbf{PY}$. The previous discussions show that $(1/T)\mathbf{PY}(\mathbf{PY})' \approx \mathbf{PB}(\mathbf{PB})'$. Because $(\mathbf{PB})'\mathbf{PB}$ is a diagonal matrix, the columns of $\mathbf{PB}$ are the eigenvectors of the $p \times p$ matrix $(1/T)\mathbf{PY}(\mathbf{PY})'$, up to a factor $\sqrt{p}$. Next, consider the $T \times T$ matrix $(1/T)(\mathbf{PY})'\mathbf{PY} \approx (1/T)\mathbf{F}(\mathbf{PB})'(\mathbf{PB})\mathbf{F}'$. This implies

$$\frac{1}{T}(\mathbf{PY})'\mathbf{PYF} \approx \mathbf{F}(\mathbf{PB})'(\mathbf{PB}).$$

Still, by the diagonality of $(\mathbf{PB})'\mathbf{PB}$, we infer that the columns of $\mathbf{F}$ are approximately the eigenvectors of the $T \times T$ sample covariance matrix $(1/T)(\mathbf{PY})'\mathbf{PY}$, up to a factor $\sqrt{T}$. In addition, because the diagonal elements of $(\mathbf{PB})'\mathbf{PB}$ grow fast as the dimensionality diverges, the corresponding eigenvalues are asymptotically the first K leading eigenvalues of $(1/T)(\mathbf{PY})'\mathbf{PY}$. This motivates the so-called projected PCA (Fan et al., 2016b), a new framework of estimating the parameters for factor analysis in the presence of a known space $\mathcal{X}$. The projected PCA can be more accurate than the usual PCA in the HDLSS scenario. It applies the PCA to the projected data (smoothed data) $\mathbf{PY}$.

Let $\widetilde{\mathbf{V}}$ be a $T \times K$ matrix, whose columns are the eigenvectors of the $T \times T$ matrix $(1/T)\mathbf{Y}'\mathbf{PY}$ corresponding to the largest K eigenvalues. Following the previous discussions, we respectively estimate the projected loading matrix $\mathbf{PB}$ and latent factors $\mathbf{F}$ by

$$\widetilde{\mathbf{G}}(\mathbf{X}) = \frac{1}{T}\mathbf{PY}\widetilde{\mathbf{F}}, \quad \widetilde{\mathbf{F}} = \sqrt{T}\widetilde{\mathbf{V}}.$$

A nice feature of the projected PCA is that the consistency of latent factors is achieved even when the sample size T is finite so long as p goes to infinity, as shown in Fan et al. (2016b). Thus, it is particularly appealing in the HDLSS context. Intuitively, the major approximation error $\mathbf{PU}$ vanishes without requiring a large sample size T. This implies the consistency under a finite T. See Fan et al. (2016b) for more detailed discussions on this aspect.

7.3. Semi-parametric factor models

In the model (7.3), let $\mathbf{G}(\mathbf{X})$ and $\mathbf{\Gamma}$ denote the $p \times K$ matrices of $g_k(\boldsymbol{X}_i)$ and γ_{ij}, respectively. Then the matrix form of the model can be written as

$$\mathbf{Y} = [\mathbf{G}(\mathbf{X}) + \mathbf{\Gamma}]\mathbf{F}' + \mathbf{U}.$$

So the model assumes that the loading matrix can be decomposed into two parts: the part that can be explained by $\mathbf{X}$ and the part that cannot. To deal with the curse of dimensionality, we assume $g_k(\cdot)$ to be additive: $g_k(\boldsymbol{X}_i) = \sum_{l=1}^{d} g_{kl}(X_{il})$, with $d = \dim(\boldsymbol{X}_i)$.

Applying the projected PCA on to the semi-parametric factor model, Fan et al. (2016b) showed that as $p, J \to \infty$, T may either grow or stay constant,

$$\frac{1}{\sqrt{T}}\|\widetilde{\mathbf{F}} - \mathbf{F}\|_2 = O_P\Big(\frac{1}{p}\Big), \quad \frac{1}{\sqrt{p}}\|\widetilde{\mathbf{G}}(\mathbf{X}) - \mathbf{G}(\mathbf{X})\|_2 = O_P\Big(\frac{1}{(p\min\{T, p\})^{1/2-1/(2\kappa)}}\Big),$$

where κ is the degree of smoothness constant for $g_k(\cdot)$. Clearly under the high dimensionality, the rate of convergence is fast even if T is finite. We refer the readers to Fan et al. (2016b) for more detailed discussions of the impacts of improved rates of convergence in factor models.

8. DISCUSSION AND OUTLOOK

This paper introduces several recent developments on estimating large covariance and precision matrices. We focus on two general approaches: the rank-based method and the factor-model-based method. We also extend the usual factor model to a projected PCA set-up, and show that the newly introduced projected PCA is appealing in the HDLSS scenario. Such an approach has attracted growing attention in the recent literature on high-dimensional PCA; see, e.g., Jung and Marron (2009), Shen et al. (2013a), Shen et al. (2013b) and Ahn et al. (2007). In addition, we introduce the rank-based approaches, including the EPIC and EC2 estimators, for estimating large precision and covariance matrices under the elliptical distribution family. These rank-based methods are robust to heavy-tailed data and achieve the nearly optimal rates of convergence under the spectral norm. The ellipticity assumptions can further be removed by using RA-mean estimators in Section 4.1.3 and the resulting covariance estimators can be further regularized.

A promising future direction is to combine the factor-model-based approach, rank-based analysis and RA-mean-based method into an integrated framework. For instance, consider the factor model

$$\boldsymbol{Y}_t = \mathbf{B}\boldsymbol{f}_t + \boldsymbol{u}_t$$

with factors $\{\boldsymbol{f}_t\}$. Instead of applying PCA or projected PCA on the sample covariance matrix, we can also apply these methods on the RA-mean based covariance matrix estimator in Section 4.1.3. If the idiosyncratic components $\boldsymbol{u}_t$ are heavy-tailed but follow the elliptical distribution, we can apply PCA or projected PCA on the rank-based covariance matrix estimator $\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{D}}\widehat{\mathbf{R}}\widehat{\mathbf{D}}$, where $\widehat{\mathbf{R}}$ and $\widehat{\mathbf{D}}$ are defined in Sections 4.1.1 and 4.1.2, respectively.

When the factors are observable, the residuals should be obtained by estimating $\mathbf{B}$. The robust regression estimator $\widehat{\mathbf{B}}$ can be employed, using the RA-quadratic loss as in Fan et al. (2016a), which provides robust estimation for the mean regression even with asymmetric error

distributions. With the estimated $\mathbf{B}$, we set $\widehat{\boldsymbol{u}}_t = \boldsymbol{Y}_t - \widehat{\mathbf{B}}\boldsymbol{f}_t$. The final factor-based covariance estimator is then given by

$$\widehat{\boldsymbol{\Sigma}} = \widehat{\mathbf{B}}\,\widehat{\mathrm{Cov}}(\boldsymbol{f}_t)\widehat{\mathbf{B}}' + \widehat{\boldsymbol{\Sigma}}_u.$$

The resulting estimator is expected to naturally handle heavy-tailed data.

When the common factors are latent, they need to be estimated using robust PCA (i.e., applying PCA to robustly estimated covariance matrix of $\boldsymbol{Y}_t$). The theoretical properties of such hybrid estimators are left for future investigations.

REFERENCES

Ahn, S. C. and A. R. Horenstein (2013). Eigenvalue ratio test for the number of factors. *Econometrica 81*, 1203–27.

Ahn, J., J. Marron, K. M. Muller and Y-Y. Chi (2007). The high-dimension, low-sample-size geometric representation holds under mild conditions. *Biometrika 94*, 760–6.

Aıt-Sahalia, Y. and D. Xiu (2015). Using principal component analysis to estimate a high dimensional factor model with high-frequency data. Working Paper, University of Chicago.

Alessi, L., M. Barigozzi and M. Capasso (2010). Improved penalization for determining the number of factors in approximate factor models. *Statistics and Probability Letters 80*, 1806–13.

Ang, A., J. Liu and K. Schwarz (2008). Using individual stocks or portfolios in tests of factor models. Working Paper, Columbia University.

Antoniadis, A. (1997). Wavelets in statistics: a review. *Journal of the Italian Statistical Society 6*, 97–130.

Antoniadis, A. and J. Fan (2001). Regularization of wavelet approximations. *Journal of the American Statistical Association 96*, 939–67.

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

Bai, J. and K. Li (2012). Statistical analysis of factor models of high dimension. *Annals of Statistics 40*, 436–65.

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 70*, 191–221.

Banerjee, O., L. El Ghaoui and A. d'Aspremont (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research 9*, 485–516.

Bickel, P. and E. Levina (2008). Covariance regularization by thresholding. *Annals of Statistics 36*, 2577–604.

Bickel, P. J., Y. Ritov and A. B. Tsybakov (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics 37*, 1705–32.

Boivin, J. and S. Ng (2005). Understanding and comparing factor-based forecasts. *International Journal of Central Banking 1*, 117–51.

Boivin, J. and S. Ng (2006). Are more data always better for factor analysis? *Journal of Econometrics 132*, 169–94.

Breitung, J. and J. Tenhofen (2011). GLS estimation of dynamic factor models. *Journal of the American Statistical Association 106*, 1150–66.

Cai, T. and W. Liu (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association 106*, 672–84.

Cai, T. and H. Zhou (2012). Optimal rates of convergence for sparse covariance matrix estimation. *Annals of Statistics 40*, 2389–420.

Cai, T., W. Liu and X. Luo (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association 106*, 594–607.

Cai, T., Z. Ma and Y. Wu (2013). Sparse PCA: optimal rates and adaptive estimation. *Annals of Statistics 41*, 3074–110.

Campbell, J. Y., A. W.-C. Lo and A. C. MacKinlay (1996). *The Econometrics of Financial Markets*. Princeton, NJ: Princeton University Press.

Candès, E. and T. Tao (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics 35*, 2313–51.

Candès, E. J. and B. Recht (2009). Exact matrix completion via convex optimization. *Foundations of Computational Mathematics 9*, 717–72.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques 48*, 1148–85.

Chamberlain, G. and M. Rothschild (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica 51*, 1281–304.

Choi, I. (2012). Efficient estimation of factor models. *Econometric Theory 28*, 274–308.

Connor, G. and L. Oliver (2007). Semiparametric estimation of a characteristic-based factor model of stock returns. *Journal of Empirical Finance 14*, 694–717.

Connor, G., H. Matthias and L. Oliver (2012). Efficient semiparametric estimation of the Fama–French model and extensions. *Econometrica 80*, 713–54.

Davis, C. (1963). The rotation of eigenvectors by a perturbation. *Journal of Mathematical Analysis and Applications 6*, 159–73.

Donoho, D. L., I. M. Johnstone, G. Kerkyacharian and D. Picard (1995). Wavelet shrinkage: asymptopia? (with discussion). *Journal of the Royal Statistical Society*, Series B *57*, 301–69.

Doz, C., D. Giannone and L. Reichlin (2012). A quasi-maximum likelihood approach for large, approximate dynamic factor models. *Review of Economics and Statistics 94*, 1014–24.

El Karoui, N. (2010). High-dimensionality effects in the markowitz problem and other quadratic programs with linear constraints: risk underestimation. *Annals of Statistics 38*, 3487–566.

Fama, E. and K. French (1992). The cross-section of expected stock returns. *Journal of Finance 47*, 427–65.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*, 1348–60.

Fan, J. and H. Liu (2013). Statistical analysis of big data on pharmacogenomics. *Advanced Drug Delivery Reviews 65*, 987–1000.

Fan, J. and H. Peng (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics 32*, 928–61.

Fan, J., Y. Fan and J. Lv (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics 147*, 186–97.

Fan, J., Y. Feng and Y. Wu (2009). Network exploration via the adaptive LASSO and SCAD penalties. *Annals of Applied Statistics 3*, 521–41.

Fan, J., Y. Liao and M. Mincheva (2011). High dimensional covariance matrix estimation in approximate factor models. *Annals of Statistics 39*, 3320–56.

Fan, J., J. Zhang and K. Yu (2012). Vast portfolio selection with gross-exposure constraints. *Journal of the American Statistical Association 107*, 592–606.

Fan, J., Y. Liao and M. Mincheva (2013). Large covariance estimation by thresholding principal orthogonal complements (with discussion). *Journal of the Royal Statistical Society*, Series B *75*, 603–80.

Fan, J., F. Han and H. Liu (2014). Challenges of big data analysis. *National Science Review 1*, 293–314.

Fan, J., H. Liu and W. Wang (2015). Large covariance estimation through elliptical factor models. Working paper, Princeton University (arXiv:1507.08377).

Fan, J., Q. Li and Y. Wang (2016a). Estimation of high-dimensional mean regression in absence of symmetry and light-tail assumptions. Forthcoming in *Journal of Royal Statistical Society*, Series B.

Fan, J., Y. Liao and W. Wang (2016b). Projected principal component analysis in factor models. *Annals of Statistics 44*, 219–54.

Fang, K.-T., S. Kotz and K. W. Ng (1990). *Symmetric Multivariate and Related Distributions, Monographs on Statistics and Applied Probability, 36*. London: Chapman and Hall.

Forni, M. and M. Lippi (2001). The generalized dynamic factor model: representation theory. *Econometric Theory 17*, 1113–41.

Forni, M., M. Hallin, M. Lippi and L. Reichlin (2000). The generalized dynamic-factor model: identification and estimation. *Review of Economics and Statistics 82*, 540–54.

Frahm, G. and U. Jaekel (2008). Tyler's M-estimator, random matrix theory, and generalized elliptical distributions with applications to finance. Working paper, Helmut Schmidt University.

Friedman, J., T. Hastie and R. Tibshirani (2008). Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics 9*, 432–41.

Gautier, E. and A. B. Tsybakov (2011). High-dimensional instrumental variables regression and confidence sets. Working Paper, CREST Paris.

Hallin, M. and R. Liška (2007). Determining the number of factors in the general dynamic factor model. *Journal of the American Statistical Association 102*, 603–17.

Hamada, M. and E. A. Valdez (2008). CAPM and option pricing with elliptically contoured distributions. *Journal of Risk and Insurance 75*, 387–409.

Han, F. and H. Liu (2013). Optimal rates of convergence for latent generalized correlation matrix estimation in transelliptical distribution. Forthcoming in *Bernoulli*.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics 29*, 295–327.

Jung, S. and J. Marron (2009). PCA consistency in high dimension, low sample size context. *Annals of Statistics 37*, 4104–30.

Kapetanios, G. (2010). A testing procedure for determining the number of factors in approximate factor models with large datasets. *Journal of Business and Economic Statistics 28*, 397–409.

Koltchinskii, V., K. Lounici and A. B. Tsybakov (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Annals of Statistics 39*, 2302–29.

Lam, C. and J. Fan (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of Statistics 37*, 4254–78.

Lam, C. and Q. Yao (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *Annals of Statistics 40*, 694–726.

Lange, K., D. R. Hunter and I. Yang. (2000). Optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics 9*, 1–20.

Lawley, D. and A. Maxwell (1971). Factor analysis as a statistical method. *Journal of the Royal Statistical Society*, Series D *12*, 209–29.

Ledoit, O. and M. Wolf (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance 10*, 603–21.

Ledoit, O. and M. Wolf (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis 88*, 365–411.

Liu, W. and X. Luo (2012). High-dimensional sparse precision matrix estimation via sparse column inverse operator. Working paper, Brown University (arXiv:1203.3896).

Liu, H. and L. Wang (2012). TIGER: a tuning-insensitive approach for optimally estimating Gaussian graphical models. Working Paper, Department of Operations Research and Financial Engineering, Princeton University.

Liu, H., F. Han, M. Yuan, J. Lafferty and L. Wasserman (2012). High-dimensional semiparametric Gaussian copula graphical models. *Annals of Statistics 40*, 2293–326.

Liu, H., L. Wang and T. Zhao (2014). Sparse covariance matrix estimation with eigenvalue constraints. *Journal of Computational and Graphical Statistics 23*, 439–59.

Ma, Z. (2013). Sparse principal component analysis and iterative thresholding. *Annals of Statistics 41*, 772–801.

Meinshausen, N. and P. Bühlmann (2006). High dimensional graphs and variable selection with the Lasso. *Annals of Statistics 34*, 1436–62.

Mitra, R. and C.-H. Zhang (2014). Multivariate analysis of nonparametric estimates of large correlation matrices. Working paper, Rutgers University (arXiv:1403.6195).

Negahban, S. and M. J. Wainwright (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Annals of Statistics 39*, 1069–97.

Onatski, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics 168*, 244–58.

Owen, J. and R. Rabinovitch (1983). On the class of elliptical distributions and their applications to the theory of portfolio choice. *Journal of Finance 38*, 745–52.

Park, B. U., E. Mammen, W. Härdle and S. Borak (2009). Time series modelling with semiparametric factor dynamics. *Journal of the American Statistical Association 104*, 284–98.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica 17*, 1617–42.

Pourahmadi, M. (2013). *High-Dimensional Covariance Estimation: with High-Dimensional Data*. New York, NY: Wiley.

Qi, H. and D. Sun (2006). A quadratically convergent Newton method for computing the nearest correlation matrix. *SIAM Journal on Matrix Analysis and Applications 28*, 360–85.

Rigollet, P. and A. Tsybakov (2012). Estimation of covariance matrices under sparsity constraints. Working Paper, Department of Operations Research and Financial Engineering, Princeton University (arXiv:1205.1210).

Rothman, A. J., P. J. Bickel, E. Levina and J. Zhu (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics 2*, 494–515.

Rothman, A. J., E. Levina and J. Zhu (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association 104*, 177–86.

Shen, X., W. Pan and Y. Zhu (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association 107*, 223–32.

Shen, D., H. Shen and J. Marron (2013a). Consistency of sparse PCA in high dimension, low sample size contexts. *Journal of Multivariate Analysis 115*, 317–33.

Shen, D., H. Shen, H. Zhu and J. Marron (2013b). Surprising asymptotic conical structure in critical sample eigen-directions. Working Paper, University of North Carolina.

Song, S., W. K. Härdle and Y. Ritov (2014). Generalized dynamic semi-parametric factor models for high-dimensional non-stationary time series. *Econometrics Journal 17*, S101–S131.

Stock, J. H. and M. W. Watson (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association 97*, 1167–79.

Stoer, J., R. Bulirsch, R. Bartels, W. Gautschi and C. Witzgall (1993). *Introduction to Numerical Analysis*, *Volume 2*. New York, NY: Springer.

Sun, T. and C.-H. Zhang (2013). Sparse matrix inversion with scaled Lasso. *Journal of Machine Learning Research 14*, 3385–418.

Tokuda, T., B. Goodrich, I. Van Mechelen, A. Gelman and F. Tuerlinckx (2011). Visualizing distributions of covariance matrices. Working Paper, Columbia University.

Tsai, H. and R. S. Tsay (2010). Constrained factor models. *Journal of the American Statistical Association 105*, 1593–605.

Vanderbei, R. (2008). *Linear Programming, Foundations and Extensions*. Berlin: Springer.

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In Y. C. Eldar and G. Kutyniok (Eds.), *Compressed Sensing, Theory and Applications*, 210–68. Cambridge: Cambridge University Press.

Wainwright, M. (2009). Sharp thresholds for high dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming. *IEEE Transactions on Information Theory 55*, 2183–201.

Wang, Y. and J. Zou (2010). Vast volatility matrix estimation for high-frequency financial data. *Annals of Statistics 38*, 943–78.

Wegkamp, M. and Y. Zhao (2013). Adaptive estimation of the copula correlation matrix for semiparametric elliptical copulas. Forthcoming in *Bernoulli*.

Wille, A., P. Zimmermann, E. Vranová, A. Fürholz, O. Laule, S. Bleuler, L. Hennig, A. Prelić, P. von Rohr, L. Thiele, E. Zitzler, W. Gruissem and P. Bühlmann (2004). Sparse graphical Gaussian modeling of the isoprenoid gene network in *Arabidopsis thaliana*. *Genome Biology 5*: R92.

Wu, W. B. and M. Pourahmadi (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika 90*, 831–44.

Xue, L. and H. Zou (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Annals of Statistics 40*, 2541–71.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research 11*, 2261–86.

Yuan, M. and Y. Lin (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika 94*, 19–35.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics 38*, 894–942.

Zhao, T. and H. Liu (2014). Calibrated precision matrix estimation for high-dimensional elliptical distributions. *IEEE Transactions on Information Theory 60*, 7874–87.

Zhao, P. and B. Yu (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research 7*, 2541–63.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association 101*, 1418–29.