An Overview on the Estimation of Large Covariance and Precision Matrices

Jianqing Fan*, Yuan Liao‡ and Han Liu*

*Department of Operations Research and Financial Engineering, Princeton University

‡ Department of Mathematics, University of Maryland

October 5, 2017

Abstract

Estimating large covariance and precision matrices are fundamental in modern multivariate analysis. The problems arise from statistical analysis of large panel economics and finance 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 estimating large covariance and precision matrices. We focus on two general approaches: rank based method and factor model based method. Theories and applications of both approaches are presented. These methods are expected to be widely applicable to analysis of economic and financial data.

Keywords: High-dimensionality, graphical model, approximate factor model, principal components, sparse matrix, low-rank matrix, thresholding, heavy-tailed, elliptical distribution, rank based methods.

1 Introduction

Estimating large covariance and precision (inverse covariance) matrices becomes fundamental problems in modern multivariate analysis, which find applications in many

*Address: Department of ORFE, Sherrerd Hall, Princeton University, Princeton, NJ 08544, USA, e-mail: *jqfan@princeton.edu*, *yuanliao@umd.edu*, *hanliu@princeton.edu*. Fan's research was supported by National Institutes of Health grants R01-GM072611 and R01GM100474-01 and National Science Foundation grants DMS-1206464 and DMS-1308566.

fields, ranging from economics and finance to biology, social networks, and health sciences (Fan et al., 2014a). 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 massive amount of estimation errors can make considerable adverse impacts on the estimation accuracy. Therefore, estimating large covariance and precision matrices attracts rapidly growing research attentions in 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, namely, many entries are zero or nearly so (Bickel and Levina, 2008; Lam and Fan, 2009; El Karoui, 2010; 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 for instance, Banerjee et al. (2008); Yuan and Lin (2007); Friedman et al. (2008); Rothman et al. (2008). To further reduce the estimation bias, Lam and Fan (2009); Shen et al. (2012) proposed non-convex penalties for sparse precision matrix estimation and studied their theoretical properties. For more general theory on penalized likelihood methods, see Fan and Li (2001); Fan and Peng (2004); Zou (2006); Zhao and Yu (2006); Bickel et al. (2009); Wainwright (2009).

The literature has been further expanded into robust estimation based on regularized rank-based approaches (Liu et al., 2012a; 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. It is particularly appealing for analysis of financial data. The literature includes, for instance, Han and Liu (2013); Wegkamp and Zhao (2013); Mitra and Zhang (2014), etc. The heavy-tailed data are often modeled by the elliptical distribution family, which has been widely used for financial data analysis. See Owen and Rabinovitch (1983); Hamada and Valdez (2004) and Frahm and Jaekel (2008).

In addition, in many applications the sparsity property is not directly applicable. For example, financial returns depend on the equity market risks, housing prices depend on the economic health, gene expressions can be stimulated by cytokines, among others. Due to the presence of common factors, it is unrealistic to assume that many outcomes are uncorrelated. A natural extension is the *conditional sparsity*, namely, conditional on the common factors, the covariance matrix of the remaining components of the outcome variables is sparse. In order to do so, we consider 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 excessive returns of financial assets over time, and has been found to produce good out-

of-sample forecast for macroeconomic variables (Boivin and Ng, 2005; 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, for instance, Stock and Watson (2002); Bai (2003); Bai and Ng (2002); Fan et al. (2008); Breitung and Tenhofen (2011); Onatski (2012); Lam and Yao (2012); Fan et al. (2013), 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, for example, Candès and Recht (2009); Koltchinskii et al. (2011); Fan et al. (2011); Negahban and Wainwright (2011); Cai et al. (2013); Ma (2013).

In this paper, we provide a selective review of several recent developments on estimating large covariance and precision matrices. We focus on two general approaches: rank-based method and 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, e.g., the shrinkage method (Ledoit and Wolf, 2003, 2004). We refer to Fan and Liu (2013), Pourahmadi (2013) and the references therein for reviews of other commonly used methods.

This paper is organized as follows. Section 2 presents methods of estimating sparse covariance matrices. Section 3 reviews methods of estimating sparse precision matrices. Section 4 discusses robust covariance and precision matrix estimations using rank-based estimators. Sections 5 and 6 respectively presents factor models based method, respectively in the cases of observable and unobservable factors. Section 7 introduces the structured factor model. Finally, Section 8 provides further discussions.

Let $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ respectively denote the minimum and maximum eigenvalues of $\mathbf{A}$. Let $\psi_{\max}(\mathbf{A})$ be the largest singular value of $\mathbf{A}$. We shall 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 $\lambda^{1/2}_{\max}(\mathbf{A}'\mathbf{A})$ and $\mathrm{tr}^{1/2}(\mathbf{A}'\mathbf{A})$. Throughout this paper, we shall use p and T to respectively denote the dimension of the covariance matrix of interest, and the sample size. Let $\mathbf{v} = (v_1, \ldots, v_p)' \in \mathbb{R}^p$ be a real valued vector, we define the vector norms: $\|\mathbf{v}\|_1 = \sum_{j=1}^p |v_j|$, $\|\mathbf{v}\|_2^2 = \sum_{j=1}^p v_j^2$, $\|\mathbf{v}\|_\infty = \max_{1\le j\le p} |v_j|$. Let $\mathcal{S}$ be a subspace of $\mathbb{R}^p$, we use $\mathbf{v}_{\mathcal{S}}$ to denote the projection of $\mathbf{v}$ onto $\mathcal{S}$: $\mathbf{v}_{\mathcal{S}} = \mathrm{argmin}_{\mathbf{u}\in\mathcal{S}} \|\mathbf{u}-\mathbf{v}\|_2^2$. We also define the orthogonal complement of $\mathcal{S}$ as $\mathcal{S}^\perp = \left\{\mathbf{u} \in \mathbb{R}^p \;\middle|\; \mathbf{u}'\mathbf{v} = 0, \text{ for any } \mathbf{v} \in \mathcal{S}\right\}$. 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 $\mathbf{A}_{*j} = (\mathbf{A}_{1j}, ..., \mathbf{A}_{pj})'$ and $\mathbf{A}_{k*} = (\mathbf{A}_{k1}, ..., \mathbf{A}_{kp})'$ denote the j^{th} column and k^{th} row of $\mathbf{A}$ in vector forms, we define the matrix norms: $\|\mathbf{A}\|_1 = \max_j \|\mathbf{A}_{*j}\|_1$, $\|\mathbf{A}\|_\infty = \max_k \|\mathbf{A}_{k*}\|_1$, $\|\mathbf{A}\|_{\max} = \max_j \|\mathbf{A}_{*j}\|_\infty$. We also define matrix elementwise (pseudo-) norms: $\|\mathbf{A}\|_{1,\text{off}} = \sum_{j\neq k} |\mathbf{A}_{jk}|$ and $\|\mathbf{A}\|_{\infty,\text{off}} = \max_{j\neq k} |\mathbf{A}_{jk}|$. 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 matrix

Let Y_{it} be the observed data for the i^{th} $(i = 1, ..., p)$ individual at time $t = 1, ..., T$ (or the t^{th} observation for the i^{th} variable). We are interested in estimating the $p \times p$ covariance matrix $\mathbf{\Sigma} = (\sigma_{ij})_{p\times p}$ of $\boldsymbol{Y}_t = (Y_{1t}, ..., Y_{pt})'$, 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 $q \geq 0$, so that the following defined quantity

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

is either bounded or grow 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 the thresholding, which sets small estimated elements to zero (Bickel and Levina, 2008). Let s_{ij} be the $(i, j)^{th}$ element of $\mathbf{S}$. For a pre-specified thresholding value ω_T, define

$$\widehat{\mathbf{\Sigma}} = (\widehat{\sigma}_{ij})_{p\times p}, \quad \widehat{\sigma}_{ij} = \begin{cases} s_{ij}, & \text{if } i = j \\ s_{ij}1\{|s_{ij}| > \omega_T\}, & \text{if } i \neq j \end{cases}. \tag{2}$$

The thresholding value should dominate the maximum estimation error $\max_{i\neq j}|s_{ij}-\sigma_{ij}|$. When the data are 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.

The advantage of thresholding is that it avoids estimating small elements so that noise does not accumulate. The decision of 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

$$\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_2 = O_P(m_p\omega_T^{1-q}) \text{ and } \|\widehat{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}^{-1}\|_2 = O_P(m_p\omega_T^{1-q}), \tag{3}$$

where m_p and q are as defined in (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 $O_P(\sqrt{\frac{\log p}{T}})$ if m_p is bounded. Since each element in the covariance matrix can be estimated with an error of order $O_P(T^{-1/2})$, it hence only costs us a $\log(p)$ factor to learn the unknown locations of the non-zero elements.

2.2 Adaptive thresholding and entry-dependent thresholding

The simple thresholding (2) does not take the varying scales of the marginal standard deviations into account. One way to account 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):

$$\widehat{\mathbf{\Sigma}} = (\widehat{\sigma}_{ij})_{p\times p}, \quad \widehat{\sigma}_{ij} = \begin{cases} s_{ij}, & \text{if } i = j \\ s_{ij}1\{|s_{ij}|/\mathrm{SE}(s_{ij}) > \omega_T\}, & \text{if } i \neq j \end{cases}, \tag{4}$$

where $\mathrm{SE}(s_{ij})$ is the estimated standard error of s_{ij}.

A simpler method to take the scale into account is to directly apply thresholding on the correlation matrix. Let $\mathbf{R} = \mathrm{diag}(\mathbf{S})^{-1/2}\mathbf{S}\mathrm{diag}(\mathbf{S})^{-1/2} = (r_{ij})_{p\times p}$ 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

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

In particular, when $\omega_T = 0$, it is exactly the sample covariance matrix since no thresh-

olding 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 since 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) and (4) are based on a simple thresholding rule, known as the *hard-thresholding*. In regression and wavelet shrinkage contexts (see, for example, Donoho et al. (1995)), hard thresholding performs worse than some more flexible regularization methods, such as the soft-thresholding and the smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), which combine thresholding with shrinkages. The estimates resulting from such shrinkage typically are continuous functions of the maximum likelihood estimates (under Gaussianity), a desirable property that is not shared by the hard thresholding method.

Therefore, the *generalized thresholding* rules of Antoniadis and Fan (2001) can be applied to estimating large covariance matrices. The 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

$$(i) \quad |h(z,\omega_T)| \leq |z|; \qquad (ii) \quad h(z;\omega_T) = 0 \text{ for } |z| \leq \omega_T; \qquad (iii) \quad |h(z;\omega_T) - z| \leq \omega_T.$$

There are a number of useful thresholding functions that are commonly used in the literature. For instance, the soft-thresholding takes $h(z;\omega_T) = \text{sgn}(z)(|z| - \omega_T)_+$, where $(x)_+ = \max\{x, 0\}$. Moreover, the SCAD thresholding is a compromise between hard and soft thresholding, whose amount of shrinkage decreases as $|z|$ increases and hence results in a nearly unbiased estimation. Another example is the MCP thresholding, proposed by Zhang (2010).

We can then define a generalized thresholding covariance estimator:

$$\widehat{\mathbf{\Sigma}} = (\widehat{\sigma}_{ij})_{p\times p}, \quad \widehat{\sigma}_{ij} = \begin{cases} s_{ij}, & \text{if } i = j \\ h(s_{ij};\omega_T), & \text{if } i \neq j \end{cases}. \tag{5}$$

Note that this admits the hard-thresholding estimator (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 $h(s_{ij}, \text{SE}(s_{ij})\omega_T)$ 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, it is

shown by Rothman et al. (2009) that the use of generalized thresholding rules does not affect the rate of convergence in (3), but it increases the family of shrinkages.

2.4 Positive definiteness

If the covariance matrix is sparse, it then follows from (3) that the thresholding estimator $\widehat{\mathbf{\Sigma}}$ is asymptotically positive definite. On the other hand, 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 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 $\omega_T = C\sqrt{\frac{\log p}{T}}$. 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 decreases the choice of C until it reaches

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

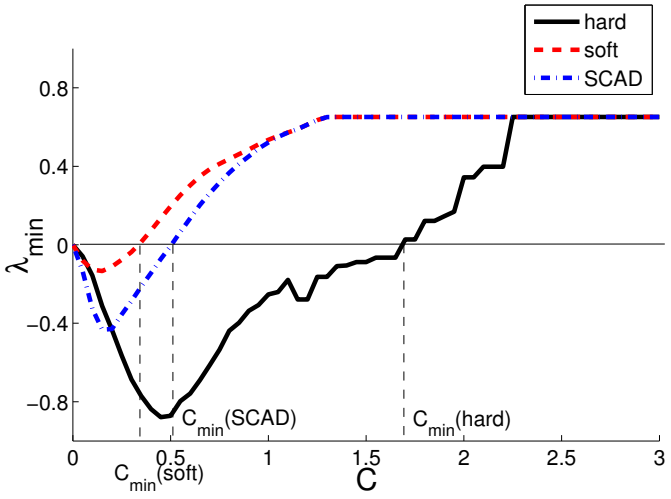
Thus, $C_{\min}$ is well defined and for all $C > C_{\min}$, $\widehat{\mathbf{\Sigma}}(C)$ is positive definite under finite sample. We can obtain $C_{\min}$ by solving $\lambda_{\min}(\widehat{\mathbf{\Sigma}}(C)) = 0, C \neq 0$. Figure 1 plots the minimum eigenvalue of $\widehat{\mathbf{\Sigma}}(C)$ as a function of C for a random simple 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. 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 constraint 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 $\widehat{\mathbf{\Sigma}}(C)$ as a function of C for three choices of thresholding rules. When the minimum eigenvalue reaches its maximum value, the covariance estimator becomes diagonal.



We can then transform back to the covariance matrix as: $\widehat{\mathbf{\Sigma}} = \text{diag}(\mathbf{S})^{1/2}\widehat{\mathbf{R}}\text{diag}(\mathbf{S})^{1/2}$. Note that if $\mathbf{R}^{\mathcal{T}}$ itself is positive semi-definite, $\widehat{\mathbf{R}} = \mathbf{R}^{\mathcal{T}}$; otherwise $\widehat{\mathbf{R}}$ is the nearest positive semi-definite correlation matrix. 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 when $\mathbf{R}^{\mathcal{T}}$ is not positive definite. Liu et al. (2014a) introduced a covariance estimation method named EC2 (Estimation of Covariance with Eigenvalue Constraints). To motivate this method, note that the thresholding method (5) can be equivalently casted as the solution to a penalized least squares problem:

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

where $P_{\omega_T}(\cdot)$ 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) (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:

$$\widetilde{\mathbf{\Sigma}} = \underset{\lambda_{\min}(\Sigma) \geq \tau}{\text{argmin}} \left\{ \frac{1}{2}\|\mathbf{S} - \mathbf{\Sigma}\|_{\text{F}}^2 + \sum_{i \neq j} P_{\omega_T}(\sigma_{ij}) \right\} \tag{6}$$

where $\tau > 0$ is a pre-specified tuning parameter that controls the smallest eigenvalue of the estimated covariance matrix $\widetilde{\mathbf{\Sigma}}$. As a result, both sparsity and positive definiteness are guaranteed. Liu et al. (2014a) showed that the problem (6) is convex when the penalty function is convex, and develops an efficient algorithm to solve it. More details on the algorithm and theory of this estimator will be explained in Section 4.

3 Estimating sparse precision matrix

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 $\boldsymbol{Y}_t = (Y_{1t}, \ldots, Y_{pt})'$. More specifically, 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

$$\mathbf{\Sigma} = \begin{pmatrix} 1.05 & -0.23 & 0.05 & -0.02 & 0.05 \\ -0.23 & 1.45 & -0.25 & 0.10 & -0.25 \\ 0.05 & -0.25 & 1.10 & -0.24 & 0.10 \\ -0.02 & 0.10 & -0.24 & 1.10 & -0.24 \\ 0.05 & -0.25 & 0.10 & -0.24 & 1.10 \end{pmatrix} \quad \text{and} \quad \mathbf{\Theta} = \begin{pmatrix} 1 & 0.2 & 0 & 0 & 0 \\ 0.2 & 1 & 0.2 & 0 & 0.2 \\ 0 & 0.2 & 1 & 0.2 & 0 \\ 0 & 0 & 0.2 & 1 & 0.2 \\ 0 & 0.2 & 0 & 0.2 & 1 \end{pmatrix}.$$

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 $\mathbf{\Sigma}$ dense, which implies that every pair of variables are marginally correlated. Thus the covariance matrix and inverse covariance matrix encode different relationships. 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).

Estimating 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 maximum likelihood. When $\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_T$ are independently and identically distributed as $N(\mathbf{0}, \boldsymbol{\Sigma})$, the negative Gaussian log-likelihood function is given by $\ell(\boldsymbol{\Theta}) = \mathrm{tr}\,(\mathbf{S}\boldsymbol{\Theta}) - \log|\boldsymbol{\Theta}|$. When either the data are non-Gaussian or the data are weakly dependent, $\ell(\boldsymbol{\Theta})$ becomes the quasi negative log-likelihood. Nevertheless, we then consider the following penalized likelihood method:

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

where 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 penalty is the ℓ_1 penalty $P_{\omega_T}(t) = \omega_T|t|$, and the problem is then well studied in the literature (e.g., Yuan and Lin (2007); Friedman et al. (2008); Banerjee et al. (2008)). Other related works are found in, e.g., Meinshausen and Bühlmann (2006a); Wille et al. (2004).

In general, we recommend to use folded concave penalties such as SCAD and MCP, as these penalties do not introduce extra bias for estimating nonzero entries with large absolute values (Lam and Fan, 2009). Using local linear approximations, the penalized likelihood can be computed by an iterated reweighed Lasso: Given the estimate $\widehat{\boldsymbol{\Theta}}^{(k)} = (\widehat{\theta}_{ij}^{(k)})$ at the k^{th} iteration, by the Taylor's 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

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

Then the next iteration is approximated by

$$\widehat{\boldsymbol{\Theta}}^{(k+1)} = \arg\min_{\boldsymbol{\Theta}=(\theta_{ij})} \left\{ \mathrm{tr}\,(\mathbf{S}\boldsymbol{\Theta}) - \log|\boldsymbol{\Theta}| + \sum_{i\neq j} P'_{\omega_T}(|\widehat{\theta}_{ij}^{(k)}|)|\theta_{ij}| \right\} + c, \tag{7}$$

where c is a constant that does not depend on $\boldsymbol{\Theta}$. The problem (7) is convex and can be solved by the graphical Lasso algorithm of Friedman et al. (2008). Such an algorithm is called majorization-minimization algorithm (Lange et al., 2000). Since the penalty function is majorized from above, it can easily be shown that the original objective function is decreasing in the iterations. Indeed, let $f(\boldsymbol{\Theta}) = \mathrm{tr}\,(\mathbf{S}\boldsymbol{\Theta}) - \log|\boldsymbol{\Theta}| + \sum_{i\neq j} P_{\omega_T}(|\theta_{ij}|)$ be the target value and $g(\boldsymbol{\Theta})$ 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 method

Under the Gaussian model $\boldsymbol{Y}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, another approach to estimating the precision matrix $\boldsymbol{\Theta}$ is through column-by-column regressions. For this, Yuan (2010) and Cai et al. (2011) propose the graphical Dantzig selector and CLIME respectively, which can be solved by linear programming. More recently, Liu and Luo (2012) and Sun and Zhang (2012) propose the SCIO and scaled-Lasso methods. Compared to the penalized likelihood methods, the column-by-column estimation methods are computationally simpler and more amenable to theoretical analysis.

The column-by-column precision matrix estimation method exploits the relationship between conditional distribution of multivariate Gaussian and linear regression. More specifically, let $\boldsymbol{Y} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, the conditional distribution of Y_j given $\boldsymbol{Y}_{\backslash j}$ satisfies

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

where $\boldsymbol{\alpha}_j = (\boldsymbol{\Sigma}_{\backslash j,\backslash j})^{-1}\boldsymbol{\Sigma}_{\backslash j,j} \in \mathbb{R}^{p-1}$ and $\sigma_j^2 = \boldsymbol{\Sigma}_{jj} - \boldsymbol{\Sigma}_{\backslash j,j}(\boldsymbol{\Sigma}_{\backslash j,\backslash j})^{-1}\boldsymbol{\Sigma}_{\backslash j,j}$. Hence, we can write

$$Y_j = \boldsymbol{\alpha}_j' \boldsymbol{Y}_{\backslash j} + \epsilon_j, \tag{8}$$

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

$$\boldsymbol{\Theta}_{jj} = \sigma_j^{-2},, \qquad \boldsymbol{\Theta}_{\backslash j,j} = -\sigma_j^{-2}\boldsymbol{\alpha}_j. \tag{9}$$

Therefore, we can recover $\boldsymbol{\Theta}$ in a column-by-column manner by regressing Y_j on $\boldsymbol{Y}_{\backslash j}$ for $j = 1, 2, \cdots, p$. For example, let $\mathbf{Y} \in \mathbb{R}^{T\times p}$ be the data matrix. We denote by $\boldsymbol{\alpha}_j := (\alpha_{j1}, \ldots, \alpha_{j(p-1)})' \in \mathbb{R}^{p-1}$. Meinshausen and Bühlmann (2006b) propose to estimate each $\boldsymbol{\alpha}_j$ by solving the Lasso regression:

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

where λ_j is a tuning parameter. Once $\widehat{\boldsymbol{\alpha}}_j$ is obtained, we get the neighborhood edges by reading out the nonzero coefficients of $\boldsymbol{\alpha}_j$. The final graph estimate $\widehat{G}$ is obtained by either the "AND" or "OR" rule on combining the neighborhoods for all the N nodes.

To estimate $\mathbf{\Theta}$, we also estimate the σ_j^2's using the fitted sum of squared residuals $\widehat{\sigma}_j^2 = T^{-1}\big\|\mathbf{Y}_{*j} - \mathbf{Y}_{*\backslash j}\boldsymbol{\alpha}_j\big\|_2^2$, then plug it into (9).

In another work, Yuan (2010) proposes to estimate $\boldsymbol{\alpha}_j$ by solving the Dantzig selector:

$$\widehat{\boldsymbol{\alpha}}_j = \underset{\boldsymbol{\alpha}_j \in \mathbb{R}^{p-1}}{\operatorname{argmin}} \big\|\boldsymbol{\alpha}_j\big\|_1 \;\text{ subject to }\; \big\|\mathbf{S}_{\backslash j,j} - \mathbf{S}_{\backslash j,\backslash j}\boldsymbol{\alpha}_j\big\|_\infty \leq \gamma_j,$$

where $\mathbf{S} := T^{-1}\mathbf{Y}'\mathbf{Y}$ is the sample covariance matrix and γ_j is a tuning parameter. The constraint corresponds to a sample version of $\mathbf{\Sigma}_{\backslash j,j} - \mathbf{\Sigma}_{\backslash j,\backslash j}\boldsymbol{\alpha}_j = 0$, with γ_j indicating the estimation error. Once $\widehat{\boldsymbol{\alpha}}_j$ is given, we can estimate σ_j^2 by $\widehat{\sigma}_j^2 = \big[1 - 2\widehat{\boldsymbol{\alpha}}_j'\mathbf{S}_{\backslash j,j} + \widehat{\boldsymbol{\alpha}}_j'\mathbf{S}_{\backslash j,\backslash j}\widehat{\boldsymbol{\alpha}}_j\big]^{-1}$. We then obtain an estimator $\widehat{\mathbf{\Theta}}$ of $\mathbf{\Theta}$ by plugging $\widehat{\boldsymbol{\alpha}}_j$ and $\widehat{\sigma}_j^2$ into (9). Yuan (2010) analyzes the L_1-norm error $\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\|_1$ and shows its minimax optimality over certain model space.

More recently, Sun and Zhang (2012) propose to estimate $\boldsymbol{\alpha}_j$ and σ_j by solving a scaled-Lasso problem:

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

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 $\mathbf{\Theta}$ by plugging $\widehat{\boldsymbol{\alpha}}_j$ and $\widehat{\sigma}_j$ into(9). Sun and Zhang (2012) provide 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) propose the CLIME estimator, which stands for "Constrained ℓ_1-Minimization for Inverse Matrix Estimation". This method directly estimates the j^{th} column of $\mathbf{\Theta}$ by solving

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

where $\mathbf{e}_j$ is the j^{th} canonical vector (i.e., the vector with the j^{th} element being 1, while the remaining elements being 0) and δ_j is a tuning parameter. Again, the constraint represent a sample version of $\mathbf{\Sigma}\mathbf{\Theta}_{*j} - \mathbf{e}_j = 0$. 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 $\widehat{\mathbf{\Theta}}$ is asymptotically positive definite, and derive its rate of convergence.

In a closely related work of CLIME, Liu and Luo (2012) propose the SCIO estimator, which estimates the j^{th} column of $\mathbf{\Theta}$ by

$$\widehat{\mathbf{\Theta}}_{*j} = \underset{\mathbf{\Theta}_{*j}}{\operatorname{argmin}} \left\{ \frac{1}{2}\mathbf{\Theta}_{*j}'\mathbf{S}\mathbf{\Theta}_{*j} - \mathbf{e}_j'\mathbf{\Theta}_{*j} + \lambda_j \big\|\mathbf{\Theta}_{*j}\big\|_1 \right\}.$$

The SCIO estimator can be solved efficiently by the pathwise coordinate descent algorithm

(Friedman et al., 2007).

3.3 Tuning-insensitive precision matrix estimation

Most of the methods described in the former subsection require choosing some tuning parameters that control the bias-variance tradeoff. Their theoretical justifications are usually built on some theoretical choices of tuning parameters that cannot be implemented in practice. For example, in the neighborhood pursuit method and the graphical Dantzig selector, the theoretically optimal tuning parameters λ_j and γ_j depend on σ_j^2, which is unknown. The optimal tuning parameters of the CLIME and SCIO depend on $\|\mathbf{\Theta}\|_1$, which is unknown.

3.3.1 The TIGER method

To handle the challenge of tuning parameter selection, Liu and Wang (2012) propose the TIGER (Tuning-Insensitive Graph Estimation and Regression) method, which is asymptotically tuning-free and only requires very few efforts to choose the regularization parameter in finite sample settings.

The idea of TIGER is to estimate the precision matrix $\mathbf{\Theta}$ in a column-by-column fashion. This idea has been adopted by many methods described in Section 3.2. These methods differ from each other mainly in how they solve the sparse regression subproblem. The only difference between TIGER and these methods is that TIGER solves its column-wise sparse regression problem using the SQRT-Lasso (Belloni et al., 2012).

The SQRT-Lasso is a penalized optimization algorithm for solving high dimensional linear regression problems. For a linear regression problem $\widetilde{\boldsymbol{Y}} = \widetilde{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\widetilde{\boldsymbol{Y}} \in \mathbb{R}^T$ is the response vector, $\widetilde{\mathbf{X}} \in \mathbb{R}^{T\times p}$ is the design matrix, $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of unknown coefficients, and $\boldsymbol{\epsilon} \in \mathbb{R}^T$ is the noise vector. The SQRT-Lasso estimates $\boldsymbol{\beta}$ by solving

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \Big\{ \frac{1}{\sqrt{T}} \|\widetilde{\boldsymbol{Y}} - \widetilde{\mathbf{X}}\boldsymbol{\beta}\|_2 + \lambda \|\boldsymbol{\beta}\|_1 \Big\},$$

where λ is a tuning parameter. It is shown in Belloni et al. (2012) that the choice of λ for the SQRT-Lasso method is asymptotically universal in the sense that it does not depend on any unknown parameters such as the noise variance. In contrast, most of other methods, including the Lasso and Dantzig selector, rely heavily on variance of the noise. Moreover, the SQRT-Lasso method achieves near oracle performance for the estimation of $\boldsymbol{\beta}$.

In Liu and Wang (2012), they show that the objective function of the scaled-Lasso is a variational upper bound of the SQRT-Lasso. Thus the TIGER method is numerically equivalent to the method in Sun and Zhang (2012). However, the SQRT-Lasso is more amenable to theoretical analysis and allows us to simultaneously establish optimal rates

of convergence for the precision matrix estimation under many different norms.

In our setting, recall that $\mathbf{S}$ is the sample covariance matrix of $\boldsymbol{Y}_t = (Y_{1t}, \ldots, Y_{pt})'$. Let $\widehat{\boldsymbol{\Gamma}} = \mathrm{diag}(\mathbf{S})$ be a p-dimensional diagonal matrix with the diagonal elements be the same as those in $\mathbf{S}$. Consider the marginally standardized variables

$$\boldsymbol{Z} := (Z_1, \ldots, Z_p)' = \boldsymbol{Y}\widehat{\boldsymbol{\Gamma}}^{-1/2}.$$

By (8), we have

$$Z_j \widehat{\boldsymbol{\Gamma}}_{jj}^{1/2} = \boldsymbol{\alpha}_j' \boldsymbol{Z}_{\backslash j} \widehat{\boldsymbol{\Gamma}}_{\backslash j,\backslash j}^{1/2} + \epsilon_j. \tag{10}$$

We define

$$\boldsymbol{\beta}_j := \widehat{\boldsymbol{\Gamma}}_{\backslash j,\backslash j}^{1/2} \widehat{\boldsymbol{\Gamma}}_{jj}^{-1/2} \boldsymbol{\alpha}_j \quad \text{and} \quad \tau_j^2 = \sigma_j^2 \widehat{\boldsymbol{\Gamma}}_{jj}^{-1}.$$

Therefore, we have

$$Z_j = \boldsymbol{\beta}_j' \boldsymbol{Z}_{\backslash j} + \widehat{\boldsymbol{\Gamma}}_{jj}^{-1/2} \epsilon_j. \tag{11}$$

We define $\widehat{\mathbf{R}}$ to be the sample correlation matrix: $\widehat{\mathbf{R}} := \big(\mathrm{diag}(\mathbf{S})\big)^{-1/2} \mathbf{S} \big(\mathrm{diag}(\mathbf{S})\big)^{-1/2}$.

Motivated by the model in (11), we propose the following precision matrix estimator.

TIGER Estimator

For $j = 1, \ldots, p,$ we estimate the j^{th} column of $\boldsymbol{\Theta}$ by solving :

$$\widehat{\boldsymbol{\beta}}_j := \underset{\boldsymbol{\beta}_j \in \mathbb{R}^{p-1}}{\mathrm{argmin}} \left\{ \sqrt{1 - 2\boldsymbol{\beta}_j' \widehat{\mathbf{R}}_{\backslash j,j} + \boldsymbol{\beta}_j' \widehat{\mathbf{R}}_{\backslash j,\backslash j} \boldsymbol{\beta}_j} + \pi \sqrt{\frac{\log p}{2T}} \big\| \boldsymbol{\beta}_j \big\|_1 \right\}, \tag{12}$$

$$\widehat{\tau}_j := \sqrt{1 - 2\widehat{\boldsymbol{\beta}}_j' \widehat{\mathbf{R}}_{\backslash j,j} + \widehat{\boldsymbol{\beta}}_j' \widehat{\mathbf{R}}_{\backslash j,\backslash j} \widehat{\boldsymbol{\beta}}_j}, \tag{13}$$

$$\widehat{\boldsymbol{\Theta}}_{jj} = \widehat{\tau}_j^{-2} \widehat{\boldsymbol{\Gamma}}_{jj}^{-1} \quad \text{and} \quad \widehat{\boldsymbol{\Theta}}_{\backslash j,j} = -\widehat{\tau}_j^{-2} \widehat{\boldsymbol{\Gamma}}_{jj}^{-1/2} \widehat{\boldsymbol{\Gamma}}_{\backslash j,\backslash j}^{-1/2} \widehat{\boldsymbol{\beta}}_j.$$

Note that the first term in (12) is just the square-root of the the sum of the square loss for the standardized variable under model (11); see (14). We see that the TIGER procedure is tuning free. If a symmetric precision matrix estimate is preferred, we conduct the following correction: $\widetilde{\boldsymbol{\Theta}}_{jk} = \min\big\{\widehat{\boldsymbol{\Theta}}_{jk}, \widehat{\boldsymbol{\Theta}}_{kj}\big\}$ for all $k \neq j$. Another symmetrization method is

$$\widetilde{\boldsymbol{\Theta}} = \frac{\widehat{\boldsymbol{\Theta}} + \widehat{\boldsymbol{\Theta}}'}{2}.$$

Cai et al. (2011) show that, if $\widehat{\boldsymbol{\Theta}}$ is a good estimator, then $\widetilde{\boldsymbol{\Theta}}$ will also be a good estimator: they achieve the same rates of convergence in the asymptotic settings.

Let $\mathbf{Z} \in \mathbb{R}^{T \times p}$ be the normalized data matrix, i.e., $\mathbf{Z}_{*j} = \mathbf{Y}_{*j} \widehat{\boldsymbol{\Gamma}}_{jj}^{-1/2}$ for $j = 1, \ldots, p$.

An equivalent form of (12) and (13) is

$$\widehat{\boldsymbol{\beta}}_j = \underset{\boldsymbol{\beta}_j \in \mathbb{R}^{p-1}}{\operatorname{argmin}} \left\{ \frac{1}{\sqrt{T}} \left\| \mathbf{Z}_{*j} - \mathbf{Z}_{*\backslash j} \boldsymbol{\beta}_j \right\|_2 + \lambda \left\| \boldsymbol{\beta}_j \right\|_1 \right\}, \tag{14}$$

$$\widehat{\tau}_j = \frac{1}{\sqrt{T}} \left\| \mathbf{Z}_{*j} - \mathbf{Z}_{*\backslash j} \widehat{\boldsymbol{\beta}}_j \right\|_2. \tag{15}$$

Once $\widehat{\boldsymbol{\Theta}}$ is estimated, we can also estimate the graph $\widehat{G} := (V, \widehat{E})$ based on the sparsity pattern of $\widehat{\boldsymbol{\Theta}}_{jk} \neq 0$.

Liu and Wang (2012) establish the rates of convergence of the TIGER estimator $\widehat{\boldsymbol{\Theta}}$ to the true precision matrix $\boldsymbol{\Theta}$ under different norms. Under the assumption that the condition number of $\boldsymbol{\Theta}$ is bounded by a constant, we have

$$\left\| \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta} \right\|_{\max} = O_P \left(\|\boldsymbol{\Theta}\|_1 \sqrt{\frac{\log p}{T}} \right). \tag{16}$$

Under mild conditions, the obtained rate in (16) is minimax optimal over the model class consisting of precision matrices with bounded condition numbers.

The result in (16) implies that the Frobenious norm error and spectral norm error between $\widehat{\boldsymbol{\Theta}}$ and $\boldsymbol{\Theta}$ satisfy the following: let $s := \sum_{j \neq k} 1\{\boldsymbol{\Theta}_{jk} \neq 0\}$ be the number of nonzero off-diagonal elements of $\boldsymbol{\Theta}$; let $k := \max_{i=1,\ldots,p} \sum_j 1\{\boldsymbol{\Theta}_{ij} \neq 0\}$,

$$\left\| \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta} \right\|_{\mathrm{F}} = O_P \left(\|\boldsymbol{\Theta}\|_1 \sqrt{\frac{(p+s)\log p}{T}} \right), \tag{17}$$

$$\left\| \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta} \right\|_2 = O_P \left(k \|\boldsymbol{\Theta}\|_2 \sqrt{\frac{\log p}{T}} \right). \tag{18}$$

The obtained rates in (18) and (17) are minimax optimal over the same model class as before.

3.3.2 The EPIC method

Another tuning-insensitive precision matrix estimation method is EPIC (Estimating Precision matrIx with Calibration), proposed by Zhao and Liu (2014). While TIGER can be viewed as a tuning-insensitive extension of the nodewise Lasso method proposed by Meinshausen and Bühlmann (2006b), EPIC can be viewed as a tuning-insensitive extension of the CLIME estimator proposed by Cai et al. (2011). Unlike the TIGER method which relies on the normality assumption, the EPIC method can be used to handle both sub-Gaussian and heavy-tailed data. We postpone the details of the EPIC method to Section 4 where we discuss robust estimators of covariance and precision matrices for heavy-tailed data.

4 Robust precision and covariance estimators

The methods introduced in Section 2 and Section 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, many types of financial data are believed to follow the elliptical distributions, which 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-tailness and conducts parameter estimation under the elliptical models. 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 tradeoff between model flexibility and interpretability.

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

Definition 1 (Elliptical Distribution)**.** *Given $\boldsymbol{\mu} \in \mathbb{R}^p$ and a symmetric positive semidefinite matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ with* $\text{rank}(\boldsymbol{\Sigma}) = r \leq p$*, a p-dimensional random vector $\boldsymbol{Y} = (Y_1, ..., Y_p)'$ 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*

$$\boldsymbol{Y} \overset{d}{=} \boldsymbol{\mu} + \xi \mathbf{A}\mathbf{u}, \tag{19}$$

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 notation convenience, we use ξ instead the distribution of ξ in the notation $EC(\boldsymbol{\mu}, \xi, \boldsymbol{\Sigma})$. Note that the model in (19) is not identifiable since 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 precision matrix estimation

To estimate $\boldsymbol{\Theta}$, our key observation is that the covariance matrix $\boldsymbol{\Sigma}$ can be decomposed as $\boldsymbol{\Sigma} = \mathbf{D}\mathbf{R}\mathbf{D}$, where $\mathbf{R}$ is the Pearson's correlation matrix, and $\mathbf{D} = \text{diag}(\sigma_1, ..., \sigma_p)$ where σ_j is the standard deviation of Y_j. Since $\mathbf{D}$ 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. Based on this relationship, the EPIC method of Zhao and Liu (2014) has three steps:

first obtain estimators $\widehat{\mathbf{R}}$ and $\widehat{\mathbf{D}}$ for $\mathbf{R}$ and $\mathbf{D}$; then apply a calibrated inverse correlation matrix estimation procedure on $\widehat{\mathbf{R}}$ to obtain $\widehat{\mathbf{\Delta}}$, an estimator for $\mathbf{\Delta}$. Finally, assemble $\widehat{\mathbf{\Delta}}$ and $\widehat{\mathbf{D}}$ to obtain a sparse precision matrix estimator $\widehat{\mathbf{\Theta}}$.

For light-tailed distributions (e.g., Gaussian or sub-Gaussian), we can directly use the sample correlation matrix and sample standard deviation to estimate the matrices $\mathbf{R}$ 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 transformed Kendall's tau estimator and Catoni's M-estimator, which will be explained in details in the following subsections.

4.1.1 Robust estimation of correlation matrix

To estimate $\mathbf{R}$, we adopt a transformed Kendall's tau estimator proposed in Fang et al. (1990). Define the population Kendall's tau correlation between Y_{jt} and Y_{kt} as

$$\tau_{kj} = \mathbb{P}\left((Y_{jt} - \widetilde{Y}_{jt})(Y_{kt} - \widetilde{Y}_{k}) > 0\right) - \mathbb{P}\left((Y_{jt} - \widetilde{Y}_{j})(Y_{kt} - \widetilde{Y}_{k}) < 0\right),$$

where $\widetilde{Y}_j$ and $\widetilde{Y}_k$ are independent copies of Y_{jt} and Y_{kt} respectively. For elliptical distributions, it is a well known result that $\mathbf{R}_{kj}$ and τ_{kj} satisfy[1]

$$\mathbf{R} = [\mathbf{R}_{kj}] = \left[\sin\left(\frac{\pi}{2}\tau_{kj}\right)\right]. \tag{20}$$

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

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

for all $k \neq j$, and $\widehat{\tau}_{kj} = 1$ otherwise. We can plug $\widehat{\tau}_{kj}$ into (20) and obtain a rank-based correlation matrix estimator

$$\widehat{\mathbf{R}} = [\widehat{\mathbf{R}}_{kj}] = \left[\sin\left(\frac{\pi}{2}\widehat{\tau}_{kj}\right)\right]. \tag{21}$$

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 function where $\text{sign}(0) = 0$. 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

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

equations:

$$\sum_{t=1}^{T} \psi\left((Y_{jt} - \mu_j)\sqrt{\frac{2}{TK_{\max}}}\right) = 0, \tag{22}$$

$$\sum_{t=1}^{T} \psi\left((Y_{jt}^2 - m_j)\sqrt{\frac{2}{TK_{\max}}}\right) = 0, \tag{23}$$

where $K_{\max}$ is an upper bound of $\max_j \mathrm{Var}(Y_{jt})$ and $\max_j \mathrm{Var}(Y_{jt}^2)$. We assume $K_{\max}$ is known. Catoni (2012) shows that the solutions to (22) and (23) 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

$$\widehat{\sigma}_j = \sqrt{\max\{\widehat{m}_j - \widehat{\mu}_j^2, K_{\min}\}}, \tag{24}$$

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. (2012b) show that

$$\|\widehat{\mathbf{R}} - \mathbf{R}\|_{\max} = O_P\Big(\sqrt{\frac{\log p}{T}}\Big) \quad \text{and} \quad \max_{1\le j\le p} |\widehat{\sigma}_j - \sigma_j| = O_P\Big(\sqrt{\frac{\log p}{T}}\Big). \tag{25}$$

In contrast, the sample correlation matrix and sample standard deviation do not have the above properties for heavy-tailed data.

4.1.3 The EPIC method for inverse correlation matrix estimation

Once $\widehat{\mathbf{R}}$ and $\widehat{\mathbf{D}}$ are obtained, we need to estimate the inverse correlation matrix $\mathbf{\Delta} = \mathbf{R}^{-1}$. In this subsection, we introduce the EPIC method for estimating $\mathbf{\Delta}$, which estimates the j^{th} column of $\mathbf{\Delta}$ by plugging the transformed Kendall's tau estimator $\widehat{\mathbf{R}}$ into the convex program,

$$(\widehat{\mathbf{\Delta}}_{*j}, \widehat{\tau}_j) = \underset{\mathbf{\Delta}_{*j}, \tau_j}{\operatorname{argmin}} \ \|\mathbf{\Delta}_{*j}\|_1 + \frac{1}{2}\tau_j, \ \text{ s.t. } \|\widehat{\mathbf{R}}\mathbf{\Delta}_{*j} - \mathbf{I}_{*j}\|_\infty \le \lambda\tau_j, \ \|\mathbf{\Delta}_{*j}\|_1 \le \tau_j. \tag{26}$$

Here τ_j serves as an auxiliary variable which ensures that we can use the same regularization parameter λ for estimating different columns of $\mathbf{\Delta}$ (Gautier and Tsybakov, 2011). Both the objective function and constraints in (26) contain τ_j, which ensures that τ_j is bounded. Zhao and Liu (2014) show that the regularization parameter λ in (26) does not depend on the unknown quantity $\mathbf{\Delta}$. Thus we can use the same λ to estimate different columns of $\mathbf{\Delta}$.

The optimization problem in (26) can be equivalently formulated as a linear program.

For notational simplicity, we omit the index j in (26). We denote $\mathbf{\Delta}_{*j}$, $\mathbf{I}_{*j}$, and τ_j by $\boldsymbol{\gamma}$, $\mathbf{e}$, and τ respectively. Let $\boldsymbol{\gamma}^+$ and $\boldsymbol{\gamma}^-$ be the positive and negative parts of $\boldsymbol{\gamma}$. By reparametrizing $\boldsymbol{\gamma} = \boldsymbol{\gamma}^+ - \boldsymbol{\gamma}^-$, we rewrite (26) as the following linear program

$$
\begin{aligned}
(\widehat{\boldsymbol{\gamma}}^+, \widehat{\boldsymbol{\gamma}}^-, \widehat{\tau}) = & \underset{\boldsymbol{\gamma}^+, \boldsymbol{\gamma}^-, \tau}{\operatorname{argmin}} \quad \mathbf{1}'\boldsymbol{\gamma}^+ + \mathbf{1}'\boldsymbol{\gamma}^- + c\tau \\
& \text{s.t.} \quad \begin{bmatrix} \widehat{\mathbf{R}} & -\widehat{\mathbf{R}} & -\boldsymbol{\lambda} \\ -\widehat{\mathbf{R}} & \widehat{\mathbf{R}} & -\boldsymbol{\lambda} \\ \mathbf{1}' & \mathbf{1}' & -1 \end{bmatrix} \begin{bmatrix} \boldsymbol{\gamma}^+ \\ \boldsymbol{\gamma}^- \\ \tau \end{bmatrix} \leq \begin{bmatrix} \mathbf{e} \\ -\mathbf{e} \\ 0 \end{bmatrix}, \\
& \boldsymbol{\gamma}^+ \geq \mathbf{0}, \ \boldsymbol{\gamma}^- \geq \mathbf{0}, \ \tau \geq 0,
\end{aligned} \tag{27}
$$

where $\boldsymbol{\lambda} = \lambda \mathbf{1}$. The optimization problem in (27) 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.

4.1.4 Symmetric precision matrix estimation

Once we obtain the inverse correlation matrix estimate $\widehat{\mathbf{\Delta}}$, we can estimate $\mathbf{\Theta}$ by

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

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

$$
\widehat{\mathbf{\Theta}} = \underset{\mathbf{\Theta}}{\operatorname{argmin}} \, \| \mathbf{\Theta} - \widetilde{\mathbf{\Theta}} \|_* \quad \text{s.t. } \mathbf{\Theta} = \mathbf{\Theta}', \tag{28}
$$

where $\| \cdot \|_*$ can be the matrix ℓ_1-, Frobenius, or elementwise max norm. For both the Frobenius and elementwise max norms, (28) has a closed form solution

$$
\widehat{\mathbf{\Theta}} = \frac{1}{2} \left(\widetilde{\mathbf{\Theta}} + \widetilde{\mathbf{\Theta}}' \right).
$$

When using the matrix ℓ_1-norm, the optimization problem in (28) 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

$$
\mathcal{U}(s, M, \kappa_u) = \left\{ \mathbf{\Delta} \in \mathbb{R}^{p \times p} \;\middle|\; \mathbf{\Delta} \succ 0, \ \max_j \sum_k 1 \left\{ \mathbf{\Delta}_{kj} \neq 0 \right\} \leq s, \| \mathbf{\Delta} \|_1 \leq M, \Lambda_{\max}(\mathbf{\Delta}) \leq \kappa_u \right\},
$$

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 (28), then for large enough T, we have

$$\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_2 = O_P\Big(M \cdot s\sqrt{\frac{\log p}{T}}\Big). \tag{29}$$

Moreover, if we choose the Forbenius norm as $\|\cdot\|_*$ in (28), then for large enough T,

$$\frac{1}{p}\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_{\mathrm{F}}^2 = O_P\Big(M^2\frac{s\log p}{T}\Big). \tag{30}$$

4.2 Robust covariance matrix estimation

In this subsection, we consider the problem of estimating the covariance matrix $\boldsymbol{\Sigma}$ under the elliptical model (19). Similar to Section 2, we impose sparsity assumption on $\boldsymbol{\Sigma}$. To estimate $\boldsymbol{\Sigma}$, Liu et al. (2014b) introduce a regularized rank-based estimation method named EC2 (Estimation of Covariance with Eigenvalue Constraints), which can be viewed as an 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.

4.2.1 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 (21) and (24). 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 transformed Kendall's tau matrix, the $\widehat{\mathbf{R}}^{\mathrm{EC2}}$ is calculated as

$$\widehat{\mathbf{R}}^{\mathrm{EC2}} := \underset{\mathrm{diag}(\mathbf{R})=1}{\mathrm{argmin}} \ \frac{1}{2}\|\widehat{\mathbf{R}} - \mathbf{R}\|_{\mathrm{F}}^2 + \lambda\|\mathbf{R}\|_{1,\mathrm{off}} \quad \text{s.t. } \tau \le \Lambda_{\min}(\mathbf{R}) \tag{31}$$

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. (2014b) prove the convexity of the formulation in (31). Alternatively, one can apply thresholding on $\widehat{\mathbf{R}}$ to obtain a positive definite estimator.

4.2.2 Asymptotic properties of the EC2 estimator

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

$$\mathcal{M}\left(q, M_p, \delta\right) := \left\{\mathbf{R} : \max_{1\leq j\leq p} \sum_{k\neq j} \left|\mathbf{R}_{jk}\right|^q \leq M_p \text{ and } \mathbf{R}_{jj} = 1 \text{ for all } j, \Lambda_{\min}\big(\mathbf{R}\big) \geq \delta\right\}.$$

We also define a class of covariance matrices:

$$\mathcal{U}(\kappa, q, M_p, \delta) := \left\{\mathbf{\Sigma} : \max_j \mathbf{\Sigma}_{jj} \leq \kappa \text{ and } \mathbf{D}^{-1}\mathbf{\Sigma}\mathbf{D}^{-1} \in \mathcal{M}\left(q, M_p, \delta\right)\right\}, \tag{32}$$

where $\mathbf{D} = \text{diag}(\sqrt{\mathbf{\Sigma}_{11}}, ..., \sqrt{\mathbf{\Sigma}_{pp}})$. 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. (2014b) show that, for large enough T, the EC2 estimator $\widehat{\mathbf{\Sigma}}$ satisfies

$$\sup_{\mathbf{\Sigma}\in\mathcal{U}(\kappa,q,M_p,\delta_{\min})} \mathbb{E}\big\|\widehat{\mathbf{\Sigma}}^{\text{EC2}} - \mathbf{\Sigma}\big\|_2 \leq c_1 \cdot M_p\Big(\frac{\log p}{T}\Big)^{\frac{1-q}{2}}. \tag{33}$$

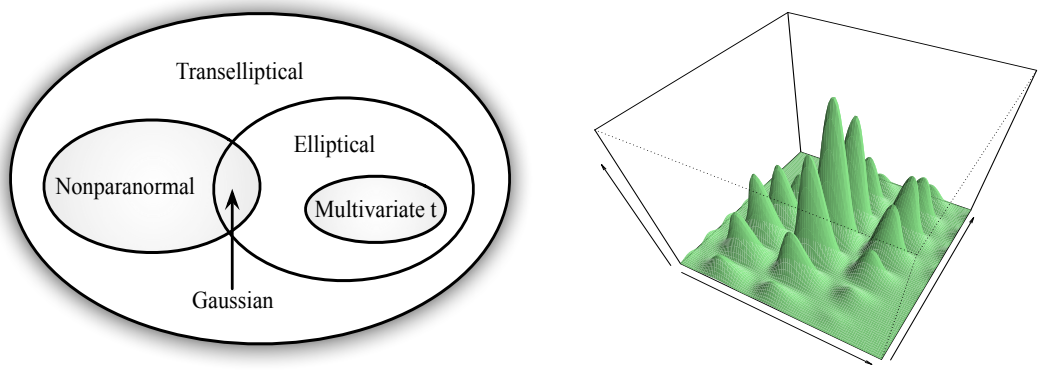
Cai and Zhou (2012) show that the rate in (33) attains the minimax lower bound over the class $\mathcal{U}(\kappa, q, M_d, \delta_{\min})$ under the Gaussian model. Thus the EC2 estimator is asymptotically rate optimal under the flexible elliptical model with covariance matrix in $\mathcal{U}(\kappa, q, M_d, \delta_{\min})$.

4.3 Extension to the elliptical copula family

In Sections 4.1 and 4.2, we introduced the regularized rank-based covariance and precision matrix estimation methods by assuming the underlying distribution of $\boldsymbol{Y} = (Y_1, \ldots, Y_p)'$ is elliptical. In fact, these rank-based procedures also work within the more general transelliptical family (Liu et al., 2012c), which is exactly the elliptical copula family but with different identifiability conditions. More specifically, we say $\boldsymbol{Y} = (Y_1, \ldots, Y_p)'$ follows a transelliptical distribution, denoted by $\boldsymbol{Y} \sim TE(\boldsymbol{\mu}, \mathbf{\Sigma}, \xi; f)$, if there exists a set of strictly increasing functions $\{f_j\}_{j=1}^p$ such that $f(\boldsymbol{Y}) = (f_1(Y_1), \ldots, f_p(Y_p))'$ follows the elliptical distribution $EC(\boldsymbol{\mu}, \xi, \mathbf{\Sigma})$. To ensure the model is identifiable, Liu et al. (2012c) impose the identifiability condition that, for $j \in \{1, \ldots, p\}$,

$$\mathbb{E}f_j(Y_j) = \mathbb{E}Y_j \quad \text{and} \quad \text{Var}(f_j(Y_j)) = \text{Var}(Y_j). \tag{34}$$

As the Kendall's tau statistics in (21) are invariant under the monotonic transform, the Kendall's tau statistics for the elliptical data $f(\boldsymbol{Y})$ are the same as those for the transellipitical data $\boldsymbol{Y}$. Therefore, we do not need to estimate the monotonic trans-



(a) The Vein diagram (b) The perspective plot of a transelliptical density

Figure 2: Transelliptical family. (a) The Vein diagram illustrating the relationships of the distribution families (The Nonparanomral family is equivalent to the Gaussian copula family). (b) The perspective plot of a transelliptical density.

formations f for computing the Kendall's tau. On the other hand, these monotonic transforms are not hard to estimate. For example, for the Gaussian copula such that the marginal distribution $f_j(Y_j) \sim N(0,1)$, then based on the empirical distribution of the observed data Y_j and the known marginal distribution $N(0,1)$, we can easily estimate f_j.

Figure 2(a) illustrates the relationships between the transelliptical, elliptical, and nonparanormal families (Liu et al., 2009, 2012b). The nonparanormal family is a a proper subset of the transelliptical family. We define $\boldsymbol{Y} = (Y_1, \ldots, Y_p)'$ to be a nonparanormal distribution, denoted by $\boldsymbol{Y} \sim NPN(\boldsymbol{\mu}, \boldsymbol{\Sigma}, f)$, if there exists a set of strictly increasing functions $\{f_j\}_{j=1}^p$ such that $f(\boldsymbol{Y}) = (f_1(Y_1), \ldots, f_p(Y_p))'$ follows the Gaussian distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Liu et al. (2012c) show that the intersection between the nonparanomral family and elliptical family is the Gaussian family. Figure 2(b) visualizes the perspective plot of a bivariate transelliptical density with certain marginal transformations. The transelliptical family is much richer than the elliptical family and its density function does not have to be symmetric.

The rank-based EPIC and EC2 methods can be directly applied to the transelliptical family. To understand the semantics of a transelliptical graphical model, Liu et al. (2012c) proved that a transelliptical distribution admits a three-layer hierarchical latent variable representation as illustrated in Figure 3: The observed vector, denoted by $\boldsymbol{Y} = (Y_1, \ldots, Y_p)'$ as presented in the first layer, has a transelliptical distribution, and a latent random vector, $\boldsymbol{Z} = (Z_1, \ldots, Z_p)'$ in the second-layer, is elliptically distributed. Variables in the first and second layers are related through the transformation $Z_j = f_j(Y_j)$ with f_j being an unknown strictly increasing function. The latent vector $\boldsymbol{Z}$ can be further represented by a third-layer latent random vector $\boldsymbol{X} = (X_1, \ldots, X_p)'$, which is a multivariate Gaussian with a covariance matrix $\boldsymbol{\Sigma}$ (called latent covariance matrix) and

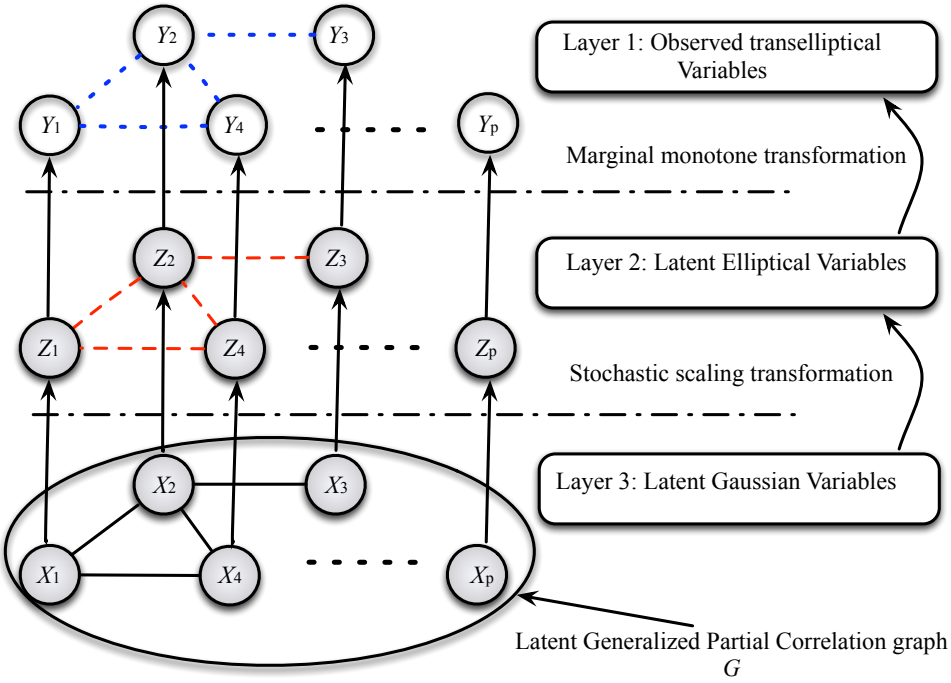


Figure 3: The hierarchical latent variable representation of the transelliptical graphical model with the latent variables grey-colored. Here the first layer is composed of observed Y_j's, and the second and third layers are composed of latent variables Z_j's and X_j's. The solid undirected lines in the third layer encode the conditional independence graph of $X_1, \ldots, X_p$ (Adapted from a manuscript that is under review).

an inverse covaraince matrix $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$ (called latent precision matrix).

We define the transelliptical graph $G = (V, E)$ with the node set $V = \{1, \ldots, p\}$ and the edge set E encoding the nonzero entries of $\mathbf{\Theta}$. The interpretations of the graph G are different for the variables in different layers: (i) For the observed variables in the first layer, the absence of an edge between two variables means the absence of a certain rank-based association (e.g., Kendall's tau) of the pair given other variables; (ii) For the latent variables in the second layer, the absence of an edge means the absence of the conditional Pearson's correlation of the pair; (iii) For the third layer variables, the absence of an edge means the conditional independence of the pair. Compared with the Gaussian and elliptical graphical model, the transelliptical graphical model has richer structure with more relaxed modeling assumptions. The three layers of hierarchy also reflects an interesting tradeoff between model flexibility and interpretability. In the third layer, the model is the most restrictive Gaussian family, but we can get strong conditional independence arguments. In contrast, in the first layer, the model is the much more flexible transelliptical family, but we can only get weaker conditional uncorrelatedenss (with respect to the rank correlation) statements.

Since the Kendall's tau statistic is monotone transformation invariant, it is easy to see that the theory and methods of the EPIC and EC2 procedures introduced in Section 4.1 and Section 4.2 are also applicable to the transelliptical distributions, though the

interpretations of the fitted results are different (as explained in this section).

5 Factor model-based covariance estimation with observable factors

Most of the aforementioned methods of estimating $\mathbf{\Sigma}$ assumes that the covariance matrix is sparse. Though this assumption is reasonable for many 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 unrealistic in these situations.

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

$$Y_{it} = \mathbf{b}_i' \boldsymbol{f}_t + u_{it}. \tag{35}$$

Here $\mathbf{b}_i$ 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 *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); Campbell et al. (1997). It is also commonly used in macroeconomics for forecasting diffusion index (e.g., Stock and Watson (2002)). We allow $p, T \to \infty$ and that p can grow much faster than T does. In contrast, the number of factors K needs to be either bounded or grows slowly.

This section introduces a method of estimating $\mathbf{\Sigma}$ using factor models. We will 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.

5.1 Conditional sparsity

The factor model (35) can be put in a matrix form as

$$\boldsymbol{Y}_t = \mathbf{B}\boldsymbol{f}_t + \boldsymbol{u}_t. \tag{36}$$

where $\mathbf{B} = (\mathbf{b}_1, ..., \mathbf{b}_p)'$ and $\boldsymbol{u}_t = (u_{1t}, ..., u_{pt})'$. We are interested in $\mathbf{\Sigma}$, the $p \times p$ covariance matrix of $\boldsymbol{Y}_t$, and its inverse $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$, which are assumed to be time-invariant. Under model (36) and the independence assumption between $\boldsymbol{f}_t$ and $\boldsymbol{u}_t$, $\mathbf{\Sigma}$ is given by

$$\mathbf{\Sigma} = \mathbf{B}\,\mathrm{Cov}(\boldsymbol{f}_t)\mathbf{B}' + \mathbf{\Sigma}_u, \tag{37}$$

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

Fan et al. (2008) studied model (37) when $p \to \infty$ possibly faster than T. They assumed $\mathbf{\Sigma}_u$ to be a diagonal matrix, which corresponds to the classical "strict factor model", and might be restrictive in practical applications. On the other hand, factor models are often only justified as being "approximate", in which the $Y_{1t}, ..., Y_{pt}$ 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, $\mathbf{\Sigma}_u$ is a non-diagonal covariance matrix, and admits many small off-diagonal entries.

In the decomposition (37), we assume $\mathbf{\Sigma}_u$ to be sparse. This can be interpreted as the conditional sparse covariance model: Given the common factors $\boldsymbol{f}_1, ..., \boldsymbol{f}_T$, 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

$$m_{u,p} = \begin{cases} \max_{i\le p} \sum_{j=1}^p 1\{\sigma_{u,ij} \neq 0\}, & \text{if } q = 0 \\ \max_{i\le p} \sum_{j=1}^p |\sigma_{u,ij}|^q, & \text{if } 0 < q < 1 \end{cases}. \tag{38}$$

We require $m_{u,p}$ be either bounded or 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: the idiosyncratic components are mostly uncorrelated. In contrast, note that in the presence of common factors, $\mathbf{\Sigma}$ itself is hardly a sparse matrix.

5.2 Estimation

When the factors are observable, one can estimate $\mathbf{B}$ by the ordinary least squares (OLS): $\widehat{\mathbf{B}} = (\widehat{\mathbf{b}}_1, ..., \widehat{\mathbf{b}}_N)'$, where,

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

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:

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

Since $\mathbf{\Sigma}_u$ 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 $\omega_{T,ij} = \mathrm{SE}(s_{u,ij})\omega_T$ and $\omega_{T,ij} = \sqrt{s_{u,ii}s_{u,jj}}\omega_T$, 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 regularization since the number of factor is assumed to be small. Therefore we obtain a substitution estimator:

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

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

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

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

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

On the other hand, it is difficult to obtain a satisfactory convergence rate for $\widehat{\boldsymbol{\Sigma}}$ under either the operator or the Frobenius norm. We illustrate this problem in the following example. Let $\mathbf{0}_d$ be a d-dimensional row vector of zeros.

Example 1. *Consider the specific case $K = 1$ with the known loading $\boldsymbol{B} = \mathbf{1}_p$ and $\boldsymbol{\Sigma}_u = \mathbf{I}$. Then $\boldsymbol{\Sigma} = \mathrm{Var}(f_1)\mathbf{1}_p\mathbf{1}_p^T + \mathbf{I}$, 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*

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

Therefore, it follows from the central limit theorem that $\frac{\sqrt{T}}{p}\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_2$ is asymptotically normal. Hence $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_2$ 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.

On the other hand, 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 $\mathbf{\Sigma}$ may diverge fast.

6 Factor models-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:

$$\mathbf{\Sigma} = \mathbf{B}\,\mathrm{Cov}(\boldsymbol{f}_t)\mathbf{B}' + \mathbf{\Sigma}_u, \tag{39}$$

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 (39) is not identifiable under fixed (p, T), since $\boldsymbol{Y}_t$ 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 (39). 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 $\mathbf{\Sigma}_u$ to be either uniformly bounded or grow slowly relative to p.

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

Assumption 1. *The eigenvalues of the $K \times K$ matrix $p^{-1}\mathbf{B}'\mathbf{B} = \frac{1}{p}\sum_{i=1}^{p}\mathbf{b}_i\mathbf{b}_i'$ 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 impact on most of the cross-sectional individuals. It then follows that the

first K eigenvalues of $\mathbf{B}\,\mathrm{Cov}(\boldsymbol{f}_t)\mathbf{B}'$ are bounded from below by $c\lambda_{\min}(\mathrm{Cov}(\boldsymbol{f}_t))p$ for some $c > 0$, and should grow fast with p. On the other hand,

$$\|\mathbf{\Sigma}_u\|_2 \leq \max_{i\leq p} \sum_{j=1}^{p} |\sigma_{u,ij}|^q |\sigma_{u,ii}\sigma_{u,jj}|^{(1-q)/2} \leq m_{u,p} \max_{i\leq p} \sigma_{u,ii}^{1-q}. \tag{40}$$

Hence when $m_{u,p}$ grows slower than $O(p)$, the leading eigenvalues of the two components on the right hand side of (39) 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, since the pair $(\mathbf{B}, \boldsymbol{f}_t)$ is equivalent to the pair $(\mathbf{B}\mathbf{H}^{-1}, \mathbf{H}\boldsymbol{f}_t)$ for any $K \times K$ nonsingular 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 (39) that

$$\mathbf{\Sigma} = \mathbf{B}\mathbf{B}' + \mathbf{\Sigma}_u.$$

Let $\widetilde{\mathbf{b}}_1, ..., \widetilde{\mathbf{b}}_K$ be the columns of $\mathbf{B}$. Since the columns of $\mathbf{B}$ are orthogonal,

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

Therefore, $\widetilde{\mathbf{b}}_1/\|\widetilde{\mathbf{b}}_1\|_2, \cdots, \widetilde{\mathbf{b}}_K/\|\widetilde{\mathbf{b}}_K\|_2$ are the eigenvectors of $\mathbf{B}\mathbf{B}'$, corresponding to the largest K eigenvalues $\{\|\widetilde{\mathbf{b}}_j\|_2^2\}_{j=1}^K$; the rest $p - K$ eigenvalues of $\mathbf{B}\mathbf{B}'$ 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 $\mathbf{B}\mathbf{B}'$ are the same as those of $\mathbf{B}'\mathbf{B}$. Hence it follows from the pervasive assumption (Assumption 1) that

$$\|\widetilde{\mathbf{b}}_j\|_2^2 \geq cp, \quad j = 1, ..., K. \tag{41}$$

Next, let us associate the leading eigenvalues of $\mathbf{B}\mathbf{B}'$ with those of $\mathbf{\Sigma}$. Let $\lambda_1, ..., \lambda_K$

denote the K largest eigenvalues of $\mathbf{\Sigma}$, and let $\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_K$ be the corresponding eigenvectors. Applying Wely's theorem and the $\sin(\theta)$-theorem of Davis (1963), Fan et al. (2013) showed

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

and

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

These results demonstrate:

1. The leading eigenvectors of $\mathbf{\Sigma}$ are approximately equal to the normalized columns of $\mathbf{B}$, as $p \to \infty$. In other words, the factor analysis and the principal analysis are approximately the same.

2. The leading eigenvalues of $\mathbf{\Sigma}$ grow at rate $O(p)$. This can be seen from applying the triangular inequality and (40), (41):

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

3. The latent factor f_{jt} is approximately $\boldsymbol{\xi}_j' \boldsymbol{Y}_t/\sqrt{\lambda_j}$ for $j = 1, ..., K$. To see this, left-multiplying $\widetilde{\mathbf{b}}_j'/\|\widetilde{\mathbf{b}}_j\|_2^2$ to $\boldsymbol{Y}_t = \mathbf{B}\boldsymbol{f}_t + \boldsymbol{u}_t$, 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 $p \to \infty$, $f_{jt} \approx \boldsymbol{\xi}_j' \boldsymbol{Y}_t/\sqrt{\lambda_j}$.

Therefore, we conclude that the first K eigenvalues of $\mathbf{\Sigma}$ 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 $\mathbf{\Sigma}$ and $\boldsymbol{Y}_t$. This builds the connection between the PCA and high-dimensional factor models.

6.3 POET estimator

Fan et al. (2013) proposed a nonparametric estimator of $\mathbf{\Sigma}$ when the factors are unobservable, named POET (Principal Orthogonal complEment Thresholding). To motivate their estimator, note that $\mathbf{B}\mathbf{B}' = \sum_{j=1}^K \widetilde{\mathbf{b}}_j \widetilde{\mathbf{b}}_j'$. 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 fact, it can be formally proved that

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

which can be understood as the (asymptotic) identification for $\mathbf{B}\mathbf{B}'$. In addition, note that $\boldsymbol{\Sigma}$ 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 $\widehat{\boldsymbol{\xi}}_1, ..., \widehat{\boldsymbol{\xi}}_p$ be the corresponding eigenvectors. Then the sample covariance has the following spectral decomposition:

$$\mathbf{S} = \sum_{i=1}^{K} \widehat{\lambda}_i \widehat{\boldsymbol{\xi}}_i \widehat{\boldsymbol{\xi}}_i' + \mathbf{S}_u,$$

where $\mathbf{S}_u = \sum_{k=K+1}^{p} \widehat{\lambda}_k \widehat{\boldsymbol{\xi}}_k \widehat{\boldsymbol{\xi}}_k' = (s_{u,ij})$, 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

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

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 $\frac{1}{\sqrt{p}}$ 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 estimated accurately enough as if they were observable.

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

$$\widehat{\mathbf{\Sigma}}_K = \sum_{i=1}^{K} \widehat{\lambda}_i \widehat{\boldsymbol{\xi}}_i \widehat{\boldsymbol{\xi}}_i' + \widehat{\mathbf{\Sigma}}_u. \tag{42}$$

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

Note that $\widehat{\mathbf{\Sigma}}_K$ requires the knowledge of K, which is the number of factors and practically unknown. There has been a large 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 over-estimating 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 sequel, we suppress the subscript K, and simply write $\widehat{\mathbf{\Sigma}}$ 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

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

On the other hand, 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 2. *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 $\widehat{\mathbf{\Sigma}}_u^{-1}$.

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 individual loadings can be estimated no faster than $O_P(T^{-1/2})$. But 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 short time horizon to infer as good as possible the risk factors as well as their associated loading matrix. To achieve this, we need additional data covariate information and modeling of the factor loadings.

Suppose that there is a d-dimensional vector of observed covariates associated with the i^{th} 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 paramters in $\boldsymbol{B}$. Specifically, they proposed and studied the following semi-parametric factor model:

$$Y_{it} = \sum_{k=1}^{K} g_k(\boldsymbol{X}_i) f_{kt} + u_{it}, \quad i = 1, \cdots, p, t = 1, \cdots, T. \tag{43}$$

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

$$g_k(\boldsymbol{X}_i) = g_{k1}(X_{i1}) + \cdots + g_{kd}(X_{id}). \tag{44}$$

Fan et al. (2014b) recognized that the above semi-parametric model (43) 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 semiparametric mixed effect model:

$$Y_{it} = \sum_{k=1}^{K} [g_k(\boldsymbol{X}_i) + \gamma_{ik}] f_{kt} + u_{it}, \quad i = 1, \cdots, p, t = 1, \cdots, T. \tag{45}$$

Here γ_{ik} is an unobservable random component with mean zero. They developed econometric techniques to test the model specifications (43) and (45). Their empirical results,

using the returns of the components of the S&P500 index and 4 exogenous variables (size, value, momentum, and volatility) as in Connor et al. (2012), provide stark evidence that model (43) can not be validated empirically whereas (45) 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 run 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 (44). This results in a smooth or projected observation matrix $\widehat{\mathbf{Y}}$, which will also be denoted by $\mathbf{PY}$. The projected PCA is then to run PCA based on the projected data $\widehat{\mathbf{Y}}$.

To provide the rationale behind this idea, we now generalize model (45) further to illustrate the idea behind the projected PCA. Specifically, consider the factor model

$$\mathbf{Y} = \mathbf{B}\mathbf{F}' + \mathbf{U}$$

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 i^{th} variable: $\boldsymbol{X}_i = (X_{i1}, \cdots, X_{id})$, which is independent of u_{it}. For a pre-determined J, let $\phi_1, ..., \phi_J$ be a set of basis functions. Let $\phi(\boldsymbol{X}_i)' = (\phi_1(X_{i1}), ..., \phi_J(X_{i1}), ..., \phi_J(X_{id}))$ and $\Phi(\mathbf{X}) = (\phi(\boldsymbol{X}_1), ..., \phi(\boldsymbol{X}_p))'$ be a $p \times (Jd)$ matrix of the sieve-transformed $\mathbf{X}$. Then the projection matrix on the space spanned by $\mathbf{X} = (\boldsymbol{X}_1, ..., \boldsymbol{X}_p)$ can be taken as

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

This corresponds to modeling $g_k(\boldsymbol{X}_i)$ in (45) by the additive model (44) and approximating each term using the series expansion. The projected data $\mathbf{PY}$ is the fitted value of the additive model (44) with basis functions $\phi_1, ..., \phi_J$:

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

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

We make the following key assumptions:

Assumption 2. (i) Pervasiveness: *With probability approaching one, all the eigenvalues of* $\frac{1}{p}(\mathbf{PB})'\mathbf{PB}$ *are bounded away from both zero and infinity as* $p \to \infty$.

(ii) Orthogonality: $\mathbb{E}(u_{it}|X_{i1}, ..., X_{id}) = 0$, *for all* $i \leq p, t \leq T$.

The above conditions require that the strengths of the loading matrix should be as strong after the projection, and $\mathbf{B}$ should be associated with $\mathbf{X}$. Condition (ii) implies that if we apply $\mathbf{P}$ to both sides of $\mathbf{Y} = \mathbf{B}\mathbf{F}' + \mathbf{U}$, then

$$\mathbf{P}\mathbf{Y} \approx \mathbf{P}\mathbf{B}\mathbf{F}',$$

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

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

$$\frac{1}{T}(\mathbf{P}\mathbf{Y})'\mathbf{P}\mathbf{Y}\mathbf{F} \approx \mathbf{F}(\mathbf{P}\mathbf{B})'(\mathbf{P}\mathbf{B}).$$

Still by the diagonality of $(\mathbf{P}\mathbf{B})'\mathbf{P}\mathbf{B}$, we infer that the columns of $\mathbf{F}$ are approximately the eigenvectors of the $T \times T$ sample covariance matrix $\frac{1}{T}(\mathbf{P}\mathbf{Y})'\mathbf{P}\mathbf{Y}$, up to a factor $\sqrt{T}$. In addition, since the diagonal elements of $(\mathbf{P}\mathbf{B})'\mathbf{P}\mathbf{B}$ grow fast as the dimensionality diverges, the corresponding eigenvalues are asymptotically the first K leading eigenvalues of $\frac{1}{T}(\mathbf{P}\mathbf{Y})'\mathbf{P}\mathbf{Y}$. This motivates the so-called "projected PCA" (Fan et al. (2014b)), 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 really the PCA to the projected data (smoothed data) $\mathbf{P}\mathbf{Y}$.

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

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

A nice feature of the projected-PCA is that the consistency is achieved even when the sample size T is finite, as shown in Fan et al. (2014b). Thus, it is particularly appealing in the HDLSS context. Intuitively, there are two sources of the approximation errors: (i) $\mathbf{P}$ approximates $\mathbf{P}$ and (ii) the normalized $\mathbf{B}$ approximates the leading eigenvectors of $\mathbf{\Sigma}$. Neither of the approximation errors require a large sample size T in order to be asymptotically negligible. This implies the consistency under a finite T. See Fan et al. (2014b) for more detailed discussions on this aspect.

7.3 Semi-parametric factor model

In the model (45), let $\mathbf{G}(\mathbf{X})$ and $\mathbf{\Gamma}$ respectively denote the $p \times K$ matrices of $g_k(\boldsymbol{X}_i)$ and γ_{ij}. 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: a part that can be explained by $\mathbf{X}$ and the part 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 onto the semi-parametric factor model, Fan et al. (2014b) showed that as $p, J \to \infty$, T may either grow or stay constant,

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

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. (2014b) for more detailed discussions on the impacts of improved rates of convergence in factor models.

8 Discussions

This paper introduces several recent developments on estimating large covariance and precision matrices. We focus on two general approaches: rank-based method and factor model based method. We also extend the usual factor model to a projected PCA setup, and show that the newly introduced projected PCA is appealing in the high-dimensional-low-sample-size scenario. Such an approach has drawn growing attentions in the recent literature on high-dimensional PCA (e.g., Jung and Marron (2009); Shen et al. (2013a,b); 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 in terms of spectral norm errors.

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

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

with observed factors $\{\boldsymbol{f}_t\}$. Here the idiosyncratic components $\boldsymbol{u}_t$'s are heavy-tailed but follow the elliptical distribution. Define the population Kendall's tau correlation between

u_{jt} and u_{kt} as

$$\tau_{u,kj} = \mathbb{P}\left((u_{jt} - \widetilde{u}_{jt})(u_{kt} - \widetilde{u}_k) > 0\right) - \mathbb{P}\left((u_{jt} - \widetilde{u}_j)(u_{kt} - \widetilde{u}_k) < 0\right),$$

where $\widetilde{u}_j$ and $\widetilde{u}_k$ are independent copies of u_{jt} and u_{kt} respectively. Let $\mathbf{R}_u$ be the correlation matrix of $\boldsymbol{u}_t$, and $\mathbf{D}_u$ be the diagonal matrix of the individual standard deviations of $\{u_{jt}\}$. Then $\boldsymbol{\Sigma}_u = \mathbf{D}_u \mathbf{R}_u \mathbf{D}_u$. For elliptical distributions, we have

$$\mathbf{R}_u = [\mathbf{R}_{u,kj}] = \left[\sin\left(\frac{\pi}{2}\tau_{u,kj}\right)\right]. \tag{46}$$

Under the conditional sparsity condition, $\mathbf{R}_u$ is a sparse matrix.

Given the "estimated residuals" $\{\widehat{u}_{it}\}$, the sample version Kendall's tau statistic is

$$\widehat{\tau}_{u,kj} = \frac{2}{T(T-1)} \sum_{t<t'} \operatorname{sign}\left((\widehat{u}_{kt} - \widehat{u}_{kt'})(\widehat{u}_{jt} - \widehat{u}_{jt'})\right)$$

for all $k \neq j$, and $\widehat{\tau}_{u,kj} = 1$ otherwise. We can plug $\widehat{\tau}_{u,kj}$ into (46) and obtain a rank-based error correlation estimator $\widehat{\mathbf{R}}_u = [\widehat{\mathbf{R}}_{u,kj}] = \left[\sin\left(\frac{\pi}{2}\widehat{\tau}_{u,kj}\right)\right]$. We then apply thresholding on $\widehat{\mathbf{R}}_u$ to produce a sparse matrix estimator:

$$\widehat{\mathbf{R}}_u^{\mathcal{T}} = (\widehat{\mathbf{R}}_{u,ij}^{\mathcal{T}})_{p\times p}, \quad \widehat{\mathbf{R}}_{u,ij}^{\mathcal{T}} = \begin{cases} 1, & i = j; \\ h(\widehat{\mathbf{R}}_{u,kj}; \omega_T), & i \neq j. \end{cases}$$

Here $h(.;\omega_T)$ is a general thresholding rule as described in Section 2, with a properly chosen threshold value ω_T. The entry-dependent threshold can also be used. Alternatively, we can apply the nearest positive definite projection to produce a sparse covariance estimator based on $\widehat{\mathbf{R}}_u$.

Given the estimated residuals, standard deviations in $\mathbf{D}_u$ can be estimated similarly as before. Specifically, let $\widehat{m}_j$ be the estimators of $\mathbb{E}u_{jt}^2$ by solving:

$$\sum_{t=1}^{T} \psi\left((\widehat{u}_{jt}^2 - m_j)\sqrt{\frac{2}{TK_{\max}}}\right) = 0, \tag{47}$$

where $K_{\max}$ is an upper bound of $\max_j \operatorname{Var}(u_{jt}^2)$. Then the rank-based estimator of $\mathbf{D}_u$ is a diagonal matrix $\widehat{\mathbf{D}}_u$, whose diagonal elements are $\widehat{\sigma}_{u,j} = \sqrt{\max\{\widehat{m}_j, K_{\min}\}}$, where $K_{\min}$ is a lower bound of $\min_j \mathbb{E}u_{jt}^2$ and is assumed to be known. This leads to the rank-based error covariance estimator:

$$\widehat{\boldsymbol{\Sigma}}_u = \widehat{\mathbf{D}}_u \widehat{\mathbf{R}}_u^{\mathcal{T}} \widehat{\mathbf{D}}_u.$$

When the factors are observable, the residuals should be obtained by estimating $\mathbf{B}$. The robust regression estimator $\widehat{\mathbf{B}}$ can be employed, e.g., L_1 regression. With the esti-

mated $\mathbf{B}$, 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 (that is, applying PCA on the rank covariance matrix of $\boldsymbol{Y}_t$). The theoretical properties of such hybrid estimators are left for future investigations.

References

Ahn, J., Marron, J., Muller, K. M. and Chi, Y.-Y. (2007). The high-dimension, low-sample-size geometric representation holds under mild conditions. *Biometrika* **94** 760–766.

Ahn, S. C. and Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica* **81** 1203–1227.

Alessi, L., Barigozzi, M. and Capasso, M. (2010). Improved penalization for determining the number of factors in approximate factor models. *Statistics & Probability Letters* **80** 1806–1813.

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

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

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

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

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

Belloni, A., Chernozhukov, V. and Wang, L. (2012). Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika* **98** 791–806.

Bickel, P. and Levina, E. (2008). Covariance regularization by thresholding. *Annals of Statistics* **36** 2577–2604.

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics* **37** 1705–1732.

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

Breitung and Tenhofen (2011). Gls estimation of dynamic factor models. *Journal of the American Statistical Association* **106** 1150–1166.

Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association* **106** 672–684.

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

Cai, T. and Zhou, H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *Annals of Statistics* **40** 2389–2420.

Cai, T. T., Ma, Z. and Wu, Y. (2013). Sparse pca: Optimal rates and adaptive estimation. *The Annals of Statistics* **41** 3074–3110.

Campbell, J. Y., Lo, A. W.-C., MacKinlay, A. C. et al. (1997). *The econometrics of financial markets*, vol. 2. princeton University press Princeton, NJ.

Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational Mathematics* **9** 717–772.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* **48** 1148–1185.

Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica* **51** 1281–1304.

Connor, G., Matthias, H. and Oliver, L. (2012). Efficient semiparametric estimation of the fama-french model and extensions. *Econometrica* **80** 713–754.

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

Davis, C. (1963). The rotation of eigenvectors by a perturbation. *Journal of Mathematical Analysis and Applications* **6** 159–173.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: asymptopia? (with discussion). *Journal of the Royal Statistical Society. Series B* 301–369.

El Karoui, N. (2010). High-dimensionality effects in the markowitz problem and other quadratic programs with linear constraints: risk underestimation. *The Annals of Statistics* **38** 3487–3566.

Fama, E. and French, K. (1992). The cross-section of expected stock returns. *Journal of Finance* **47** 427–465.

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

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

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association* **96** 1348–1360.

Fan, J., Liao, Y. and Mincheva, M. (2011). High dimensional covariance matrix estimation in approximate factor models. *Annals of Statistics* **39** 3320–3356.

Fan, J., Liao, Y. and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements (with discussion). *Journal of the Royal Statistical Society, Series B* **75** 603–680.

Fan, J., Liao, Y. and Wang, W. (2014b). Projected principal component analysis in factor models. *Available at SSRN 2450770* .

Fan, J. and Liu, H. (2013). Statistical analysis of big data on pharmacogenomics. *Advanced drug delivery reviews* **65** 987–1000.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics* **32** 928–961.

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

Fang, K.-T., Kotz, S. and Ng, K. W. (1990). *Symmetric Multivariate and Related Distributions, Monographs on Statistics and Applied Probability, 36*. London: Chapman and Hall Ltd. MR1071174.

Frahm, G. and Jaekel, U. (2008). Tyler's m-estimator, random matrix theory, and generalized elliptical distributions with applications to finance. *Random Matrix Theory, and Generalized Elliptical Distributions with Applications to Finance (October 21, 2008)* .

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

Friedman, J., T. Hastie, H. H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics* **1** 302–332.

Gautier, E. and Tsybakov, A. B. (2011). High-dimensional instrumental variables regression and confidence sets. Tech. rep., ENSAE ParisTech.

Hallin, M. and Liška, R. (2007). Determining the number of factors in the general dynamic factor model. *Journal of the American Statistical Association* **102** 603–617.

Hamada, M. and Valdez, E. (2004). *CAPM and option pricing with elliptical distributions*. School of Finance and Economics, University of Technology, Sydney.

Han, F. and Liu, H. (2013). Optimal rates of convergence for latent generalized correlation matrix estimation in transelliptical distribution. *arXiv preprint arXiv:1305.6916* .

Jung, S. and Marron, J. (2009). Pca consistency in high dimension, low sample size context. *The Annals of Statistics* **37** 4104–4130.

Koltchinskii, V., Lounici, K. and Tsybakov, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics* **39** 2302–2329.

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

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

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

Lawley, D. and Maxwell, A. (1971). Factor analysis as a statistical method .

Ledoit, O. and Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of empirical finance* **10** 603–621.

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

Liu, H., Han, F., Yuan, M., Lafferty, J. and Wasserman, L. (2012a). High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics* **40** 2293–2326.

Liu, H., Han, F., Yuan, M., Lafferty, J. and Wasserman, L. (2012b). High dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics (to appear)* .

Liu, H., Han, F. and Zhang, C.-H. (2012c). Transelliptical graphical models. In *Advances in Neural Information Processing Systems 25*.

Liu, H., Lafferty, J. and Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research* **10** 2295–2328.

Liu, H. and Wang, L. (2012). Tiger: A tuning-insensitive approach for optimally estimating gaussian graphical models. Tech. rep., Department of Operations Research and Financial Engineering, Princeton University.

Liu, H., Wang, L. and Zhao, T. (2014a). Sparse covariance matrix estimation with eigenvalue constraints. *Journal of Computational and Graphical Statistics* **23** 439–459.

Liu, H., Wang, L. and Zhao, T. (2014b). Sparse covariance matrix estimation with eigenvalue constraints. *Journal of Computational and Graphical Statistics* **23** 439–459.

Liu, W. and Luo, X. (2012). High-dimensional sparse precision matrix estimation via sparse column inverse operator. *arXiv/1203.3896* .

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

Meinshausen, N. and Bühlmann, P. (2006a). High dimensional graphs and variable selection with the lasso. *The Annals of Statistics* **34** 1436–1462.

Meinshausen, N. and Bühlmann, P. (2006b). High dimensional graphs and variable selection with the lasso. *Annals of Statistics* **34** 1436–1462.

Mitra, R. and Zhang, C.-H. (2014). Multivariate analysis of nonparametric estimates of large correlation matrices. *arXiv preprint arXiv:1403.6195* .

Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics* **39** 1069–1097.

Onatski, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics* **168** 244–258.

Owen, J. and Rabinovitch, R. (1983). On the class of elliptical distributions and their applications to the theory of portfolio choice. *Journal of Finance* **38** 745–752.

Pourahmadi, M. (2013). *High-Dimensional Covariance Estimation: With High-Dimensional Data*. John Wiley & Sons.

Qi, H. and Sun, D. (2006). A quadratically convergent newton method for computing the nearest correlation matrix. *SIAM journal on matrix analysis and applications* **28** 360–385.

Rigollet, P. and Tsybakov, A. (2012). Estimation of covariance matrices under sparsity constraints. *arXiv preprint arXiv:1205.1210* .

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

Rothman, A. J., Levina, E. and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association* **104** 177–186.

Shen, D., Shen, H. and Marron, J. (2013a). Consistency of sparse pca in high dimension, low sample size contexts. *Journal of Multivariate Analysis* **115** 317–333.

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

Shen, X., Pan, W. and Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association* **107** 223–232.

Stock, J. H. and Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American statistical association* **97** 1167–1179.

Stoer, J., Bulirsch, R., Bartels, R., Gautschi, W. and Witzgall, C. (1993). *Introduction to numerical analysis*, vol. 2. Springer New York.

Sun, T. and Zhang, C.-H. (2012). Sparse matrix inversion with scaled lasso. Tech. rep., Department of Statistics, Rutgers University.

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

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

Wainwright, M. (2009). Sharp thresholds for highdimensional and noisy sparsity recovery using $\ell_1$constrained quadratic programming. *IEEE Transactions on Information Theory* **55** 2183–2201.

Wegkamp, M. and Zhao, Y. (2013). Adaptive estimation of the copula correlation matrix for semiparametric elliptical copulas. *arXiv preprint arXiv:1305.6526* .

Wille, A., Zimmermann, P., Vranová, E., Fürholz, A., Laule, O., Bleuler, S., Hennig, L., Prelic, A., von Rohr, P., Thiele, L. et al. (2004). Sparse graphical gaussian modeling of the isoprenoid gene network in arabidopsis thaliana. *Genome Biol* **5** R92.

Wu, W. B. and Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika* **90** 831–844.

Xue, L. and Zou, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *The Annals of Statistics* **40** 2541–2571.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research* **11** 2261–2286.

Yuan, M. and Lin, Y. (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. *The Annals of Statistics* 894–942.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research* **7** 2541–2563.

Zhao, T. and Liu, H. (2014). Calibrated precision matrix estimation for high-dimensional elliptical distributions. *IEEE Transactions on Information Theory* **60** 7874–7887.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101** 1418–1429.