Published in final edited form as:

Ann Stat. 2018 August ; 46(4): 1383–1414. doi:10.1214/17-AOS1588.

LARGE COVARIANCE ESTIMATION THROUGH ELLIPTICAL FACTOR MODELS

Jianqing Fan[*,‡], **Han Liu**[†,‡], and **Weichen Wang**[‡]

[‡]Dept of Operations Research & Financial Engineering, Sherrerd Hall, Princeton University, Princeton, NJ 08544, USA

Abstract

We propose a general Principal Orthogonal complEment Thresholding (POET) framework for large-scale covariance matrix estimation based on the approximate factor model. A set of high level sufficient conditions for the procedure to achieve optimal rates of convergence under different matrix norms is established to better understand how POET works. Such a framework allows us to recover existing results for sub-Gaussian data in a more transparent way that only depends on the concentration properties of the sample covariance matrix. As a new theoretical contribution, for the first time, such a framework allows us to exploit conditional sparsity covariance structure for the heavy-tailed data. In particular, for the elliptical distribution, we propose a robust estimator based on the marginal and spatial Kendall's tau to satisfy these conditions. In addition, we study conditional graphical model under the same framework. The technical tools developed in this paper are of general interest to high dimensional principal component analysis. Thorough numerical results are also provided to back up the developed theory.

Keywords and phrases

principal component analysis; approximate factor model; sub-Gaussian family; elliptical distribution; conditional graphical model; marginal and spatial Kendall's tau

1. Introduction

This paper considers factor model based covariance matrix estimation for heavy-tailed data. Factor model is a powerful tool for dimension reduction and latent factor extraction, which gained its popularity in various applications from finance to biology. When applied to covariance matrix estimation, it assumes a conditional sparse covariance structure, i.e., conditioning on the low dimensional spiked factors, the covariance matrix of the

[*]Jianqing Fan's research was partially supported by NSF grants DMS-1206464 and DMS-1406266 and NIH grants R01GM072611-10 and NIH R01GM100474-04.

[†]Han Liu's research was supported by NSF CAREER Award DMS1454377, NSF IIS1408910, NSF IIS1332109, NIH R01MH102339, NIH R01GM083084, and NIH R01HG06841.

MSC 2010 subject classifications: Primary 62H25; secondary 62H12

SUPPLEMENTARY MATERIAL

Supplement: Technical proofs Fan, Liu and Wang (2015)

(). This supplementary material contains all the remaining proofs and technical lemmas and the comparison of relative error norms.

idiosyncratic errors is sparse. To be specific, consider the approximate factor model in Bai and Ng (2002):

$$y_{it} = \mathbf{b}_i' \mathbf{f}_t + u_{it}, \quad (1.1)$$

where y_{it} is the observed data for the i^{th} ($i = 1,\ldots, p$) dimension at time $t = 1,\ldots, n$; $\mathbf{f}_t$ is an unknown m-dimensional vector of common factors, and $\mathbf{b}_i$ is the factor loading for the i^{th} variable; u_{it} is the idiosyncratic error, uncorrelated with the common factors. Previous works are limited by only considering factors and noises whose distributions are exponentially decayed. In this paper we aim to move beyond this limitation and consider heavy-tailed distributions. More specifically, we will consider as an example the case where factors and noises are elliptically distributed. Under this broad class of heavy-tailed distributions, we aim to understand the underlying mathematical structure of large covariance matrix estimation.

Large-scale covariance matrix estimation has been pioneered by Bickel and Levina (2008a,b) and Fan, Fan and Lv (2008). After that, substantial amount of work has focused on the inference of high-dimensional covariance matrices under unconditional sparsity, that is, the covariance matrix itself is sparse (Cai and Liu, 2011; Cai, Ren and Zhou, 2013; Cai, Zhang and Zhou, 2010; Karoui, 2008; Lam and Fan, 2009; Ravikumar et al., 2011) or conditional sparsity, that is, the covariance matrix is sparse after subtraction by a low-rank component (Amini and Wainwright, 2008; Berthet and Rigollet, 2013a,b; Birnbaum et al., 2013; Cai, Ma and Wu, 2013, 2015; Johnstone and Lu, 2009; Levina and Vershynin, 2012; Rothman, Levina and Zhu, 2009; Ma, 2013; Shen, Shen and Marron, 2013; Paul and Johnstone, 2012; Vu and Lei, 2013; Zou, Hastie and Tibshirani, 2006). This research area is very active, and as a result, this list of references is illustrative rather than comprehensive. To emphasize, Fan and his collaborators proposed to use factor model, which entails a conditional sparsity structure, for covariance matrix estimation (Fan, Fan and Lv, 2008; Fan, Liao and Mincheva, 2011, 2013; Fan, Liao and Wang, 2014). The model encompasses the situation of unconditional sparse covariance by setting the number of factors to zero. Thus it is more general and realistic given the fact that the observed data are usually driven by some common factors.

Another line of research that is related to our work is robust estimation. The idea of robust location estimation dates back to Huber (1964). Adaptive location estimation was later considered for nonparametric symmetric distributions (e.g. Beran (1978); Bickel (1982)). Recently, Catoni (2012) developed the non-asymptotic concentration bound for more general distributions with bounded variance. The result is very useful for high-dimensional problems. Fan, Li and Wang (2016) studied the same problem with a simpler Huber influence function. The ideas of Catoni (2012) and Fan, Li and Wang (2016) can also be used to estimate variances for heavy-tailed data. Another approach for robust variance estimation is to use quantile estimator, such as median absolute deviation (Hampel, 1974) and the Q_n estimator (Rousseeuw and Croux, 1993). In multiple dimensions, robust factor analysis has been studied when the dimension is fixed, for example see Pison et al. (2003)

and Lozeron and Victoria-Feser (2010). The above literature is obviously only a small portion of many tremendous contributions to robust statistics.

In high dimensions, robust covariance estimation has recently received significant attention from the literature. For example, Han and Liu (2013a, 2014) proposed to use the marginal Kendall's tau statistics for estimating large covariance matrix under the elliptical and transelliptical (or elliptical copula) distributions. In addition, spatial Kendall's tau was considered by Han and Liu (2013b) to estimate eigenspaces of covariance matrices in high dimensions. Those methods, applied to PCA or sparse PCA, can be potentially useful for dealing with factor models with heavy-tailed factors and noises. More related references on robust covariance estimation will be provided in Section 4. The goal of this paper is to develop a unified theory that allows us to extend these robust rank-based covariance estimation procedures to handle heavy-tailed data with conditional covariance sparsity.

1.1. Background on approximate factor model

To illustrate how to use factor model as a dimension reduction tool for covariance matrix estimation, let us write model (1.1) in its vector form:

$$\mathbf{y}_t = \mathbf{B}\mathbf{f}_t + \mathbf{u}_t, \quad (1.2)$$

where $\mathbf{y}_t$ contains all observed individuals at time $t = 1,\ldots, n$ and $\mathbf{B} = (\mathbf{b}_1,\ldots, \mathbf{b}_p)'$ is the factor loading matrix. The matrix form of (1.1) is

$$\mathbf{Y} = \mathbf{B}\mathbf{F}' + \mathbf{U}, \quad (1.3)$$

where $\mathbf{Y}_{p\times n}$, $\mathbf{B}_{p\times m}$, $\mathbf{F}_{n\times m}$, $\mathbf{U}_{p\times n}$ are matrices of observed data, factor loadings, factors, and errors with $\mathbf{Y} = (\mathbf{y}_1,\ldots, \mathbf{y}_n)$, $\mathbf{F} = (\mathbf{f}_1,\ldots, \mathbf{f}_n)'$ and $\mathbf{U} = (\mathbf{u}_1,\ldots, \mathbf{u}_n)$. We consider the case where the dimension p is no smaller than the sample size n (i.e., $p \geq n$) and for simplicity we assume n samples are independent and identically distributed. An extension to the dependent setting is possible but with more technicality. We assume the factor matrix $\mathbf{F}$ is unobservable. To make the model (1.1) identifiable, we impose the following condition as in Bai and Ng (2013) and Bai and Li (2012):

$$\mathrm{cov}(\mathbf{f}_t) = \mathbf{I}_m. \quad (1.4)$$

The condition in (1.4) is common in the factor model literature. Under (1.4), the covariance matrix of $\mathbf{y}_t$ is

$$\sum = \mathrm{cov}(\mathbf{y}_t) = \mathbf{B}\mathbf{B}' + \sum_u, \quad (1.5)$$

where $\mathbf{\Sigma}_u$ is the covariance matrix of the idiosyncratic error $\mathbf{u}_t$. In this paper we are interested in separately recovering the low-rank and sparse components. To recover the low-rank

component $\mathbf{BB}'$, we only need to estimate the loading matrix $\mathbf{B}$ up to an orthogonal transformation. In addition, as will be pointed out in Section 2, (1.4) is sufficient only for asymptotic identifiability of recovering $\mathbf{BB}'$ rather than exact identifiability.

1.2. Major contributions of this paper

Under model (1.2), Fan, Liao and Mincheva (2013) proposed the Principal Orthogonal complEment Thresholding (POET) estimator for $\mathbf{\Sigma}$ under the assumption that factors and noises are exponentially decayed. Under the *pervasiveness* condition that $p^{-1}\mathbf{B}'\mathbf{B}$ has spectrum bounded from above and below (Assumption 1 in their paper), implying that the leading eigenvalues of $\mathbf{\Sigma}$ diverge linearly with the dimension p (Assumption 2.1 in the current paper), Fan, Liao and Mincheva (2013) established the consistency and rates of convergence of the POET estimator. However, their proofs are mathematically involved and do not transparently explain why POET works in estimating large covariance matrices. The idea of pervasiveness originates from Chamberlain and Rothschild (1982). It has been pointed out by Fan and Wang (2015) how pervasive factors help in estimating the low-rank component $\mathbf{BB}'$ in (1.5). In the current paper, we further explain the benefit of pervasive structure in a more transparent way, along with a weaker sub-Gaussian assumption (see discussions after Theorem 3.2). A surprising result is that the diverging signal of spiked eigenvalues excludes the necessity of the sparse principal component assumption in sparse PCA literature, comparing with for example Cai, Ma and Wu (2013).

The main contributions of the current paper are two folds. First, we summarize a unified theoretical framework in Section 2.2 for applying POET to various potentially heavy-tailed distributions. The key Theorem 2.1 provides a set of high level interface conditions (1.6) explaining how to design a POET covariance estimator according to factor and error distributions. More specifically, the POET procedure needs three components: initial pilot estimators $\hat{\sum}$, $\hat{\mathbf{\Lambda}}$, $\hat{\mathbf{\Gamma}}$ for covariance matrix $\mathbf{\Sigma}$, its leading eigenvalues $\Lambda = \mathrm{diag}(\lambda_1,\ldots,\lambda_m)$ and their corresponding leading eigenvectors $\mathbf{\Gamma}_{p\times m} = (\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_m)$. Note we will assume distinct leading eigenvalues so that the leading eigenvectors are uniquely determined if we choose the sign by a certain rule, e.g. the j^{th} entry of $\boldsymbol{\xi}_j$ is positive. With these components, a generic POET estimator can be constructed. We will show that such an estimator attains desired rates of convergence as long as

$$\|\hat{\sum} - \sum\|_{\max} = O_P(\sqrt{\log p/n}),\ \|(\hat{\mathbf{\Lambda}} - \mathbf{\Lambda})\mathbf{\Lambda}^{-1}\|_{\max} = O_P(\sqrt{\log p/n}),\ \|\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}\|_{\max} = O_P\ (1.6)$$
$$(\sqrt{\log p/(np)})\,.$$

These conditions are relatively easy to verify, as they involve only the componentwise maximums. Through those sufficient conditions, we are able to separate the deterministic analysis of the estimation procedure based on the theoretical inference in (1.6), and the probabilistic guarantee of the design of initial pilot estimators.

Second, for both sub-Gaussian and elliptical distributions, we provide methods to construct those initial estimators. For sub-Gaussian, it is natural to employ the sample covariance

matrix and its eigenvalues and eigenvectors as the estimates for $\mathbf{\Sigma}$, $\mathbf{\Lambda}$ and $\mathbf{\Gamma}$. We show that such an initialization indeed satisfies the above conditions for sub-Gaussian data, which gives a more transparent explanation on why POET in previous literature works. For elliptical distributions, constructing estimators with the desired rates are nontrivial. We need to use the marginal Kendall's tau to obtain $\hat{\Sigma}$ and $\hat{\mathbf{\Lambda}}$ while a different method involving spatial Kendall's tau is applied to construct $\hat{\mathbf{\Gamma}}$. The final POET estimator separately estimates the eigenvectors and eigenvalues using different methods. To the best of our knowledge, this is the first robust estimator constructed for high dimensional elliptical factor models. This result also illustrates the usefulness of the general theoretical interface in (1.6) as a guide for designing new estimators.

1.3. Notations

If $\mathbf{M}$ is a general matrix, we denote its matrix entry-wise maximum value as $\|\mathbf{M}\|_{\max} = \max_{i,j} |M_{i,j}|$ and define the quantities $\|\mathbf{M}\|_2 = \lambda_{\max}^{1/2}(\mathbf{M}'\mathbf{M})$ (or $\|\mathbf{M}\|$ for short), $\|\mathbf{M}\|_F = \left(\sum_{i,j} M_{i,j}^2\right)^{1/2}$, $\|\mathbf{M}\|_\infty = \max_i \sum_j |M_{i,j}|$ and $\|\mathbf{M}\|_{1,1} = \sum_i \sum_j |M_{i,j}|$ to be its spectral, Frobenius, induced ℓ_∞ and element-wise ℓ_1 norms. If furthermore $\mathbf{M}$ is symmetric, we define $\lambda_j(\mathbf{M})$ to be the j^{th} largest eigenvalue of $\mathbf{M}$ and $\lambda_{\max}(\mathbf{M})$, $\lambda_{\min}(\mathbf{M})$ to be the maximal and minimal eigenvalues respectively. We denote tr($\mathbf{M}$) to be the trace of $\mathbf{M}$. We denote diag($\mathbf{M}_1$,…, $\mathbf{M}_n$) to be the block diagonal matrix with the diagonal block entries as $\mathbf{M}_1$,…, $\mathbf{M}_n$. Here $\mathbf{M}_1$,…, $\mathbf{M}_n$ can be either matrices or just numbers. For any vector $\mathbf{v}$, its ℓ_2 norm is represented by $\|\mathbf{v}\|$ while ℓ_1 and ℓ_0 norms are written as $\|\mathbf{v}\|_1$ and $\|\mathbf{v}\|_\infty$. For two random matrices $\mathbf{A}$, $\mathbf{B}$ of the same size, we say $\mathbf{A} = \mathbf{B} + O_P(\delta)$ if $\|\mathbf{A} - \mathbf{B}\| = O_P(\delta)$ and $\mathbf{A} = \mathbf{B} + O_P(\delta)$ if $\|\mathbf{A} - \mathbf{B} = o_P(\delta)$. The inner product of them is defined as $\langle \mathbf{A}, \mathbf{B} \rangle = \text{tr}(\mathbf{A}'\mathbf{B})$. Similarly for two random vectors $\mathbf{a}$, $\mathbf{b}$ of the same length, we denote $\mathbf{a} = \mathbf{b} + O_P(\delta)$ if $\|\mathbf{a} - \mathbf{b}\| = O_P(\delta)$ and $\mathbf{a} = \mathbf{b} + o_P(\delta)$ if $\|\mathbf{a} - \mathbf{b}\| = o_P(\delta)$. We denote $\mathbf{a} \overset{d}{=} \mathbf{b}$ if random vectors $\mathbf{a}$ and $\mathbf{b}$ have the same distribution. The inner product of them is defined as $\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}'\mathbf{b}$. For two sequences $s_{n,p}$ and $w_{n,p}$, we denote $s_{n,p} \asymp w_{n,p}$ if there exists constant $c_1, c_2 > 0$ such that $c_1 s_{n,p} \leq w_{n,p} \leq c_2 s_{n,p}$ In the sequel, C is a generic constant that may differ from line to line.

1.4. Paper organization

In Section 2, we present a generic POET estimating procedure and a high level theoretical interface (1.6) which ensures the consistency of the generic procedure for factor-based conditional sparsity models. In Section 3, we verify that the conditions in (1.6) hold with high probability for sub-Gaussian data, which provides a transparent understanding of the mechanism of the POET methodology. In Section 4, we propose a new method using a combination of the marginal and spatial Kendall's tau estimators and show that these estimators satisfy the theoretical interface in (1.6) under elliptical factor models. Thorough numerical simulations are conducted to illustrate the merits of our proposed method in Section 5. In Section 6, we conclude the paper with a short discussion. Some technical proofs are relegated to the appendix and supplementary material.

2. A high-level theoretical interface

In this section, we summarize a generic POET procedure and provide a set of high level sufficient conditions for consistent covariance estimation when $p \geq n$. Before doing that, let us review what has been achieved in literature where both the factors and noises are assumed sub-Gaussian.

2.1. Spiked covariance model

Suppose the observed random variables $\{\mathbf{y}_i\}_{i=1}^n$ have zero mean and covariance matrix $\mathbf{\Sigma}_{p\times p}$ where the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ of $\mathbf{\Sigma}$ are ordered in descending order. We consider the spiked population model as suggested by the approximate factor structure (1.5). Specifically we have the following assumption on the eigvenvalues.

Assumption 2.1 (Spiked covariance model)—Let m be a fixed constant that does not change with n and p such that $m \leq p$. As $p \to \infty$, $\lambda_1 > \lambda_2 > \cdots > \lambda_m \gg \lambda_{m+1} \geq \cdots \geq \lambda_p > 0$, where the spiked eigenvalues are linearly proportional to dimension p while the non-spiked eigenvalues are bounded, i.e., $c_0 \leq \lambda_j \leq C_0$, $j > m$ for constants c_0, $C_0 > 0$. In addition, the average of the non-spiked eigenvalues $(p-m)^{-1} \sum_{j=m+1}^{p} \lambda_j = \bar{c} + o(1)$.

Assumption 2.1 divides the eigenvalues into the diverging and bounded ones. For simplicity, we only consider distinguishable eigenvalues (multiplicity 1) for the largest m eigenvalues. This assumption is typically satisfied by the factor model (1.1) with pervasive factors. More specifically, if the factor loadings $\{\mathbf{b}_j\}_{j=1}^p$ are i.i.d. samples from a population with a finite second moment, then by the strong law of large numbers, $p^{-1}\mathbf{B}'\mathbf{B} = p^{-1}\sum_{j=1}^p \mathbf{b}_j\mathbf{b}_j' \to \sum_b$ almost surely, where $\sum_b = \mathbb{E}(\mathbf{b}_j\mathbf{b}_j')$. In other words, the eigenvalues of $\mathbf{BB}'$ are approximately

$$p\lambda_1(\sum_b)(1+o(1)), \cdots, p\lambda_m(\sum_b)(1+o(1)), 0, \cdots, 0,$$

where $\lambda_j(\mathbf{\Sigma}_b)$ is the j^{th} eigenvalue of $\mathbf{\Sigma}_b$. If we further assume that $\|\mathbf{\Sigma}_u\|$ is bounded, by Weyl's theorem, we conclude

$$\lambda_j = p\lambda_j(\sum_b)(1+o(1)), \quad \text{for } j = 1, \cdots, m, \qquad (2.1)$$

and the remaining are bounded.

2.2. A review of POET procedure for covariance estimation

We see from (1.5) that the population covariance of the factor model (1.1) exhibits a low-rank plus sparse structure, if $\mathbf{\Sigma}_u$ is sparse, whose sparsity level is measured by

$$m_p := \max_{i \le p} \sum_{j \le p} |\sigma_{u,ij}|^q$$

for some $q \in [0, 1]$. In particular, with $q = 0$, m_p corresponds to the maximum number of nonzero elements in each row of $\mathbf{\Sigma}_u$.

To estimate the covariance matrix $\mathbf{\Sigma}$ with the approximate factor structure (1.5), Fan, Liao and Mincheva (2013) proposed the POET method to recover the factor matrix as well as the factor loadings. The idea is to first decompose the sample covariance matrix into the spiked and non-spiked parts,

$$\hat{\mathbf{\Sigma}}_Y = \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i \mathbf{y}_i' = \sum_{j=1}^{m} \hat{\lambda}_j \hat{\boldsymbol{\xi}}_j \hat{\boldsymbol{\xi}}_j' + \hat{\mathbf{\Sigma}}_u, \quad (2.2)$$

where $\hat{\Sigma}_u = \sum_{j=m+1}^{p} \hat{\lambda}_j \hat{\boldsymbol{\xi}}_j \hat{\boldsymbol{\xi}}_j'$ is called the principal orthogonal complement. Then by employing adaptive thresholding on $\hat{\Sigma}_u$ to get $\hat{\Sigma}_u^{\mathcal{T}}$ (Cai and Liu, 2011), they obtain a final covariance estimator $\hat{\Sigma}^{\mathcal{T}}$ defined as

$$\hat{\mathbf{\Sigma}}^{\mathcal{T}} = \sum_{j=1}^{m} \hat{\lambda}_j \hat{\boldsymbol{\xi}}_j \hat{\boldsymbol{\xi}}_j' + \hat{\mathbf{\Sigma}}_u^{\mathcal{T}}. \quad (2.3)$$

The above procedure can be equivalently viewed as a least-square approach. That is, the factor and loading matrices are estimated by solving the following nonconvex minimization problem:

$$(\hat{\mathbf{B}}, \hat{\mathbf{F}}) = \arg\min_{\mathbf{B},\mathbf{F}} \|\mathbf{Y} - \mathbf{B}\mathbf{F}'\|_F^2 \, \text{s.t.} \, \frac{1}{n}\mathbf{F}'\mathbf{F} = \mathbf{I}_m, \mathbf{B}'\mathbf{B} \text{ is diagonal}. \quad (2.4)$$

It is shown that the columns of $\hat{\mathbf{F}}/\sqrt{n}$ are the eigenvectors corresponding to the m largest eigenvalues of the $n \times n$ matrix $n^{-1}\mathbf{Y}'\mathbf{Y}$ and $\hat{\mathbf{B}} = n^{-1}\mathbf{Y}\hat{\mathbf{F}}$. Note that the estimator $\hat{\mathbf{B}}$ given by minimizing (2.4), after normalization, is actually the first m empirical eigenvectors of the sample covariance matrix $\hat{\Sigma}_Y = n^{-1}\mathbf{Y}\mathbf{Y}'$. The optimizer can be obtained via the Eckart-Young theorem. Given $\hat{\mathbf{B}}, \hat{\mathbf{F}}$, we define $\hat{\mathbf{U}} = \mathbf{Y} - \hat{\mathbf{B}}\hat{\mathbf{F}}'$ and $\hat{\Sigma}_u = n^{-1}\hat{\mathbf{U}}\hat{\mathbf{U}}' = \hat{\Sigma} - \hat{\mathbf{B}}\hat{\mathbf{B}}'$ Finally adaptive thresholding is applied to $\hat{\Sigma}_u$ to obtain $\hat{\Sigma}_u^{\mathcal{T}} = (\hat{\sigma}_{u,ij}^{\mathcal{T}})_{p \times p}$ with

$$\hat{\sigma}^{\mathcal{T}}_{u,ij} = \begin{cases} \hat{\sigma}_{u,ij}, & i = j \\ s_{ij}(\hat{\sigma}_{u,ij}) I(|\hat{\sigma}_{u,ij}| \geq \tau_{ij}), & i \neq j \end{cases}, \quad (2.5)$$

where $s_{ij}(\cdot)$ is the generalized shrinkage function (Antoniadis and Fan, 2001; Rothman, Levina and Zhu, 2009) and $\tau_{ij} = \tau(\hat{\sigma}_{u,ii}\hat{\sigma}_{u,jj})^{1/2}$ is an entry-dependent threshold. The above adaptive threshold operator corresponds to applying thresholding with parameter τ to the correlation matrix of $\hat{\sum}_u$. The positive parameter τ will be determined based on theoretical analysis.

Let $w_n = \sqrt{\log p/n} + 1/\sqrt{p}$. Fan, Liao and Mincheva (2013) claimed that under some technical assumptions (exponentially decayed factors and noises with mixing time dependency), with $\tau \asymp w_n$, if $m_p w_n^{1-q} = o(1)$,

$$\| \hat{\sum}^{\mathcal{T}}_u - \sum_u \|_{\max} = O_P(w_n), \| \hat{\sum}^{\mathcal{T}}_u - \sum_u \|_2 = O_P\left(m_p w_n^{1-q}\right) = \| (\hat{\sum}^{\mathcal{T}}_u)^{-1} - \sum_u^{-1} \|_2, \quad (2.6)$$

and

$$\| \hat{\sum}^{\mathcal{T}} - \sum \|_{\max} = O_P(w_n), \| \hat{\sum}^{\mathcal{T}} - \sum \|_{\sum} = O_P\left(\frac{\sqrt{p}\log p}{n} + m_p w_n^{1-q}\right), \| (\hat{\sum}^{\mathcal{T}})^{-1} - \sum^{-1} \|_2 = O_P\left(m_p w_n^{1-q}\right), \quad (2.7)$$

where $\|\mathbf{A}\|_{\mathbf{\Sigma}} = p^{-1/2}\|\mathbf{\Sigma}^{-1/2}\mathbf{A}\mathbf{\Sigma}^{-1/2}\|_F$ is the relative Frobenius norm. The scaling $p^{-1/2}$ is exploited to ensure $\|\mathbf{\Sigma}\|_{\mathbf{\Sigma}} = 1$. The term $1/\sqrt{p}$ in w_n is the price we need to pay for estimating the unknown factors. The original proofs for getting the above rates are mathematically involved and it is not clear why the above rates are attained, especially with no sparsity assumption for eigenvectors imposed as in sparse PCA literature.

So our question is: why such a simple POET procedure works under the spiked covariance assumption (2.1)? Can we replace the sample covariance matrix by other pilot estimators as a starting point for the eigendecomposition if other family of distributions, such as elliptical distributions or other more general heavy-tailed distributions, are considered?

2.3. A generic procedure and a high level theoretical interface

We propose a generic POET procedure here:

1. Given three initial pilot estimators $\hat{\Sigma}$, $\hat{\mathbf{\Lambda}}$ $\hat{\mathbf{\Gamma}}$ for true covariance matrix $\mathbf{\Sigma}$, leading eigenvalues $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1,\ldots,\lambda_m)$ and leading eigenvectors $\mathbf{\Gamma}_{p\times m} = (\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_m)$ respectively, the principal orthogonal complement $\hat{\Sigma}_u$ can be computed by subtracting out the leading low-rank part, i.e.,

$$\hat{\Sigma}_u = \hat{\Sigma} - \hat{\mathbf{\Gamma}}\hat{\mathbf{\Lambda}}\hat{\mathbf{\Gamma}}';$$

2. The adaptive thresholding (2.5) is applied to $\hat{\Sigma}_u$ to obtain $\hat{\Sigma}_u^{\mathcal{T}}$;

3. The low-rank structure is added back to obtain $\hat{\Sigma}^{\mathcal{T}}$.

The advantage of the above procedure is modular: three initial components can be constructed separately. For sub-Gaussian distributions, $\hat{\mathbf{\Lambda}} = \mathrm{diag}(\hat{\lambda}_1,\ldots,\hat{\lambda}_m)$ is the diagonal matrix constructed by the first m leading empirical eigenvalues of the sample covariance matrix $\hat{\Sigma}_Y$ while $\hat{\mathbf{\Gamma}} = (\hat{\boldsymbol{\xi}}_1,\ldots,\hat{\boldsymbol{\xi}}_m)$ is the matrix of the corresponding leading empirical eigenvectors and $\hat{\mathbf{B}}\hat{\mathbf{B}}' = \hat{\mathbf{\Gamma}}\hat{\mathbf{\Lambda}}\hat{\mathbf{\Gamma}}'$ But in general, $\hat{\mathbf{\Lambda}}$ and $\hat{\mathbf{\Gamma}}$ do not have to come from the sample covariance matrix. In fact, they can even be separately estimated.

A high level explanation is provided to understand the generic POET procedure. Sufficient conditions are established for $\hat{\Sigma}_u^{\mathcal{T}}$ and $\hat{\Sigma}^{\mathcal{T}}$ to achieve the desired rates of convergence in (2.6) and (2.7). Our vital conclusion is stated in the following theorem.

Theorem 2.1—Suppose Assumption 2.1 holds, and in addition there exists $C > 0$ such that $\|\mathbf{B}\|_{\max}$, $\|\Sigma_u^{-1}\|_2$, $\|\mathbf{\Sigma}_u\|_2 \le C$. If $m_p w_n^{1-q} = o(1)$ with $w_n = \sqrt{\log p/n} + 1/\sqrt{p}$ and if we have estimators $\hat{\Sigma}$, $\hat{\mathbf{\Gamma}}$, $\hat{\mathbf{\Lambda}}$ satisfying (1.6), then the rates of convergence in (2.6) and (2.7) hold with the generic POET procedure described above. In addition, (2.8) and (2.9) below hold.

The proof of Theorem 2.1 (given in Appendix A) provides insights on how the generic POET procedure works. Note that in (A.1) of Appendix A, the max norm of the low-rank matrix estimation error is bounded by Δ_1 and Δ_2 that we briefly describe here. The term Δ_1 quantifies the estimation error of leading empirical eigen-structure $\hat{\mathbf{\Gamma}}$ $\hat{\mathbf{\Lambda}}$ $\hat{\mathbf{\Gamma}}'$ for its population counterpart, and is of order $\Delta_1 = O_P(\sqrt{\log p/n})$. The term Δ_2 measures the error of identifying the low-rank matrix $\mathbf{BB}'$ by $\mathbf{\Gamma\Lambda\Gamma}'$. This identification under pervasiveness condition is asymptotically unique with identification error $\Delta_2 = O(1/\sqrt{p})$. The asymptotic identifiability is consistent with the sufficient non-asymptotic identification condition given by Chandrasekaran et al. (2011) and Hsu, Kakade and Zhang (2011).

We comment on the optimality of rates in (2.6) and (2.7). Firstly, note that throughout the paper, we consider the high dimensional regime $p \ge n$ so that $1/\sqrt{p} \le \sqrt{\log p/n}$, making the term $1/\sqrt{p}$ negligible. We leave it there just for illustrating the asymptotic identifiability. Secondly, it is not hard to see (2.6) and the first and third conclusions of (2.7) are optimal

from an application of Cai and Zhou (2012). See also Tsybakov (2009) for more details on lower bound construction. In terms of the relative Frobenius norm $\|\cdot\|_{\mathbf{\Sigma}}$, we obtain the same rate of convergence as in Fan, Liao and Mincheva (2013). However, this rate is not optimal (We thank an anonymous referee for pointing this out). Interestingly, by using a naive estimator $\hat{\sum}_u^{\mathcal{T}}$ to estimate $\mathbf{\Sigma}$, we are able to show

$$\|\hat{\sum}_u^{\mathcal{T}} - \sum\|_{\sum} \le C_p^{-1/2}(\|\hat{\sum}_u^{\mathcal{T}} - \sum_u\|_F + 1) = O_P\big(m_p^{1/2} w_n^{1-q/2}\big), \quad (2.8)$$

which turns out to be optimal according to Cai and Zhou (2012). In (2.8), the optimal rate for $\|\hat{\sum}_u^{\mathcal{T}} - \sum_u\|_F$ is also presented. This means in terms of $\|\cdot\|_{\mathbf{\Sigma}}$, $\hat{\sum}_u^{\mathcal{T}}$ is a much better estimator in estimating $\mathbf{\Sigma}$ than $\hat{\sum}^{\mathcal{T}}$ itself. We will verify this by simulations in Section 5.1. How shall we explain this odd property? We provided a detailed discussion on this point in Appendix C in the supplementary material (Fan, Liu and Wang, 2015). The key insight is that the relative Frobenius norm is not an ideal criterion to characterize the performance of the POET procedure. To appropriately evaluate POET in a relative sense, we should not entangle the effects of the low-rank and sparse components. To solve this problem, we propose a more suitable relative norm to characterize the low-rank component in (2.10) below. Further comparison of the two relative norms can also be found in Appendix C. Thirdly, in terms of estimating the low-rank matrix, we have

$$\|\hat{\mathbf{\Gamma}}\hat{\mathbf{\Lambda}}\hat{\mathbf{\Gamma}} - \mathbf{B}\mathbf{B}'\|_{\max} = O_P(w_n), \|\hat{\mathbf{\Gamma}}\hat{\mathbf{\Lambda}}\hat{\mathbf{\Gamma}} - \mathbf{B}\mathbf{B}'\|_2 \le \hat{\mathbf{\Gamma}}\hat{\mathbf{\Lambda}}\hat{\mathbf{\Gamma}} - \mathbf{B}\mathbf{B}'\|_F = O_P\big(pw_n\big), \quad (2.9)$$

which implies the following more appropriate notion of relative error (simply a normalization of (2.9) by magnitude of leading eigenvalues):

$$\|\mathbf{L}^{-1/2}(\hat{\mathbf{\Gamma}}\hat{\mathbf{\Lambda}}\hat{\mathbf{\Gamma}}' - \mathbf{B}\mathbf{B}')\mathbf{L}^{-1/2}\|_F = O_P(\sqrt{\log p/n}), \quad (2.10)$$

where $\mathbf{L}^{-1/2} = \mathbf{\Gamma}\mathbf{\Lambda}^{-1/2}\mathbf{\Gamma}'$. Cai, Ma and Wu (2015) showed that the rate for spectral norm in (2.9) is optimal up to the log p (so is the rate for the Frobenius norm and in (2.10)), when the leading eigenvalues are of order p and leading eigenvectors are dense. Wegkamp and Zhao (2016) proved the same rate of convergence for the Frobenius norm of the low-rank matrix under the strict factor model.

In sum, under the pervasiveness condition, we are able to (near) optimally recover the low-rank and sparse matrices separately, in terms of the norms $\|\hat{\sum}_u^{\mathcal{T}} - \sum_u\|_{\max \text{ or } 2 \text{ or } F}$ and $\|\hat{\mathbf{\Gamma}}\hat{\mathbf{\Lambda}}\hat{\mathbf{\Gamma}}' - \mathbf{B}\mathbf{B}'\|_{\max \text{ or } 2 \text{ or } F}$ (thus $\|\hat{\sum}^{T} - \sum\|_{\max \text{ or } 2 \text{ or } F}$) and furthermore $\|(\hat{\sum}^{\mathcal{T}})^{-1} - \sum^{-1}\|_2$ and a better notion of the relative error $\|\mathbf{L}^{-1/2}(\hat{\mathbf{\Gamma}}\hat{\mathbf{\Lambda}}\hat{\mathbf{F}}' - \mathbf{B}\mathbf{B}')\mathbf{L}^{-1/2}\|_F$. We postpone more discussions on the role of pervasiveness in Section 6.

2.4. Conditional graphical model

In Section 2.2, m_p measures the sparsity of $\mathbf{\Sigma}_u$, but its inverse $\mathbf{\Omega}_u = \sum_u^{-1}$ is not necessarily sparse. Sometimes, the sparsity structure on $\mathbf{\Omega}_u$ reveals more interesting structure than $\mathbf{\Sigma}_u$. For example, if $\mathbf{u}_t$ follows an elliptical distribution, i.e., $\mathbf{u}_t \sim \mathrm{ED}_p(\mathbf{0}, \mathbf{\Sigma}_u, \zeta)$ using the notation (4.1) below, the sparsity of $\mathbf{\Omega}_u$ encodes the conditional uncorrelatedness relationships between all variables in the p dimensional vector $\mathbf{u}_t$. Here ED is short for elliptical distribution. More specifically, for p nodes $u_1, \ldots, u_p$, each corresponding to one element of $\mathbf{u}_t$, u_i and u_j are connected if and only if $(\mathbf{\Omega}_u)_{ij} \neq 0$, meaning that u_{it} and u_{jt} are uncorrelated conditioning on all the other $\{u_{kt}\}_{k \neq i,j}$ and $\mathbf{f}_t$. If the number of factors is zero, this reduces to the classical elliptical graphical model, studied by Vogel and Fried (2011) and Liu, Han and Zhang (2012).

In many applications, the conditional graphical model (or conditional sparse inverse covariance model) appears more natural compared to the unconditional graphical model. For example, to estimate the stock network, it is more meaningful to take out the common market factors from the return data and study the conditional independence relationships between idiosyncratic components; in genomics, the conditional independence graph after taking the confounding factors such as age and environment exposure are of better interest. The factors can also be interpreted as covariates to be adjusted before analyzing the correlatedness of the residual part (Fan et al., 2016). Rothman, Levina and Zhu (2010) and Cai et al. (2013) adopted the same idea of adjusting the factors in predicting asset returns and analyzing genomics data, but they do not impose the pervasiveness condition, instead they need to impose the constraint of a sparse factor loading matrix $\mathbf{B}$. In this paper, we only put the sparsity on $\mathbf{\Omega}_u$, measured by the quantity

$$M_p := \max_{i \leq p} \sum_{j \leq p} |\omega_{u,ij}|^q.$$

The generic POET procedure could also be modified to estimate conditional graphical model. The first step is still recovering $\hat{\sum}_u = \hat{\sum} - \hat{\mathbf{\Gamma}}\hat{\mathbf{\Lambda}}\hat{\mathbf{\Gamma}}'$ by removing the effect of the low-rank dominating factors. Then the method "constrained ℓ_1-minimization for inverse matrix estimation" (CLIME) proposed by Cai, Liu and Luo (2011) can be applied to obtain $\hat{\mathbf{\Omega}}_u$. Specifically, CLIME solves the following constrained minimization problem:

$$\hat{\mathbf{\Omega}}_u^1 = \mathrm{argmin}_{\mathbf{\Omega}} \|\mathbf{\Omega}\|_{1,1} \text{ subject to } \|\hat{\sum}_u \mathbf{\Omega} - \mathbf{I}\|_{\max} \leq \tau, \quad (2.11)$$

where $\|\mathbf{\Omega}\|_{1,1} = \sum_i \sum_j |\omega_{i,j}|$ and τ is a tuning parameter so that $\tau \asymp w_n$. A further symmetrization step can be carried out to guarantee a symmetric estimator $\hat{\mathbf{\Omega}}_u = (\hat{\omega}_{u,ij})$ where

$$\hat{\omega}_{u,ij} = \hat{\omega}^1_{u,ij} 1(|\hat{\omega}^1_{u,ij}| \le |\hat{\omega}^1_{u,ji}|) + \hat{\omega}^1_{u,ji} 1(|\hat{\omega}^1_{u,ij}| > |\hat{\omega}^1_{u,ji}|). \quad (2.12)$$

Note that (2.11) can be solved column by column using a linear program solver. We can either apply the interior point method (with polynomial time worst case complexity) or simplex algorithm (with superior average case complexity, but exponential worst case complexity). Other possible methods include the graphical Lasso, graphical SCAD, graphical Dantzig selector, and graphical neighborhood selection (Friedman, Hastie and Tibshirani, 2008; Yuan and Lin, 2007; Fan, Feng and Wu, 2009; Lam and Fan, 2009; Ravikumar et al., 2011; Yuan, 2010; Meinshausen and Bühlmann, 2006). Though substantial amount of efforts have been made to understand the graphical model, little has been done for estimating conditional graphical model, which is more general and realistic.

Once we have $\hat{\mathbf{\Omega}}_u$, the original inverse covariance matrix $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ can be estimated using the Sherman-Morrison-Woodbury formula as follows:

$$\hat{\mathbf{\Omega}} = \hat{\mathbf{\Omega}}_u - \hat{\mathbf{\Omega}}_u \hat{\mathbf{\Gamma}} (\hat{\mathbf{\Lambda}}^{-1} + \hat{\mathbf{\Gamma}}' \hat{\mathbf{\Omega}}_u \hat{\mathbf{\Gamma}})^{-1} \hat{\mathbf{\Gamma}}' \hat{\mathbf{\Omega}}_u . \quad (2.13)$$

The following theorem gives the rates of convergence for $\hat{\mathbf{\Omega}}_u$ and $\hat{\mathbf{\Omega}}$ provided good pilot estimators $\hat{\sum}$, $\hat{\mathbf{\Lambda}}$ and $\hat{\mathbf{\Gamma}}$. Its proof is in Appendix B in the supplementary material (Fan, Liu and Wang, 2015)

Theorem 2.2—Under Assumptions 2.1, if there exists $C > 0$ such that $\|\mathbf{B}\|_{\max}$, $\|\mathbf{\Omega}_u^{-1}\|_2$, $\|\mathbf{\Omega}_u\|_\infty \le C$, $M_p w_n^{1-q} = o(1)$ and we have estimators $\hat{\sum}$, $\hat{\mathbf{\Gamma}}$, $\hat{\mathbf{\Lambda}}$ satisfying conditions (1.6), then the generic POET procedure with CLIME gives

$$\|\hat{\mathbf{\Omega}}_u - \mathbf{\Omega}_u\|_{\max} = O_P(w_n) = \|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\|_{\max}, \|\hat{\mathbf{\Omega}}_u - \mathbf{\Omega}_u\|_2 = O_P(M_p w_n^{1-q}) = \|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\|_2 \quad (2.14)$$

Note the assumption of bounded $\|\mathbf{\Omega}_u\|_\infty$ is stronger than the case of estimating covariance matrix in Theorem 2.1. This condition might be relaxed if graphical model estimation methods other than CLIME are applied. We do not pursue the weakest possible condition here. Many potential applications only involve the estimation of inverse covariance matrix $\mathbf{\Omega}$, for instance classification and discriminant analysis and optimal portfolio allocation.

2.5. Positive semi-definite projection under max norm

There is an additional issue that requires careful consideration. In the generic POET procedure, if $\hat{\mathbf{\Gamma}}$ and $\hat{\mathbf{\Lambda}}$ are not estimated by the leading empirical eigenvectors and leading empirical eigenvalues of some (and the same) positive semi-definite (PSD) matrix $\hat{\sum}$, the

residual $\hat{\Sigma}_u$ may not be PSD for a given sample. Thus, the following optimization should be considered to find the nearest PSD matrix of $\hat{\Sigma}_u$ in terms of the max norm:

$$\tilde{\Sigma}_u \operatorname{argmin}_{\Sigma_u \geq \mathbf{0}} \|\hat{\Sigma}_u - \Sigma_u\|_{\max}. \quad (2.15)$$

The minimizer preserves the max norm error bound since

$$\|\tilde{\Sigma}_u - \Sigma_u\|_{\max} \leq \|\tilde{\Sigma}_u - \hat{\Sigma}_u\|_{\max} + \|\hat{\Sigma}_u - \Sigma_u\|_{\max} \leq 2\|\hat{\Sigma}_u - \Sigma_u\|_{\max},$$

and everything else in the POET procedure works with $\hat{\Sigma}_u$ replaced by $\tilde{\Sigma}_u$. The same problem occurs in conditional graphical model estimation. Although $\hat{\mathbf{\Omega}}_u$ is PSD with high probability, in practice we may get a non-PSD estimator for $\mathbf{\Omega}_u$. So we need to explicitly perform the PSD projection of $\hat{\mathbf{\Omega}}_u$ onto the PSD cone as in (2.15).

Minimization (2.15) is challenging due to its non-smoothness. An effective smooth surrogate for the max norm objective was proposed by Zhao, Roeder and Liu (2014) which can be solved efficiently. Specifically, they considered minimizing $\|\hat{\Sigma}_u - \Sigma_u\|_{\max}^{\mu}$ subject to $\Sigma_u \geq \mathbf{0}$ with

$$\|\mathbf{A}\|_{\max}^{\mu} = \max_{\|\mathbf{U}\|_{1,1} \leq 1} \langle \mathbf{U}, \mathbf{A} \rangle - \frac{\mu}{2}\|\mathbf{U}\|_F^2,$$

where $\|\mathbf{U}\|_{1,1} = \sum_{i,j} |u_{ij}|$. More details can be found in their paper. An alternative approach is to solve the dual problem of graphical lasso, that is,

$$\max_{\mathbf{W}} \log\det(\mathbf{W}) \text{ subject to } \|\mathbf{W} - \hat{\Sigma}\|_{\max} \leq \tau.$$

By choosing $\tau \asymp w_n$, the optimal solution is a PSD matrix satisfying the max norm bound. Such a projection is still valid for the generic POET procedure to get the optimal convergence rates.

3. Sub-Gaussian factor models

We have established sufficient conditions in (1.6) for optimal estimation of covariance matrices as well as conditional graphical models. The next natural question is whether these conditions hold for sub-Gaussian factor models. In this section, we validate the conditions for sample covariance matrix under sub-Gaussian conditions.

By the spectral decomposition, $\sum = \mathbf{\Gamma}_p \mathbf{\Lambda}_p \mathbf{\Gamma}_p'$ where $\mathbf{\Lambda}_p = \text{diag}(\lambda_1, \ldots, \lambda_p)$ and $\mathbf{\Gamma}_p$ is constructed by all the corresponding eigenvectors of $\mathbf{\Sigma}$. We use subscript p to explicitly denote the dependence of $\mathbf{\Lambda}_p$ and $\mathbf{\Gamma}_p$ on all eigenvalues or eigenvectors rather than just spiked ones. Let $\mathbf{x}_i = \mathbf{\Gamma}_p' \mathbf{y}_i$. So $\mathbf{x}_i$ has mean zero and diagonal covariance matrix $\mathbf{\Lambda}_p$. Since under orthogonal transformations of the data, the empirical eigenvalues of sample covariance are invariant and the empirical eigenvectors are equivariant, the analysis will be done on $\mathbf{x}_i$'s which naturally extends to our original data $\mathbf{y}_i$'s by a simple affine transformation. The following assumption on $\mathbf{x}_i$ is imposed.

Assumption 3.1 (Sub-Gaussian distribution)

Let $\mathbf{z}_i = \mathbf{\Lambda}_p^{-1/2} \mathbf{x}_i$ be the standardized version of the transformed data $\mathbf{x}_i$. $\mathbf{z}_i$'s are iid samples of sub-Gaussian isotropic random vector $\mathbf{z}$, i.e., $\|\mathbf{z}\|_{\phi_2} = \sup_{\mathbf{u} \in \mathcal{S}^{p-1}} \|\langle \mathbf{z}, \mathbf{u} \rangle\|_{\phi_2} \leq M$ for some constant $M > 0$ where the sub-Gaussian norm is defined as $\|\langle \mathbf{z}, \mathbf{u} \rangle\|_{\phi_2} = \sup_{p \geq 1} p^{-1/2} (\mathbb{E} |\langle \mathbf{z}, \mathbf{u} \rangle|^p)^{1/p}$ Furthermore, we assume there exist M_1, $M_2 > 0$ such that for $0 \leq \theta \leq M_1$,

$$\mathbb{E}\left[\exp\left(-\theta \sum_{j=1}^{p} (z_j^2 - 1)\right)\right] \leq \exp(M_2 \theta^2 p). \quad (3.1)$$

The above assumption requires a slightly stronger condition than the classical sub-Gaussian condition for $\mathbf{z}$. It has to satisfy (3.1) for technical reasons discussed in Lemma D.2 in Appendix D in the supplementary material (Fan, Liu and Wang, 2015). This assumption is clearly satisfied if $\mathbf{z}$ has independent elements of sub-Gaussian variables (Vershynin, 2010) although it could also hold for weakly dependent sub-Gaussian vectors.

Under this assumption, trivially the first condition in (1.6) holds for the sample covariance matrix $\hat{\sum}_Y$ of $\mathbf{y}_i$, i.e., $\|\hat{\sum}_Y - \sum\|_{\max} = O_P(\sqrt{\log p / n})$ because from (1.5) the maximal element-wise variance of $\mathbf{y}_i$ is bounded. Next, we present two theoretical properties respectively on leading empirical eigenvalues $\{\hat{\lambda}_j\}_{j=1}^m$ and eigenvectors o $\{\hat{\xi}_j\}_{j=1}^m$ of the sample covariance matrix $\hat{\sum}_X$ of $\mathbf{x}_i$'s. These properties are useful for us to verify the remaining conditions of the high level theoretical interface (1.6).

Theorem 3.1

Under Assumptions 2.1 and 3.1, for $j \leq m$ we have

$$\left|\hat{\lambda}_j / \lambda_j - 1\right| = O_P(n^{-1/2}),$$

where $\hat{\lambda}_j = \lambda_j(\hat{\sum}_X)$ is the j^{th} largest eigenvalue of $\hat{\sum}_X$.

Consider the empirical eigenvectors $\hat{\boldsymbol{\xi}}_j$ of $\hat{\boldsymbol{\Sigma}}_X$ for $j \le m$. Each $\hat{\boldsymbol{\xi}}_j$ is divided into two parts $\hat{\boldsymbol{\xi}}_j = (\hat{\boldsymbol{\xi}}'_{jA}, \hat{\boldsymbol{\xi}}'_{jB})'$, where $\hat{\boldsymbol{\xi}}_{jA}$ is of length m corresponding to the spiked component and $\hat{\boldsymbol{\xi}}_{jB}$ corresponds to the non-spiked component. Note that $\hat{\boldsymbol{\xi}}_j$ is uniquely determined up to sign since eigenvalues are well separated and we always choose the right sign such that $\hat{\boldsymbol{\xi}}'_j \mathbf{e}_j \ge 0$.

Theorem 3.2

Under Assumptions 2.1 and 3.1, for $j \le m$ we have

i. $\|\hat{\boldsymbol{\xi}}_{jA} - \mathbf{e}_{jA}\| = O_P(n^{-1/2})$ of length m with 1 P), where $\mathbf{e}_{jA}$ is unit vector at the j^{th} coordinate and 0 everywhere else;

ii. $\|\boldsymbol{\Omega}\hat{\boldsymbol{\xi}}_{jB}\|_{\max} = O_P(\sqrt{\log p/(np)})$ *for any* $\boldsymbol{\Omega}_{p\times(p-m)}$ s.t. $\boldsymbol{\Omega}'\boldsymbol{\Omega} = \mathbf{I}_{p-m}$.

The theorems state that under the pervasiveness condition that spiked eigenvalues are of order p, we are able to approximately recover the true leading eigenvalues and eigenvectors. In Fan and Wang (2015), the same phenomenon is observed when $\mathbf{z}_i$'s are sub-Gaussian vector with independent elements. But here we do not require element-wise independence and relax the condition to any sub-Gaussian isotropic random vectors satisfying (3.1). The proofs of the above two theorems can be found in Appendix E in the supplementary material (Fan, Liu and Wang, 2015).

Given the above two theorems, let us validate the second and third conditions in (1.6). Define $\hat{\boldsymbol{\Lambda}}_{\mathrm{SG}} = \mathrm{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_m)$ where SG is short for sub-Gaussian. The second condition holds for $\hat{\boldsymbol{\Lambda}}_{\mathrm{SG}}$ according to Theorems 3.1. Note that $\hat{\boldsymbol{\Sigma}}_Y$ and $\hat{\boldsymbol{\Sigma}}_X$ share the same set of empirical eigenvalues. To check the third one, let $\widehat{\boldsymbol{\Gamma}}_{\mathrm{SG}} = (\widehat{\boldsymbol{\xi}}_1^{(Y)}, \ldots, \widehat{\boldsymbol{\xi}}_m^{(Y)})$ be the matrix consisting of the top m leading eigenvectors of $\hat{\boldsymbol{\Sigma}}_Y$. If the whole eigen-space of $\boldsymbol{\Sigma}$ is written as $\boldsymbol{\Gamma}_p = (\boldsymbol{\Gamma}, \boldsymbol{\Omega})$, then $\widehat{\boldsymbol{\xi}}_j^{(Y)} = \boldsymbol{\Gamma}_p \widehat{\boldsymbol{\xi}}_j = \boldsymbol{\Gamma}\widehat{\boldsymbol{\xi}}_{jA} + \boldsymbol{\Omega}\widehat{\boldsymbol{\xi}}_{jB}$.Therefore $\widehat{\boldsymbol{\xi}}_j^{(Y)} - \boldsymbol{\xi}_j = \boldsymbol{\Gamma}(\widehat{\boldsymbol{\xi}}_{jA} - \mathbf{e}_{jA}) + \boldsymbol{\Omega}\widehat{\boldsymbol{\xi}}_{jB}$ and

$$\|\widehat{\boldsymbol{\Gamma}}_{\mathrm{SG}} - \boldsymbol{\Gamma}\|_{\max} = \max_j \|\widehat{\boldsymbol{\xi}}_j^{(Y)} - \boldsymbol{\xi}_j\|_{\max} \le \max_j \left(\sqrt{m}\|\boldsymbol{\Gamma}\|_{\max}\|\widehat{\boldsymbol{\xi}}_{jA} - \mathbf{e}_{jA}\| + \|\boldsymbol{\Omega}\widehat{\boldsymbol{\xi}}_{jB}\|_{\max}\right), \quad (3.2)$$

which is $O_P(\sqrt{\log p/(np)})$ due to Theorem 3.2 and the fact $\|\boldsymbol{\Gamma}\|_{\max} = O(1/\sqrt{p})$ shown in the proof of Theorem 2.1. Hence, the sample covariance based estimators $\hat{\boldsymbol{\Sigma}}_Y$, $\hat{\boldsymbol{\Lambda}}_{\mathrm{SG}}$ and $\widehat{\boldsymbol{\Gamma}}_{\mathrm{SG}}$ satisfy the sufficient conditions in (1.6). Together with Theorem 2.1, this explains why POET achieves all the desired rates (2.6) – (2.9).

We finally devote a remark to the assumption of zero mean of the observed data implied by Assumption 3.1. This condition is only made to simplify the presentation of proofs. In practice, we first center the data by $\bar{\mathbf{y}} = n^{-1}\sum_i \mathbf{y}_i$. All the conclusions of this section hold for the centered data as well.

4. Elliptical factor models

In the previous section, we assume $\mathbf{y}_i$ to be a sub-Gaussian random vector, a strong distributional assumption for many applications. In this section, we replace the sub-Gaussian assumption 3.1 by elliptical distribution assumption 4.1 and propose a novel robust estimator for the analysis of factor models.

We first briefly review the elliptical distribution family, which generalize the multivariate normal distribution and multivariate t-distribution. Compared to the sub-Gaussian setting, it is more challenging to design pilot estimators to simultaneously satisfy the three requirements in (1.6). To handle this challenge, we separately construct two estimators $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$. $\hat{\Sigma}_1$ and its leading eigenvalues satisfy the first two requirements in (1.6) while the eigenvectors of $\hat{\Sigma}_2$ satisfy the last condition of (1.6).

4.1. Elliptical distribution

We define the elliptical distribution as follows. Let $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ with $\operatorname{rank}(\boldsymbol{\Sigma}) = q \leq p$. A p-dimensional random vector $\mathbf{y}$ has an elliptical distribution, denoted by $\mathbf{y} \sim \mathrm{ED}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \zeta)$, if it has a stochastic representation

$$\mathbf{y} \overset{d}{=} \boldsymbol{\mu} + \zeta \mathbf{A}\mathbf{U}, \quad (4.1)$$

where $\mathbf{U}$ is a random vector uniformly distributed on the unit sphere $\mathcal{S}^{q-1}$ in $\mathbb{R}^q$, $\zeta \geq 0$ is a scalar random variable independent of $\mathbf{U}$, $\mathbf{A} \in \mathbb{R}^{p\times q}$ is a deterministic matrix satisfying $\mathbf{A}\mathbf{A}' = \boldsymbol{\Sigma}$. Here $\boldsymbol{\Sigma}$ is called the scatter matrix. Note that the representation in (4.1) is not identifiable since we can rescale ζ and $\mathbf{A}$. To make the model identifiable, we require $\mathbb{E}\zeta^2 = q$ so that $\operatorname{Cov}(\mathbf{y}) = \boldsymbol{\Sigma}$. In addition, we assume $\boldsymbol{\Sigma}$ is non-singular, i.e., $q = p$, following Assumption 2.1. In this paper, we only consider continuous elliptical distributions with $\mathbb{P}(\zeta = 0) = 0$.

An equivalent definition of an elliptical distribution is through its characteristic function $\exp(i\mathbf{t}'\boldsymbol{\mu})\psi(\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t})$, where ψ is a properly defined characteristic function and $i := \sqrt{-1}$. ζ and ψ are mutually determined by each other. In this setting, we denote by $\mathbf{y} \sim \mathrm{ED}_p(\mu, \boldsymbol{\Sigma}, \psi)$. The marginal and conditional distributions of an elliptical distribution are also elliptical.

Compared to the Gaussian family, the elliptical family provides more flexibility in modeling complex data. The main advantage of the elliptical family is its ability to model heavy-tailed data and the tail dependence between variables (Hult and Lindskog, 2002), which makes it useful for modeling many modern datasets, including financial data (Rachev, 2003; Cizek, Härdle and Weron, 2005), genomics data (Liu et al., 2003; Posekany, Felsen-stein and Sykacek, 2011), and fMRI data (Ruttimann et al., 1998).

The following assumption is considered in this section.

Assumption 4.1 (Elliptical distribution)—The data $\mathbf{y}_i$'s are elliptically distributed, i.e., $\mathbf{y}_i \sim \mathrm{ED}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \zeta)$ or $\mathbf{y}_i \overset{d}{=} \boldsymbol{\mu} + \zeta_i \boldsymbol{\Sigma}^{\frac{1}{2}} \mathbf{U}_i$ with $\mathbf{U}_i$ uniformly distributed on the unit sphere $\mathcal{S}^{p-1}$ and the random variable $\zeta_i \geq 0$ independent from $\mathbf{U}_i$. Additionally, we assume $\mathbb{E}[\zeta_i^2] = p$ due to identifiability and $\max_{j \leq p} \mathbb{E} y_{ij}^4$ is bounded by an absolute constant independent of p.

The above assumption is implied by imposing a joint elliptical model of the factors and noises, i.e., $(\mathbf{f}_t', \mathbf{u}_t')' \sim \mathrm{ED}_{p+m}(\mathbf{0}, \mathrm{diag}(\mathbf{I}_m, \boldsymbol{\Sigma}_u), \zeta)$. Obviously, the elliptical family is more general than the Gaussian family and contains heavy-tailed distributions. One typical example is multivariate t-distribution with degrees of freedom ν. The restriction $\nu > 4$ will be imposed so that it has a bounded fourth moment, only for the sake of estimating marginal variances by methods discussed in Section 4.2. We will discuss other possible alternative methods and the necessity of the bounded fourth moment condition in Section 4.5.

4.2. Robust estimation of variances

Let $\boldsymbol{\Sigma} = \mathbf{DRD}$ where $\mathbf{R}$ is the correlation matrix and $\mathbf{D} = \mathrm{diag}(\sigma_1, \ldots, \sigma_p)$ is the diagonal matrix consisting of standard deviations for each dimension. Our construction of $\hat{\boldsymbol{\Sigma}}_1$ is based on separately estimating $\mathbf{D}$ and $\mathbf{R}$. In this subsection, we first introduce a robust estimator $\hat{\mathbf{D}}$ to estimate $\mathbf{D}$.

Since $\mathbf{y}_i \sim \mathrm{ED}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \zeta)$ may be heavy-tailed, we need a method to robustly estimate $\boldsymbol{\mu}$ to center the data and estimate the covariance matrix. Substantial amount of research has been conducted on this subject in both low dimensional setting (Huber, 1964; Beran, 1978; Bickel, 1982; Zou and Yuan, 2008; Wu and Liu, 2009) and high dimensional setting (Belloni et al., 2011; Fan, Fan and Barut, 2014). In addition, Hampel (1974); Rousseeuw and Croux (1993); Koenker (2005) considered the problem from a quantile perspective. In this section, we introduce two M-estimators proposed by Fan, Li and Wang (2016) and Catoni (2012). The methods allow asymmetric distributions, and thus are also useful for robust estimation of variances.

Let us denote $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$ and $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})'$ for $i = 1, \ldots, n$. We estimate each μ_j using the data $\{y_{1j}, \ldots, y_{nj}\}$. The M-estimator $\hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_p)'$ of Fan, Li and Wang (2016) is obtained by solving

$$\sum_{i=1}^{n} h[\alpha(y_{ij} - \hat{\mu}_j)] = 0 \quad (4.2)$$

for each $j \leq p$, where $h : \mathbb{R} \to \mathbb{R}$ is the derivative function of the Huber loss satisfying $h(x) = x$ if $|x| \leq 1$, $h(x) = 1$ if $x > 1$ and $h(x) = -1$ if $x < -1$. The above estimator can be equivalently obtained by minimizing the Huber loss

$$\ell_\alpha(x) = \begin{cases} 2\alpha^{-1}|x| - \alpha^{-2} & : |x| > \alpha^{-1}; \\ x^2 & : |x| \le \alpha^{-1}. \end{cases}$$

According to Fan, Li and Wang (2016), choosing $\alpha = \sqrt{\log(\varepsilon^{-1})/(nv)}$ for $\varepsilon \in (0, 1)$ such that $\log(\varepsilon^{-1}) \le n/8$ and v an upper bound of $\max\left\{\sigma_1^2, \ldots, \sigma_p^2\right\}$ we have

$$\mathbb{P}\left(\left|\hat{\mu}_j - \mu_j\right| \le 4\sqrt{\frac{v\log(\varepsilon^{-1})}{n}}\right) \ge 1 - 2\varepsilon. \quad (4.3)$$

Catoni (2012) proposed another M-estimator by solving (4.2) with a different strictly increasing influence function $h(x)$ such that $-\log(1-x+x^2/2) \le h(x) \le \log(1+x+x^2/2)$. For a value $\varepsilon \in (0,1)$ such that $n > 2\log(1/\varepsilon)$, let

$$\alpha = \sqrt{\frac{2\log(\varepsilon^{-1})}{n(v + \frac{2v\log(\varepsilon^{-1})}{n - 2\log(\varepsilon^{-1})})}},$$

where v is again an upper bound of $\max\left\{\sigma_1^2, \ldots, \sigma_p^2\right\}$. Catoni (2012) showed that the solution of (4.2) satisfies

$$\mathbb{P}\left(\left|\hat{\mu}_j - \mu_j\right| \le \sqrt{\frac{2v\log(\varepsilon^{-1})}{n - 2\log(\varepsilon^{-1})}}\right) \ge 1 - 2\varepsilon. \quad (4.4)$$

Therefore, by taking $\varepsilon = 1/(n \vee p)^2$, $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_\infty \le C\sqrt{\log p/n}$ with probability at least $1 - 2(n \vee p)^{-1}$ for both methods. We implement Catoni's estimator in the simulation by taking $h(x) = \mathrm{sgn}(x)\log(1 + |x| + x^2/2)$. For the choice of v, we simply take $v = 3\max\left\{\tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_p^2\right\}$ as in Fan et al. (2015), where $\tilde{\sigma}_j^2$ is the sample variance of the j^{th} dimension. Catoni (2012) also introduced the Lepski's method for adaptively choosing v.

To estimate σ_j^2, we apply the above M-estimators on the squared data. Note that $\sigma_j^2 = \mathbb{E}(y_{ij}^2) - \mu_j^2$. We have estimated μ_j above. To estimate $\mathbb{E}(y_{ij}^2)$, we employ the M-estimator (4.2) on the squared data $\left\{y_{1j}^2, \ldots, y_{nj}^2\right\}$, and denote the resulting estimator by $\hat{\eta}_j$. This works as the fourth moment of y_{ij} is finite. The robust variance estimator is then defined as

$$\hat{\sigma}_j^2 = \max\left\{\hat{\eta}_j - \hat{\mu}_j^2, \delta_0\right\}, \quad (4.5)$$

where $\delta_0 > 0$ is a small constant ($\delta_0 < \min\{\sigma_1^2, \ldots, \sigma_p^2\}$).

Let $\widehat{\mathbf{D}} = \mathrm{diag}(\hat{\sigma}_1, \ldots, \hat{\sigma}_p)$, we have the following proposition.

Proposition 4.1—*Suppose* $n \geq C \log p$,

$$\|\widehat{\mathbf{D}} - \mathbf{D}\| = O_P(\sqrt{\log p / n}). \quad (4.6)$$

Additionaly, $\|\mathbf{D}\| = O(1)$ due to the structure $\mathbf{\Sigma} = \mathbf{B}\mathbf{B}' + \mathbf{\Sigma}_u$, *where* $\|\mathbf{\Sigma}_u\|$ and $\|\mathbf{B}\|_{\max}$ are bounded. Hence $\|\widehat{\mathbf{D}}\| = O_P(1)$.

4.3. Marginal Kendall's tau estimator

We now provide a pilot estimator to robustly estimate the correlation matrix $\mathbf{R} = (r_{jk})$ when the data follow an elliptical distribution. The idea of Kendall's tau statistic was introduced by Kendall (1948) for estimating pairwise comovement correlation. Kendall's tau correlation coefficient is defined as

$$\hat{\tau}_{jk} := \frac{2}{n(n-1)} \sum_{i < i'} \mathrm{sgn}((y_{ij} - y_{i'j})(y_{ik} - y_{i'k})), \quad (4.7)$$

whose population counterpart is

$$\tau_{jk} := \mathbb{P}((y_{1j} - y_{2j})(y_{1k} - y_{2k}) > 0) - \mathbb{P}((y_{1j} - y_{2j})(y_{1k} - y_{2k}) < 0). \quad (4.8)$$

Note that the estimator does not depend on the location $\boldsymbol{\mu}$. So without loss of generality, we assume $\boldsymbol{\mu} = \mathbf{0}$. Then $\mathbf{y} \sim \mathrm{ED}_p(\mathbf{0}, \mathbf{\Sigma}, \zeta)$ with independent and identically distributed samples $\mathbf{y}_1, \ldots, \mathbf{y}_n$.

Let $\mathbf{T} = (\tau_{jk})$ and $\hat{\mathbf{T}} = \left(\hat{\tau}_{jk}\right)$. For the elliptical family, it is known that the nonlinear relationship $r_{jk} = \sin\left(\frac{\pi}{2}\tau_{jk}\right)$ holds for the Pearson correlation and Kendall's correlation (Fang, Kotz and Ng, 1990; Han and Liu, 2014). Therefore, a natural estimator for $\mathbf{R}$ is $\hat{\mathbf{R}} = \left(\hat{r}_{jk}\right)$ where

$$\hat{r}_{jk} = \sin\left(\frac{\pi}{2}\hat{\tau}_{jk}\right). \quad (4.9)$$

By Theorem 3.2 of Han and Liu (2013a), with probability larger than $1 - 2\varepsilon - \varepsilon^2$ for any $\varepsilon \in (0, 1)$,

$$\|\hat{\mathbf{R}}-\mathbf{R}\|_2 \le \pi^2\|\mathbf{R}\|_2\left(2\sqrt{\frac{\left(\operatorname{tr}\mathbf{R}/\|\mathbf{R}\|_2+1\right)\log(p/\in)}{3n}}+\frac{\left(\operatorname{tr}\mathbf{R}/\|\mathbf{R}\|_2+1\right)\log(p/\in)}{n}\right).$$

Using the fact $\|\mathbf{D}\|^{-2}\|\Sigma\| \le \|\mathbf{R}\| \le \|\mathbf{D}^{-1}\|^2\|\Sigma\|$, we know $\|\mathbf{R}\| \asymp \|\Sigma\| \asymp p$ since all the eigenvalues of $\mathbf{D}$ are bounded away from infinity and zero. This is true because $\lambda_{\min}(\mathbf{D}^2) = \min_j \sigma_j^2 \ge \min_{\|\xi\|=1} \xi' \sum \xi = \lambda_{\min}(\sum) \ge c_0$, and $\|\mathbf{D}\| = O(1)$ in Proposition 4.1. This implies

$$\|\hat{\mathbf{R}}-\mathbf{R}\|_2 = O_P\left(\sqrt{\frac{p^2\log p}{n}}\right). \quad (4.10)$$

Wegkamp and Zhao (2016) derived the same bound as above, while Mitra and Zhang (2014) got rid of the log p term although their results cannot be directly applied to the case of $\|\mathbf{R}\| \gg \log p$.

Combining the rates in (4.6) and (4.10), we conclude if $\log p \le Cn$,

$$\left|\lambda_j(\hat{\mathbf{D}}\hat{\mathbf{R}}\hat{\mathbf{D}}) - \lambda_j(\mathbf{DRD})\right| \le \|\hat{\mathbf{D}}\hat{\mathbf{R}}\hat{\mathbf{D}} - \mathbf{DRD}\| = O_P(\|(\hat{\mathbf{D}}-\mathbf{D})\mathbf{RD}\| + \|\hat{\mathbf{D}}(\hat{\mathbf{R}}-\mathbf{R})\hat{\mathbf{D}}\| + \|(\hat{\mathbf{D}}-\mathbf{D})\mathbf{R}(\hat{\mathbf{D}}-\mathbf{D})\|)$$
$$= O_P\left(\sqrt{\frac{p^2\log p}{n}}\right).$$

Define $\hat{\sum}_1 = \hat{\mathbf{D}}\hat{\mathbf{R}}\hat{\mathbf{D}}$. The estimator $\hat{\mathbf{\Lambda}}_{\mathrm{ED}} = \operatorname{diag}(\lambda_1(\hat{\sum}_1), \ldots, \lambda_m(\hat{\sum}_1))$, which consists of the first m eigenvalues of $\hat{\sum}_1$, the following proposition holds based on the above discussion.

Proposition 4.2—Under Assumptions 2.1 and 4.1,

$$\|\hat{\sum}_1 - \sum\|_{\max} = O_P(\sqrt{\log p/n}), \|\left(\hat{\mathbf{\Lambda}}_{\mathrm{ED}} - \mathbf{\Lambda}\right)\mathbf{\Lambda}^{-1}\| = O_P(\sqrt{\log p/n}). \quad (4.11)$$

Proposition 4.2 verifies the first and second sufficient conditions in (1.6). We can easily check the first conclusion in (4.11) using concentration bound for U-statistics and Proposition 4.1.

Although $\hat{\sum}_1$ based on the marginal Kendall's tau has good properties for leading eigenvalues, it is difficult to prove the third sufficient condition for eigenvectors in (1.6) due to the complicated nonlinear sin(·) transformation. Luckily, we do not require $\hat{\mathbf{\Gamma}}$ and $\hat{\mathbf{\Lambda}}$ in (1.6) to come from the same covariance estimator. In the next section, we propose another covariance estimator $\hat{\sum}_2$ whose eigenvectors satisfy the third sufficient condition in (1.6).

4.4. Spatial Kendall's tau estimator

To find an estimator $\hat{\mathbf{\Gamma}}_{\mathrm{ED}}$ that satisfies the third condition in (1.6), we resort to the spatial Kendall's tau estimator. We focus our analysis again on the transformed data $\mathbf{x}_i = \mathbf{\Gamma}_p' \mathbf{y}_i$. The population spatial Kendall's tau matrix is defined as

$$\mathbf{K} := \mathbb{E}\left(\frac{(\mathrm{x}_1 - \mathrm{x}_2)(\mathrm{x}_1 - \mathrm{x}_2)'}{\|\mathrm{x}_1 - \mathrm{x}_2\|_2^2} \right). \quad (4.12)$$

The sample version of the spatial Kendall's tau estimator is a second-order U-statsitc:

$$\hat{\mathbf{K}} := \frac{2}{n(n-1)} \sum_{i<i'} k(\mathbf{x}_i, \mathbf{x}_{i'}) \quad (4.13)$$

where

$$k(\mathbf{x}_i, \mathbf{x}_{i'}) = \frac{(\mathbf{x}_i - \mathbf{x}_{i'})(\mathbf{x}_i - \mathbf{x}_{i'})'}{\|\mathbf{x}_i, \mathbf{x}_{i'}\|_2^2}.$$

Several important properties of the above estimator are worth mentioning. First this estimator is location invariant, which allows us to assume $\boldsymbol{\mu} = \mathbf{0}$ without loss of generality, i.e., $\mathbf{x}_i \overset{d}{=} \zeta_i \mathbf{\Lambda}_p^{\frac{1}{2}} \mathbf{U}_i$. Moreover, the eigenvectors of the estimator $\hat{\mathbf{K}}$ is equivariant to orthogonal transformation. So if we define the spatial Kendall's tau estimator based on the observed data $\mathbf{y}_i$ as

$$\hat{\sum}_2 = \frac{2}{n(n-1)} \sum_{i<i'} k(\mathbf{y}_i, \mathbf{y}_{i'}) = \mathbf{\Gamma}_p \hat{\mathbf{K}} \mathbf{\Gamma}_p', \quad (4.14)$$

we have $\hat{\boldsymbol{\xi}}_j^{(Y)} = \mathbf{\Gamma}_p \hat{\boldsymbol{\xi}}_j$, where $\hat{\boldsymbol{\xi}}_j$ and $\hat{\boldsymbol{\xi}}_j^{(Y)}$ are the j^{th} empirical eigenvector of $\hat{\mathbf{K}}$ and $\hat{\sum}_2$, respectively.

The most important feature of the U-statistic estimator in (4.13) is that its kernel $k(\mathbf{y}_i, \mathbf{y}_{i'})$ does not depend on the distribution of ζ. To see this, for elliptically distributed $\mathbf{x}$, we have

$$\mathbf{x} - \tilde{\mathbf{x}} \overset{d}{=} \zeta \mathbf{\Lambda}_p^{\frac{1}{2}} \mathbf{U} - \tilde{\zeta} \mathbf{\Lambda}_p^{\frac{1}{2}} \tilde{\mathbf{U}} \overset{d}{=} \bar{\zeta} \mathbf{\Lambda}_p^{\frac{1}{2}} \mathbf{U},$$

where $\tilde{\mathbf{x}}$ is an independent copy of $\mathbf{x}$ and the characteristic function of $\bar{\zeta}$ is determined by that of ζ. See Hult and Lindskog (2002) for a detailed expression of the characteristic function. Thus, for a multivariate standard normal vector $\mathbf{g} = (g_1, \ldots, g_p)'$,

$$k(\mathbf{x}, \tilde{\mathbf{x}}) = \frac{(\mathbf{x} - \tilde{\mathbf{x}})(\mathbf{x} - \tilde{\mathbf{x}})'}{\|\mathbf{x} - \tilde{\mathbf{x}}\|_2^2} \overset{d}{=} \frac{\boldsymbol{\Lambda}_p^{\frac{1}{2}} \mathbf{U}\mathbf{U}' \boldsymbol{\Lambda}_p^{\frac{1}{2}}}{\mathbf{U}' \boldsymbol{\Lambda}_p \mathbf{U}} \overset{d}{=} \frac{\boldsymbol{\Lambda}_p^{\frac{1}{2}} \mathbf{g}\mathbf{g}' \boldsymbol{\Lambda}_p^{\frac{1}{2}}}{\mathbf{g}' \boldsymbol{\Lambda}_p \mathbf{g}},$$

which depends only on $\mathbf{g}$. The last equality is due to $\mathbf{U} \overset{d}{=} \mathbf{g}/\|\mathbf{g}\|$. Thus $\mathbf{K}$ defined by (4.12) is a diagonal matrix by the symmetry of $\mathbf{g}$.

Write $\mathbf{K} = \mathrm{diag}(\theta_1, \ldots, \theta_p)$, where θ_j is defined as

$$\theta_j = \mathbb{E}\left(\frac{\lambda_j g_j^2}{\sum_{k=1}^p \lambda_k g_k^2} \right),$$

which is a multiple of λ_j. Obviously, $\mathbf{K}$ shares the same eigenvalue ordering as that of $\mathrm{Cov}(\mathbf{x}_i) = \boldsymbol{\Lambda}_p$, and thus the same eigenspaces as those of $\mathrm{Cov}(\mathbf{x}_i)$. So estimating the leading eigenvectors of $\mathrm{Cov}(\mathbf{x}_i)$ is equivalent to estimating those of $\mathbf{K}$. In sum, $\hat{\sum}_2$ particularly fits the goal of estimating the eigenvectors of $\boldsymbol{\Sigma}$.

The above spatial Kendall's tau statistic was first introduced in Choi and Marden (1998) and has been used for low dimensional covariance estimation (Visuri, Koivunen and Oja, 2000) and principal component estimation (Marden, 1999; Croux, Ollila and Oja, 2002). Many testing literature based on rank statistics is also related to the estimator, for example Tyler (1982); Hallin and Paindaveine (2006). The literature listed here is only illustrative rather than complete.

We now consider the theoretical properties of the eigenvectors $\hat{\boldsymbol{\xi}}_j$ of $\hat{\mathbf{K}}$. As before, we choose the proper sign for $\hat{\boldsymbol{\xi}}_j$, which is divided into the spiked part $\hat{\boldsymbol{\xi}}_{jA}$ and non-spiked part $\hat{\boldsymbol{\xi}}_{jB}$.

Theorem 4.1—Under Assumptions 2.1 and 4.1, for j ≤ m we have

- **i.** $\|\hat{\boldsymbol{\xi}}_{jA} - \mathbf{e}_{jA}\| = O_P(n^{-1/2})$, where $\mathbf{e}_{\mathrm{jA}}$ is a unit vector of length m with 1 at the j^{th} coordinate and 0 everywhere else;
- **ii.** $\|\boldsymbol{\Omega}\hat{\boldsymbol{\xi}}_{jB}\|_{\max} = O_P(\sqrt{\log p/(np)})$ *for any* $\boldsymbol{\Omega}_{p\times(p-m)}$ *s.t.* $\boldsymbol{\Omega}'\boldsymbol{\Omega} = \mathbf{I}_{p-m}$.

The proof for Theorem 4.1 is relegated to Appendix F in the supplementary material (Fan, Liu and Wang, 2015). Define $\hat{\sum}_2$ as the spatial Kendall's tau estimator of the observed data $\mathbf{y}_i$'s and $\hat{\boldsymbol{\Gamma}}_{\mathrm{ED}} = (\hat{\boldsymbol{\xi}}_1^{(Y)}, \ldots, \hat{\boldsymbol{\xi}}_m^{(Y)})$ as the leading eigenvectors of $\hat{\sum}_2$. Theorem 4.1 implies

$$\|\hat{\mathbf{\Gamma}}_{\mathrm{ED}} - \mathbf{\Gamma}\|_{\max} = O_P(\sqrt{\log p/(np)}),$$

following the same derivation in (3.2). So the third sufficient condition in (1.6) holds for $\hat{\mathbf{\Gamma}}_{\mathrm{ED}}$. Together with the estimators $\hat{\sum}_1$ and $\hat{\mathbf{\Lambda}}_{\mathrm{ED}}$ defined in Section 4.3, we are ready to apply the generic POET procedure for the elliptical factor model and achieve all the desired estimation convergence rates for both covariance and precision matrices.

We comment on the computation time for the robust procedure. Compared with the computation complexity $O(p^2n)$ of the sample covariance matrix, $\hat{\mathbf{T}}$ and $\hat{\mathbf{K}}$ can be calculated by an efficient algorithm based on sorting and balanced binary trees, which achieves the complexity $O(p^2n \log n)$. Please see Christensen (2005) for more details.

4.5. Other alternatives

Under the elliptical distributions, in addition to the above marginal and spatial Kendall's tau estimators, there are other possible methods. In this section, we discuss some alternative options.

For mean estimation, since elliptical distribution is symmetric, sample median should also work well. For marginal variance estimation, Hsu and Sabato (2014) proposed the "median-of-means" estimator, which has an optimal concentration bound similar to (4.3) and (4.4) when the fourth moment is finite, but has less efficiency compared with the M-estimators (Fan, Wang and Zhong, 2016); in addition, some quantile-based methods such as the mean absolute deviation (MAD) and Q_n (Hampel, 1974; Rousseeuw and Croux, 1993) may be used to robustly estimate the variance if the underlying distribution is known. For covariance matrix estimation, Han, Lu and Liu (2014) proposed the generalized MAD and Q_n estimators to estimate the scatter matrix, which is proportional to the covariance matrix, but we still need a proper way to estimate the one-dimensional scaling factor. Visuri, Koivunen and Oja (2000) mentioned other two-step robust procedures to estimate eigenvalues and eigenvectors. For example, one can first estimate the eigenvectors using $\hat{\sum}_2$ and then project the data onto the estimated eigenvectors to further estimate the eigenvalues (variances of the projected data); or one can first estimate the (marginal or spatial) median $\hat{\boldsymbol{\mu}}$ and then replace $\mathbf{y}_{i'}$ with $\hat{\boldsymbol{\mu}}$ in (4.7) and (4:14) and average over only sample index i (Dürre, Vogel and Tyler, 2014). Those methods are potentially applicable, although their analysis can be involved, especially in high dimensions.

Another question is about the necessity of the finite fourth moment in Assumption 4.1, which serves only for the estimation of marginal variances. This condition seemingly cannot be removed in all the alternative methods discussed above. In (4.3) and (4.4), the deviation bound depends on v, which has to be assumed bounded in order to achieve the desired rate $\|\hat{\mathbf{D}} - \mathbf{D}\| = O_P(\sqrt{\log p/n})$. For the MAD type of estimators, we still need to estimate the auxiliary scalar with a desired rate of convergence, say $O_P(n^{-1/2})$. It is not clear to us whether this can be done without the bounded fourth moment condition. Certainly, if the

factor analysis is applied to standardized variables, for example as in Wegkamp and Zhao (2016), the marginal variance estimation can be avoided. However, one may argue in some applications, scales indeed matter to explain variability.

5. Simulations

Simulations are carried out in this section to demonstrate the effectiveness of the proposed method for elliptical factor models. The robust estimators $\hat{\Sigma}_1$, $\hat{\mathbf{\Lambda}}_{\mathrm{ED}}$, $\hat{\mathbf{\Gamma}}_{\mathrm{ED}}$ proposed in Section 4 will be compared with the original POET estimator based on the sample covariance, or $\hat{\Sigma}_Y$, $\hat{\mathbf{\Lambda}}_{\mathrm{SG}}$, $\hat{\mathbf{\Gamma}}_{\mathrm{SG}}$ discussed in Section 3. We put the two sets of estimators into the generic POET framework described in Section 2 for estimating both conditional sparsity covariance and conditional graphical models. In addition, we also compare $\hat{\Sigma}^{\mathcal{T}}$ and $\hat{\Sigma}_u^{\mathcal{T}}$ in estimating $\mathbf{\Sigma}$ in relative Frobenius norm $\|\cdot\|_\Sigma$ discussed in Section 2.3.

5.1. Conditional sparse covariance estimation

We consider the factor model (1.1) with $(\mathbf{f}_t, \mathbf{u}_t)$ jointly follow a multivariate t-distribution with degrees of freedom ν. Larger ν corresponds to a lighter tail and $\nu = \infty$ corresponds to a multivariate normal distribution. We simulated n independent samples of $(\mathbf{f}_t, \mathbf{u}_t)$ from multivariate t-distribution with covariance matrix $\mathrm{diag}(\mathbf{I}_m, \mathbf{I}_p)$ and each row of $\mathbf{B}$ from $\mathcal{N}(\mathbf{0}, \mathbf{I}_m)$. The observed data is formed as $\mathbf{y}_t = \mathbf{B}\mathbf{f}_t + \mathbf{u}_t$ and the true covariance is $\mathbf{\Sigma} = \mathbf{B}\mathbf{B}' + \mathbf{I}_p$. We vary p from 100 to 1000 with sample size $n = p/2$, and fixed number of factors $m = 3$.

For each triple (p, n, m), both the original POET estimator ($\hat{\Sigma}_Y$, $\hat{\mathbf{\Lambda}}_{\mathrm{SG}}$, $\hat{\mathbf{\Gamma}}_{\mathrm{SG}}$) and the proposed robust POET estimator ($\hat{\Sigma}_1$, $\hat{\mathbf{\Lambda}}_{\mathrm{ED}}$, $\hat{\mathbf{\Gamma}}_{\mathrm{ED}}$) were employed to estimate $\mathbf{\Sigma}_u$ and $\mathbf{\Sigma}$. 100 simulations were conducted for each case. The log-ratios (base 2) of the average estimation errors using the two methods were reported in Figure 1, measured under the following norms:

- For Σ_u: $\|\hat{\Sigma}_u^{\mathcal{T}} - \Sigma_u\|_{\max}$, $\|\hat{\Sigma}_u^{\mathcal{T}} - \Sigma_u\|_2$, $\|(\hat{\Sigma}_u^{\mathcal{T}})^{-1} - \Sigma_u^{-1}\|_2$;
- For Σ: $\|\hat{\Sigma}^{\mathcal{T}} - \Sigma\|_{\max}$, $\|\hat{\Sigma}^{\mathcal{T}} - \Sigma\|_\Sigma$, $\|\hat{\Sigma}_u^{\mathcal{T}} - \Sigma\|_\Sigma$, $\|(\hat{\Sigma}^{\mathcal{T}})^{-1} - \Sigma^{-1}\|_2$;
- For $\mathbf{B}\mathbf{B}'$: $\|\hat{\mathbf{\Gamma}}\hat{\mathbf{\Lambda}}\hat{\mathbf{\Gamma}}' - \mathbf{B}\mathbf{B}'\|_2$, $\|\mathbf{L}^{-1/2}(\hat{\mathbf{\Gamma}}\hat{\mathbf{\Lambda}}\hat{\mathbf{\Gamma}}' - \mathbf{B}\mathbf{B}')\mathbf{L}^{-1/2}\|_F$;
- For initializers: $\|\hat{\Sigma} - \Sigma\|_{\max}$, $\|(\hat{\mathbf{\Lambda}} - \mathbf{\Lambda})\mathbf{\Lambda}^{-1}\|_2$, $\|\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}\|_{\max}$;

In addition, three different degrees of freedom $\nu = 4.2$, $\nu = 7$, $\nu = \infty$ were chosen, representing respectively heavy tail, moderate tail, and normal situations.

From Figure 1, when factors and noises are heavy-tailed from $t_{4.2}$ (black dotted), the original POET estimator is poorly behaved while the robust method performs well as we expected. t_7 (blue dashed) typically fits financial or biological data better than normal in practice. In this case, we also observe a significant advantage of the robust POET estimator. The errors are

reduced by roughly a half if the rank based estimation is applied. However, when the distribution is indeed normal or t_∞ (orange solid), the original POET estimator based on sub-Gaussian data performs better, though the robust method also achieves comparable performance.

In Figure 2, unlike the log ratios plotted in Figure 1, we directly plot the errors of the optimal method (original vs robust POET) in the setting of t-distribution with $\nu = 4.2$ (dotted) and $\nu = \infty$ (solid). We specifically compare $\|\hat{\Sigma}^{\mathcal{T}} - \Sigma\|_\Sigma$ (red) and $\|\hat{\Sigma}_u^{\mathcal{T}} - \Sigma\|_\Sigma$ (blue) in the left panel and compare $\|\mathbf{L}^{-1/2}\left(\hat{\mathbf{\Gamma}}\hat{\mathbf{\Lambda}}\hat{\mathbf{\Gamma}}' - \mathbf{B}\mathbf{B}'\right)\mathbf{L}^{-1/2}\|_F$ (red) and $\|\mathbf{L}^{-1/2}(0 - \mathbf{B}\mathbf{B}')\mathbf{L}^{-1/2}\|_F$ (blue) in the right panel (note that blue dotted and blue solid lines coincide). As discussed in Section 2.3, the naive estimator $\hat{\Sigma}_u^{\mathcal{T}}$ is optimal in the relative Frobenius norm and indeed superior to the POET estimator. However, POET provides good recovery for the low-rank part while $\hat{\Sigma}_u^{\mathcal{T}}$ only estimates the sparse part.

5.2. Conditional graphical model estimation

We consider the conditional graphical model described in Section 2.4. In particular, we compare the accuracy of different methods for estimating the precision matrices $\mathbf{\Omega}_u$ and $\mathbf{\Omega}$. Here we assume a block diagonal precision error matrix $\mathbf{\Omega}_u = \text{diag}(\mathbf{M}, \ldots, \mathbf{M})$ where $\mathbf{M}$ is a 2 by 2 correlation matrix with off-diagonal element equals 0.5. Then we simulate $(\mathbf{f}_t, \mathbf{u}_t)$ again from multivariate t-distribution with covariance diag $(\mathbf{I}_m, \mathbf{\Omega}_u^{-1})$. We set the dimension p to range from 50 to 500, sample size $n = 0.6p$ and a fixed number of factors $m = 3$.

For each configuration of (p, n, m), after applying POET with the original and robust pilot estimators, we estimate $\hat{\mathbf{\Omega}}_u$ and $\hat{\mathbf{\Omega}}$ as proposed in Section 2.4 using the CLIME procedure. To efficiently solve large-scale CLIME optimization (2.11), we used the R package "fastclime" developed by Pang, Liu and Vanderbei (2014), which provides a very efficient C implementation of a parametric dual simplex algorithm (Vanderbei, 2007). 100 simulations were conducted for each case. The log-ratios (base 2) of the average errors of the two methods were reported in Figure 3, measured under spectral norms $\|\hat{\mathbf{\Omega}}_u - \mathbf{\Omega}_u\|_2$ and $\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\|_2$. Three different degrees of freedom $\nu = 4.2$ (black dotted), $\nu = 7$ (blue dashed), $\nu = \infty$ (orange solid) were used as in Section 5.1. Clearly, the robust estimators outperform non-robust ones for $t_{4.2}$ and t_7, and maintains competitive for the normal case.

6. Discussions

We provide a fundamental understanding of high dimensional factor models under the pervasiveness condition. In particular, we extend the POET estimator in Fan, Liao and Mincheva (2013) to a generic procedure which could take any pilot covariance matrix estimators as initial inputs, as long as they satisfy a set of sufficient, high level conditions specified in (1.6). Transparent theoretical results are then developed. The main challenge is to check those high level conditions for given estimators. When the observed data $\mathbf{y}_i$ is sub-

Gaussian, we are able to use sample covariance matrix to construct initial estimators. However, if we encounter heavy-tailed elliptical distributions, robust estimators for the eigen-structure should be considered. The paper provides an example of separately estimating leading eigenvalues and eigenvectors under the elliptical factor models. The results may be generalized to richer families of distributions.

It is interesting to see whether it is possible to eliminate the pervasiveness condition. Based on the recent work of Fan and Wang (2015), it is possible to relax the spiked eigenvalue condition from order p to weaker signal level of order $\sqrt{p}$, by correcting the estimation biases of the empirical eigenvalues, under sub-Gaussian factor models. However, as pointed out by Johnstone and Lu (2009), bounded eigenvalues are insufficient for consistent estimation of the eigen-structure when $p \geq n$. It is an interesting topic of future investigation to understand the weakest condition on the divergence rate of the spiked eigenvalues. It is worth mentioning that the requirement on the pervasiveness condition also depends on the targeted measure of estimation error. For example, the pervasiveness condition may be eliminated if we only care about optimally estimating Σ in $\|\cdot\|_{\Sigma}$, which we perceive as an improper criterion for the POET procedure. However, in other measure of estimation errors (e.g., $\|\cdot\|_F, \|\cdot\|_2$, etc) and heavy-tailed distributions, diverging eigenvalues (if not at the diverging rate p) might still be necessary for separately recovering the low-rank and sparse components.

Agarwal, Negahban and Wainwright (2012) consider a similar type of low-rank plus sparse decomposition, but their work is based on solving the convex optimization of Frobenius loss with nuclear and sparse regularization. We remark on the comparison of the optimization approach with our generic POET approach. The POET procedure can be viewed as a one-step approximation to the optimization problem. Therefore, the optimization approach may potentially reduce the required signal level as it involves multiple iterations. But it is not clear how much the signal can be reduced for optimal recovery of both the low-rank and sparse matrices in various norms. Specifically, Agarwal, Negahban and Wainwright (2012) do not leverage pervasiveness, but assume the element-wise maximum of the low-rank component to be of order $O(1/p)$, which is not as natural as pervasive factors. In addition, they only derived the Frobenius error bounds. Chandrasekaran, Parrilo and Willsky (2012) studied an optimization based estimator for the latent graphical model, but under a typical setting ($\mu(\mathbf{\Omega}) \asymp 1$, $\xi(T) \asymp 1/\sqrt{p}$ following their notations), they require the minimal eigenvalue of the low-rank matrix to be larger than the order $p/\sqrt{n}$ and only consider the situation of $p \leq Cn$. So with a relatively smaller dimensionality, they still have a condition on the diverging rate of leading eigenvalues, almost as strong as our pervasiveness condition. Wegkamp and Zhao (2016) applied the optimization approach to the strict factor model for $p \leq n$. Their requirement on the signal level is of a similar order $p\sqrt{\log p/n}$. Hsu, Kakade and Zhang (2011) put no restriction on the signal strength, however their conclusions use very different norms from ours and are not easy to be compared with. By all means, the comparison between the optimization approach and the generic POET approach should be studied in further details. We leave this for future investigation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to acknowledge Junwei Lu and Ziwei Zhu for their helpful discussions, and the associate editor and the three anonymous reviewers for their insightful comments, which lead to a better presentation of this paper.

APPENDIX A: PROOFS IN SECTION 2

We first state a version of the Davis and Kahan's sin θ theorem as follows, which will be used now and again in the proofs.

Proposition A.1 (Davis and Kahan (1970))

Using the notations of our paper, we have

$$\|\hat{\xi}_j - \xi_j\|_2 \le \sqrt{2}\|\hat{\xi}_j\hat{\xi}_j' - \xi_j\xi_j'\|_2 \le \frac{\sqrt{2}\|\hat{\Sigma} - \Sigma\|}{\min\left\{\left|\hat{\lambda}_{j-1} - \lambda_j\right|, \left|\lambda_j - \hat{\lambda}_{j+1}\right|\right\}}$$

with the convention of choosing the right sign for eigenvectors and $\hat{\lambda}_0 = \infty$.

Proof of Theorem 2.1

We establish (2.6) here and put the remaining proofs for (2.7) – (2.9) in Appendix B in the supplementary material (Fan, Liu and Wang, 2015).

To obtain the rates of convergence in (2.6), it suffices to prove $\|\hat{\Sigma}_u - \Sigma_u\|_{\max} = Op(w_n)$. Once the max error of sparse matrix $\mathbf{\Sigma}_u$ is controlled, it is not hard to show the adaptive procedure discussed in (2.5) gives $\hat{\Sigma}_u^{\mathcal{T}}$ such that the spectral error $\|\hat{\Sigma}_u^{\mathcal{T}} - \Sigma_u\|_2 = O_p(m_p w_n^{1-q})$ (Fan, Liao and Mincheva, 2011; Cai and Liu, 2011; Rothman, Levina and Zhu, 2009). Furthermore, $\|(\hat{\Sigma}_u^{\mathcal{T}})^{-1} - \Sigma_u^{-1}\|_2 \le \|(\hat{\Sigma}_u^{\mathcal{T}})^{-1}\|_2 \|\hat{\Sigma}_u^{\mathcal{T}} - \Sigma_u\|_2 \|\Sigma_u^{-1}\|_2$. So $\|(\hat{\Sigma}_u^{\mathcal{T}})^{-1} - \Sigma_u^{-1}\|_2$ is also $O_P\left(m_p w_n^{1-q}\right)$ due to the lower boundedness of $\|\Sigma_u\|_2$.

According to the first condition in (1.6), $\|\hat{\Sigma} - \Sigma\|_{\max} = O_P(\sqrt{\log p/n})$. Therefore to show $\|\hat{\Sigma}_u - \Sigma_u\|_{\max} = O_P(w_n)$, we only need to prove the low-rank part of Σ concentrates at a desired rate under max norm, that is,

$$\|\hat{\mathbf{\Gamma}}\hat{\mathbf{\Lambda}}\hat{\mathbf{\Gamma}} - \mathbf{B}\mathbf{B}'\|_{\max} = O_P(\sqrt{\log p/n} + 1/\sqrt{p}) \quad \text{(A.1)}$$

Let $\mathbf{B} = \left(\widetilde{\mathbf{b}}_1, \ldots, \widetilde{\mathbf{b}}_m\right)$ and $\mathbf{BB}' = \widetilde{\mathbf{\Gamma}}\widetilde{\mathbf{\Lambda}}\widetilde{\mathbf{\Gamma}}'$ where $\widetilde{\mathbf{\Lambda}} = \mathrm{diag}\,(\|\widetilde{\mathbf{b}}_1\|^2, \ldots, \|\widetilde{\mathbf{b}}_m\|^2)$ and the j^{th} column of $\widetilde{\mathbf{\Gamma}}$ is $\widetilde{\mathbf{b}}_j/\|\widetilde{\mathbf{b}}_j\|$. To obtain (A.1), we bound $\Delta_1 := \|\hat{\mathbf{\Gamma}}\hat{\mathbf{\Lambda}}\hat{\mathbf{\Gamma}}' - \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}'\|_{\max}$ and $\Delta_2 := \|\widetilde{\mathbf{\Gamma}}\widetilde{\mathbf{\Lambda}}\widetilde{\mathbf{\Gamma}}' - \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}'\|_{\max}$ separately. Four useful rates of convergence are listed in the following:

$$\|\widetilde{\mathbf{\Lambda}} - \mathbf{\Lambda}\|_{\max} \le \|\sum_u\| = O(1), \|\widetilde{\mathbf{\Gamma}} - \mathbf{\Gamma}\|_{\max} \le C\|\sum_u\|/p = O(1/p), \|(\hat{\mathbf{\Lambda}} - \mathbf{\Lambda})\mathbf{\Lambda}^{-1}\|_{\max} = O_P(\sqrt{\log p/n}),$$
$$\|\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}\|_{\max} = O_P(\sqrt{\log p/(np)}).$$

The first one is due to Weyl's inequality since $\|\widetilde{\mathbf{\Lambda}} - \mathbf{\Lambda}\|_{\max} = \|\widetilde{\mathbf{\Lambda}} - \mathbf{\Lambda}\|_2$ while the second follows from trivial bound $\|\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}\|_{\max} \le \|\mathbf{\Gamma} - \mathbf{\Gamma}\|_F$, which is further bounded by $C\|\mathbf{\Sigma}_u\|/p$ according to the sin θ theorem of Davis and Kahan (1970) (see Proposition A.1). The third and fourth rates are by assumption. Next we show $\|\mathbf{\Gamma}\|_{\max} = O(1/\sqrt{p})$ and derive the rates for Δ_1 and Δ_2.

Note that

$$\|\mathbf{\Gamma}\mathbf{\Lambda}^{\frac{1}{2}} - \widetilde{\mathbf{\Gamma}}\widetilde{\mathbf{\Lambda}}^{\frac{1}{2}}\|_{\max} \le \|\mathbf{B}\widetilde{\mathbf{\Lambda}}^{-\frac{1}{2}}(\mathbf{\Lambda}^{\frac{1}{2}} - \widetilde{\mathbf{\Lambda}}^{\frac{1}{2}})\|_{\max} + \|(\mathbf{\Gamma} - \widetilde{\mathbf{\Gamma}})\mathbf{\Lambda}^{\frac{1}{2}}\|_{\max} \le C\frac{\|\mathbf{B}\|_{\max} + \|\sum_u\|}{\sqrt{p}} = o(1).$$

Since $\|\mathbf{B}\|_{\max} = \|\widetilde{\mathbf{\Gamma}}\widetilde{\mathbf{\Lambda}}^{\frac{1}{2}}\|_{\max} = O(1)$, we have $\|\mathbf{\Gamma}\mathbf{\Lambda}^{1/2}\|_{\max} = O(1)$ and $\|\mathbf{\Gamma}\|_{\max} = O(1/\sqrt{p})$. Using this fact, the following argument implies $\Delta_1 = O_P(\sqrt{\log p/n})$ and $\Delta_2 = O(\sqrt{1/p})$. More specifically,

$$\Delta_1 \le \|\hat{\mathbf{\Gamma}}(\hat{\mathbf{\Lambda}} - \mathbf{\Lambda})\hat{\mathbf{\Gamma}}'\|_{\max} + \|(\hat{\mathbf{\Gamma}} - \mathbf{\Gamma})\mathbf{\Lambda}(\hat{\mathbf{\Gamma}} - \mathbf{\Gamma})'\|_{\max} + 2\|\mathbf{\Gamma}\mathbf{\Lambda}(\hat{\mathbf{\Gamma}} - \mathbf{\Gamma})'\|_{\max} = O_p(p^{-1}\|\hat{\mathbf{\Lambda}} - \mathbf{\Lambda}\|_{\max}$$
$$+ \sqrt{p}\|\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}\|_{\max}) = O_P(\sqrt{\log p/n}),$$

$$\Delta_2 \le \|\widetilde{\mathbf{\Gamma}}(\widetilde{\mathbf{\Lambda}} - \mathbf{\Lambda})\widetilde{\mathbf{\Gamma}}'\|_{\max} + \|(\widetilde{\mathbf{\Gamma}} - \mathbf{\Gamma})\mathbf{\Lambda}(\widetilde{\mathbf{\Gamma}} - \mathbf{\Gamma})'\|_{\max} + 2\|\mathbf{\Gamma}\mathbf{\Lambda}(\widetilde{\mathbf{\Gamma}} - \mathbf{\Gamma})'\|_{\max} = O(p^{-1}\|\widetilde{\mathbf{\Lambda}} - \mathbf{\Lambda}\|_{\max}$$
$$+ \sqrt{p}\|\widetilde{\mathbf{\Gamma}} - \mathbf{\Gamma}\|_{\max}) = O(\sqrt{1/p}).$$

Combining the rates of Δ_1 and Δ_2, we prove (A.1). Thus (2.6) follows. □

References

Agarwal A, Negahban S, Wainwright MJ. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. The Annals of Statistics. 2012; 40:1171–1197.

Amini AA, Wainwright MJ. Information Theory, 2008 ISIT 2008 IEEE International Symposium on 2454–2458. IEEE; 2008. High-dimensional analysis of semidefinite relaxations for sparse principal components.

Antoniadis A, Fan J. Regularization of wavelet approximations. Journal of the American Statistical Association. 2001; 96

Bai J, Li K. Statistical analysis of factor models of high dimension. The Annals of Statistics. 2012; 40:436–465.

Bai J, Ng S. Determining the number of factors in approximate factor models. Econometrica. 2002; 70:191–221.

Bai J, Ng S. Principal components estimation and identification of static factors. Journal of Econometrics. 2013; 176:18–29.

Belloni A, Chernozhukov V, et al. ℓ_1-penalized quantile regression in high-dimensional sparse models. The Annals of Statistics. 2011; 39:82–130.

Beran R. An efficient and robust adaptive estimator of location. The Annals of Statistics. 1978:292–313.

Berthet Q, Rigollet P. Optimal detection of sparse principal components in high dimension. The Annals of Statistics. 2013a; 41:1780–1815.

Berthet Q, Rigollet P. Complexity theoretic lower bounds for sparse principal component detection. Conference on Learning Theory. 2013b:1046–1066.

Bickel PJ. On adaptive estimation. The Annals of Statistics. 1982:647–671.

Bickel PJ, Levina E. Covariance regularization by thresholding. The Annals of Statistics. 2008a:2577–2604.

Bickel PJ, Levina E. Regularized estimation of large covariance matrices. The Annals of Statistics. 2008b:199–227.

Birnbaum A, Johnstone IM, Nadler B, Paul D. Minimax bounds for sparse PCA with noisy high-dimensional data. Annals of statistics. 2013; 41:1055. [PubMed: 25324581]

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

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

Cai TT, Ma Z, Wu Y. Sparse PCA: Optimal rates and adaptive estimation. The Annals of Statistics. 2013; 41:3074–3110.

Cai T, Ma Z, Wu Y. Optimal estimation and rank detection for sparse spiked covariance matrices. Probability theory and related fields. 2015; 161:781–815. [PubMed: 26257453]

Cai TT, Ren Z, Zhou HH. Optimal rates of convergence for estimating Toeplitz covariance matrices. Probability Theory and Related Fields. 2013; 156:101–143.

Cai TT, Zhang CH, Zhou HH. Optimal rates of convergence for covariance matrix estimation. The Annals of Statistics. 2010; 38:2118–2144.

Cai TT, Zhou HH. Optimal rates of convergence for sparse covariance matrix estimation. The Annals of Statistics. 2012; 40:2389–2420.

Cai TT, Li H, Liu W, Xie J. Covariate-adjusted precision matrix estimation with an application in genetical genomics. Biometrika. 2013; 100:139–156. [PubMed: 28316337]

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

Chamberlain G, Rothschild M. Arbitrage, factor structure, and mean-variance analysis on large asset markets. 1982

Chandrasekaran V, Parrilo PA, Willsky AS. Latent variable graphical model selection via convex optimization. The Annals of Statistics. 2012; 40:1935–1967.

Chandrasekaran V, Sanghavi S, Parrilo PA, Willsky AS. Rank-sparsity incoherence for matrix decomposition. SIAM Journal on Optimization. 2011; 21:572–596.

Choi K, Marden J. A multivariate version of Kendall's τ. Journal of Nonparametric Statistics. 1998; 9:261–293.

Christensen D. Fast algorithms for the calculation of Kendall?s τ. Computational Statistics. 2005; 20:51–62.

Cizek P, Härdle WK, Weron R. Statistical tools for finance and insurance. Springer Science & Business Media; 2005.

Croux C, Ollila E, Oja H. Statistical data analysis based on the L1-norm and related methods. Springer; 2002. Sign and rank covariance matrices: statistical properties and application to principal components analysis; 257–269.

Davis C, Kahan WM. The rotation of eigenvectors by a perturbation. III. SIAM Journal on Numerical Analysis. 1970; 7:1–46.

Dürre A, Vogel D, Tyler DE. The spatial sign covariance matrix with unknown location. Journal of Multivariate Analysis. 2014; 130:107–117.

Fan J, Fan Y, Lv J. High dimensional covariance matrix estimation using a factor model. Journal of Econometrics. 2008; 147:186–197.

Fan J, Fan Y, Barut E. Adaptive robust variable selection. Annals of statistics. 2014; 42:324. [PubMed: 25580039]

Fan J, Feng Y, Wu Y. Network exploration via the adaptive LASSO and SCAD penalties. Annals of Applied statistics. 2009; 3:521–541. [PubMed: 21643444]

Fan J, Li Q, Wang Y. Estimation of high-dimensional mean regression in absence of symmetry and light-tail assumptions. Journal of the Royal Statistical Society: Series B to appear. 2016

Fan J, Liao Y, Mincheva M. High dimensional covariance matrix estimation in approximate factor models. Annals of statistics. 2011; 39:3320. [PubMed: 22661790]

Fan J, Liao Y, Mincheva M. Large covariance estimation by thresholding principal orthogonal complements. Journal of the Royal Statistical Society: Series B. 2013; 75:1–44.

Fan J, Liao Y, Wang W. Projected Principal Component Analysis in Factor Models. arXiv preprint arXiv:1406.3836. 2014

Fan J, Liu H, Wang W. Supplementary appendix to the paper “Large Covariance Estimation through Elliptical Factor Models”. 2015

Fan J, Wang W. Asymptotics of Empirical Eigen-structure for Ultra-high Dimensional Spiked Covariance Model. arXiv preprint arXiv:1502.04733. 2015

Fan J, Wang W, Zhong Y. An ℓ_∞ eigenvector perturbation bound and its application to robust covariance estimation. arXiv preprint arXiv:1603.03516. 2016

Fan J, Ke T, Liu H, Xia L. QUADRO: A Supervised Dimension Reduction Method via Rayleigh Quotient Optimization. Annals of Statistics 43 to apepar. 2015

Fan J, Liu H, Wang W, Zhu Z. Heterogeneity Adjustment with Applications to Graphical Model Inference. arXiv preprint arXiv:1602.05455. 2016

Fang KT, Kotz S, Ng KW. Symmetric multivariate and related distributions. Chapman and Hall; 1990.

Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008; 9:432–441. [PubMed: 18079126]

Hallin M, Paindaveine D. Semiparametrically efficient rank-based inference for shape. I. Optimal rank-based tests for sphericity. The Annals of Statistics. 2006; 34:2707–2756.

Hampel FR. The influence curve and its role in robust estimation. Journal of the American Statistical Association. 1974; 69:383–393.

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

Han F, Liu H. ECA: High Dimensional Elliptical Component Analysis in non-Gaussian Distributions. arXiv preprint arXiv:1310.3561. 2013b

Han F, Liu H. Scale-invariant sparse PCA on high-dimensional meta-elliptical data. Journal of the American Statistical Association. 2014; 109:275–287. [PubMed: 24932056]

Han F, Lu J, Liu H. Robust scatter matrix estimation for high dimensional distributions with heavy tails Technical Report. 2014

Hsu D, Kakade SM, Zhang T. Robust matrix decomposition with sparse corruptions. Information Theory, IEEE Transactions on. 2011; 57:7221–7234.

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

Huber PJ. Robust estimation of a location parameter. The Annals of Mathematical Statistics. 1964; 35:73–101.

Hult H, Lindskog F. Multivariate extremes, aggregation and dependence in elliptical distributions. Advances in Applied probability. 2002; 34:587–608.

Johnstone IM, Lu AY. On Consistency and Sparsity for Principal Components Analysis in High Dimensions. Journal of the American Statistical Association. 2009; 104:682–693. [PubMed: 20617121]

Karoui NE. Operator norm consistent estimation of large-dimensional sparse covariance matrices. The Annals of Statistics. 2008:2717–2756.

Kendall MG. Rank correlation methods. 1948

Koenker R. Quantile regression. Vol. 38. Cambridge university press; 2005.

Lam C, Fan J. Sparsistency and rates of convergence in large covariance matrix estimation. Annals of statistics. 2009; 37:4254. [PubMed: 21132082]

Levina E, Vershynin R. Partial estimation of covariance matrices. Probability Theory and Related Fields. 2012; 153:405–419.

Liu H, Han F, Zhang C-h. Transelliptical graphical models. Advances in Neural Information Processing Systems. 2012:809–817.

Liu L, Hawkins DM, Ghosh S, Young SS. Robust singular value decomposition analysis of microarray data. Proceedings of the National Academy of Sciences. 2003; 100:13167–13172.

Lozeron ED, Victoria-Feser MP. Robust estimation of constrained covariance matrices for confirmatory factor analysis. Computational Statistics & Data Analysis. 2010; 54:3020–3032.

Ma Z. Sparse principal component analysis and iterative thresholding. The Annals of Statistics. 2013; 41:772–801.

Marden JI. Some robust estimates of principal components. Statistics & Probability Letters. 1999; 43:349–359.

Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the lasso. The Annals of Statistics. 2006:1436–1462.

Mitra R, Zhang CH. Multivariate Analysis of Nonparametric Estimates of Large Correlation Matrices. arXiv preprint arXiv:1403.6195. 2014

Pang H, Liu H, Vanderbei R. The fastclime package for linear programming and large-scale precision matrix estimation in R. The Journal of Machine Learning Research. 2014; 15:489–493. [PubMed: 25620890]

Paul D. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. Statistica Sinica. 2007; 17:1617–1642.

Paul D, Johnstone IM. Augmented sparse principal component analysis for high dimensional data. arXiv preprint arXiv:1202.1242. 2012

Pison G, Rousseeuw PJ, Filzmoser P, Croux C. Robust factor analysis. Journal of Multivariate Analysis. 2003; 84:145–172.

Posekany A, Felsenstein K, Sykacek P. Biological assessment of robust noise models in microarray data analysis. Bioinformatics. 2011; 27:807–814. [PubMed: 21252077]

Rachev ST. Handbook of Heavy Tailed Distributions in Finance: Handbooks in Finance. Vol. 1. Elsevier; 2003.

Ravikumar P, Wainwright MJ, Raskutti G, Yu B. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. Electronic Journal of Statistics. 2011; 5:935–980.

Rothman AJ, Levina E, Zhu J. Generalized thresholding of large covariance matrices. Journal of the American Statistical Association. 2009; 104:177–186.

Rothman AJ, Levina E, Zhu J. Sparse multivariate regression with covariance estimation. Journal of Computational and Graphical Statistics. 2010; 19:947–962. [PubMed: 24963268]

Rousseeuw PJ, Croux C. Alternatives to the median absolute deviation. Journal of the American Statistical association. 1993; 88:1273–1283.

Ruttimann UE, Unser M, Rawlings RR, Rio D, Ramsey NF, Mattay VS, Hommer DW, Frank JA, Weinberger DR. Statistical analysis of functional MRI data in the wavelet domain. Medical Imaging, IEEE Transactions on. 1998; 17:142–154.

Shen D, Shen H, Marron J. Consistency of sparse PCA in high dimension, low sample size contexts. Journal of Multivariate Analysis. 2013; 115:317–333.

Tsybakov AB. Introduction to nonparametric estimation. Springer Series in Statistics; Springer, New York: 2009.

Tyler DE. Radial estimates and the test for sphericity. Biometrika. 1982; 69:429–436.

Vanderbei RJ. Linear programming: Foundations and Extensions. Springer; 2007.

Vershynin R. Introduction to the non-asymptotic analysis of random matrices. arXiv preprint arXiv: 1011.3027. 2010

Visuri S, Koivunen V, Oja H. Sign and rank covariance matrices. Journal of Statistical Planning and Inference. 2000; 91:557–575.

Vogel D, Fried R. Elliptical graphical modelling. Biometrika. 2011; 98:935–951.

Vu VQ, Lei J. Minimax sparse principal subspace estimation in high dimensions. The Annals of Statistics. 2013; 41:2905–2947.

Wegkamp M, Zhao Y. Adaptive estimation of the copula correlation matrix for semiparametric elliptical copulas. Bernoulli. 2016; 22:1184–1226.

Wu Y, Liu Y. Variable selection in quantile regression. Statistica Sinica. 2009; 19:801.

Yuan M. High dimensional inverse covariance matrix estimation via linear programming. The Journal of Machine Learning Research. 2010; 11:2261–2286.

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

Zhao T, Roeder K, Liu H. Positive semidefinite rank-based correlation matrix estimation with application to semiparametric graph estimation. Journal of Computational and Graphical Statistics. 2014; 23:895–922. [PubMed: 25382957]

Zou H, Hastie T, Tibshirani R. Sparse principal component analysis. Journal of computational and graphical statistics. 2006; 15:265–286.

Zou H, Yuan M. Composite quantile regression and the oracle model selection theory. The Annals of Statistics. 2008:1108–1126.

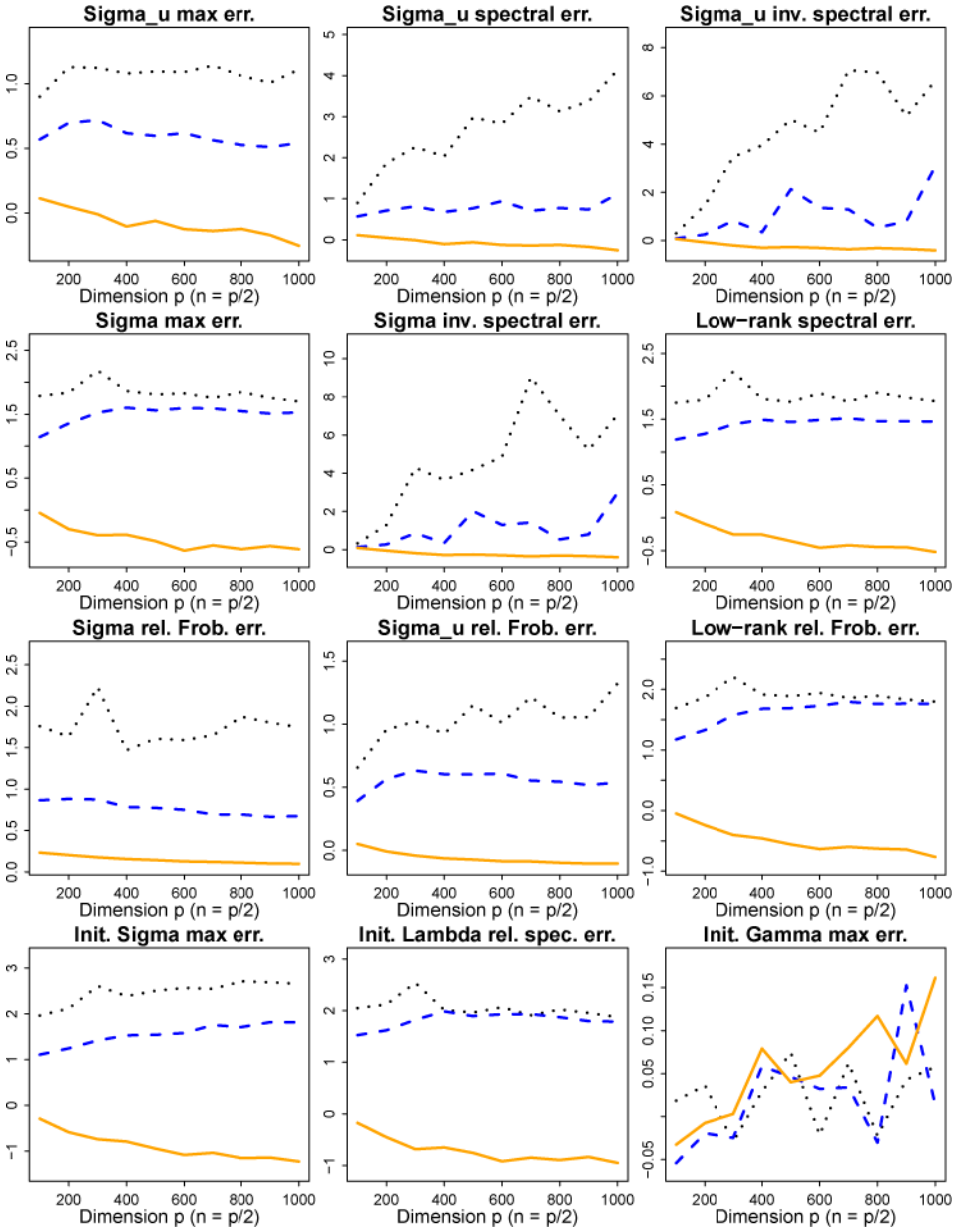

Fig 1.

Conditional sparse covariance matrix estimation. The 12 plots correspond to logarithms (base 2) of the ratios of average errors of the original and the robust POET estimators, measured in different norms. Data were generated from multivariate t-distribution with degree of freedom $\nu = 4.2$ (black dotted), $\nu = 7$ (blue dashed), $\nu = \infty$ (orange solid) with p from 100 to 1000, $n = p/2$ and $m = 3$. 100 simulations were conducted for each p.

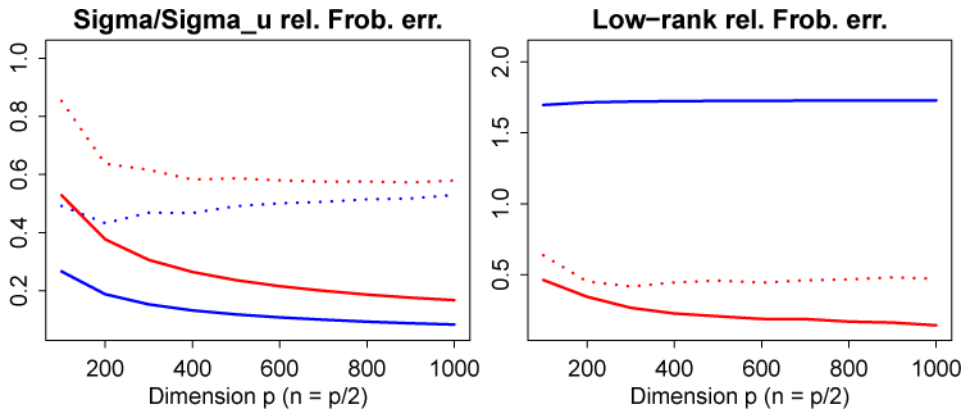

Fig 2.

Comparison of relative Frobenius norms. The plots correspond to average errors of the original POET (for $\nu=\infty$, solid) and the robust POET (for $\nu=4.2$, dotted). In each setting, we compare $\|\hat{\Sigma}^{\mathcal{T}} - \Sigma\|_{\Sigma}$ (red) and $\|\hat{\Sigma}_u^{\mathcal{T}} - \Sigma\|_{\Sigma}$ (blue) in the left panel and compare $\|\mathbf{L}^{-1/2}\left(\hat{\mathbf{\Gamma}}\hat{\mathbf{\Lambda}}\hat{\mathbf{\Gamma}}' - \mathbf{B}\mathbf{B}'\right)\mathbf{L}^{-1/2}\|_F$ (red) and $\|\mathbf{L}^{-1/2}\mathbf{B}\mathbf{B}'\mathbf{L}^{-1/2}\|_F$ (blue) in the right panel.

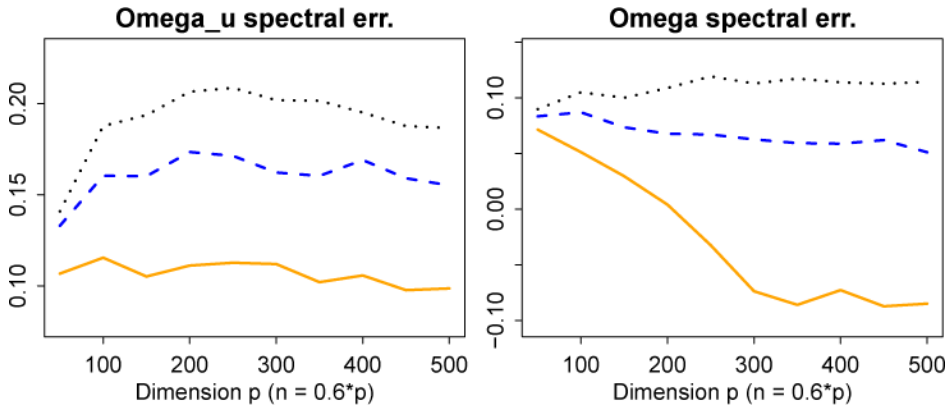

Fig 3.

Conditional graphical model estimation. The plots correspond to log ratios (base 2) of average errors of the original and the robust POET estimators for $\hat{\mathbf{\Omega}}_u$ and $\hat{\mathbf{\Omega}}$, measured in spectral norms. Data were generated from multivariate t-distribution with degree of freedom $\nu = 4.2$ (black dotted), $\nu = 7$ (blue dashed), $\nu = \infty$ (orange solid) with p from 50 to 500, $n = 0.6p$ and $m = 3$. 100 simulations were conducted for each p.