Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2013)
75, *Part* 4, *pp.* 603–680

Large covariance estimation by thresholding principal orthogonal complements

Jianqing Fan,

Princeton University, USA

Yuan Liao

University of Maryland, College Park, USA

and Martina Mincheva

Princeton University, USA

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, February 13th, 2013*, Professor G. A. Young *in the Chair*]

Summary. The paper deals with the estimation of a high dimensional covariance with a conditional sparsity structure and fast diverging eigenvalues. By assuming a sparse error covariance matrix in an approximate factor model, we allow for the presence of some cross-sectional correlation even after taking out common but unobservable factors. We introduce the principal orthogonal complement thresholding method 'POET' to explore such an approximate factor structure with sparsity. The POET-estimator includes the sample covariance matrix, the factor-based covariance matrix, the thresholding estimator and the adaptive thresholding estimator as specific examples. We provide mathematical insights when the factor analysis is approximately the same as the principal component analysis for high dimensional data. The rates of convergence of the sparse residual covariance matrix and the conditional sparse covariance matrix are studied under various norms. It is shown that the effect of estimating the unknown factors vanishes as the dimensionality increases. The uniform rates of convergence for the unobserved factors and their factor loadings are derived. The asymptotic results are also verified by extensive simulation studies. Finally, a real data application on portfolio allocation is presented.

Keywords: Approximate factor model; Cross-sectional correlation; Diverging eigenvalues; High dimensionality; Low rank matrix; Principal components; Sparse matrix; Thresholding; Unknown factors

1. Introduction

Information and technology make large data sets widely available for scientific discovery. Much statistical analysis of such high dimensional data involves the estimation of a covariance matrix or its inverse (the precision matrix). Examples include portfolio management and risk assessment (Fan *et al.*, 2008), high dimensional classification such as the Fisher discriminant (Hastie *et al.*, 2009), graphic models (Meinshausen and Bühlmann, 2006), statistical inference such as controlling false discoveries in multiple testing (Leek and Storey, 2008; Efron, 2010), finding quantitative trait loci based on longitudinal data (Yap *et al.*, 2009; Xiong *et al.*, 2011) and testing the capital asset pricing model (Sentana, 2009), among others. See Section 5 for some of those

Address for correspondence: Jianqing Fan, Department of Operations Research and Financial Engineering, Sherrerd Hall, Princeton University, Princeton, NJ 08544, USA.
E-mail: jqfan@princeton.edu

 1369–7412/13/75603

applications. Yet, the dimensionality is often either comparable with the sample size or even larger. In such cases, the sample covariance is known to have poor performance (Johnstone, 2001), and some regularization is needed.

Realizing the importance of estimating large covariance matrices and the challenges that are brought by the high dimensionality, in recent years researchers have proposed various regularization techniques to estimate $\mathbf{\Sigma}$ consistently. One of the key assumptions is that the covariance matrix is sparse, namely many entries are 0 or nearly so (Bickel and Levina, 2008; Rothman *et al.*, 2009; Lam and Fan, 2009; Cai and Zhou, 2012; Cai and Liu, 2011). In many applications, however, the sparsity assumption directly on $\mathbf{\Sigma}$ is not appropriate. For example, financial returns depend on the equity market risks, housing prices depend on the economic health and gene expressions can be stimulated by cytokines, among others. Because of the presence of common factors, it is unrealistic to assume that many outcomes are uncorrelated. An alternative method is to assume a factor model structure, as in Fan *et al.* (2008). However, they restrict themselves to the strict factor models with known factors.

A natural extension is conditional sparsity. Given the common factors, the outcomes are weakly correlated. To do so, we consider an approximate factor model, which has been frequently used in economic and financial studies (Chamberlain and Rothschild (1983), Fama and French (1992) and Bai and Ng (2002), among others):

$$y_{it} = \mathbf{b}_i' \mathbf{f}_t + u_{it}. \tag{1.1}$$

Here y_{it} is the observed response for the ith ($i = 1, \ldots, p$) individual at time $t = 1, \ldots, T$, $\mathbf{b}_i$ is a vector of factor loadings, $\mathbf{f}_t$ is a $K \times 1$ vector of common factors and u_{it} is the error term, which is usually called the *idiosyncratic component*, uncorrelated with $\mathbf{f}_t$. Both p and T diverge to ∞, whereas K is assumed fixed throughout the paper, and p is possibly much larger than T.

We emphasize that, in model (1.1), only y_{it} is observable. It is intuitively clear that the unknown common factors can only be inferred reliably when there are sufficiently many cases, i.e. $p \to \infty$. In a data rich environment, p can diverge at a rate that is faster than T. The factor model (1.1) can be put in a matrix form as

$$\mathbf{y}_t = \mathbf{B}\mathbf{f}_t + \mathbf{u}_t, \tag{1.2}$$

where $\mathbf{y}_t = (y_{1t}, \ldots, y_{pt})'$, $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_p)'$ and $\mathbf{u}_t = (u_{1t}, \ldots, u_{pt})'$. We are interested in $\mathbf{\Sigma}$, the $p \times p$ covariance matrix of $\mathbf{y}_t$, and its inverse, which are assumed to be time invariant. Under model (1.1), $\mathbf{\Sigma}$ is given by

$$\mathbf{\Sigma} = \mathbf{B}\,\mathrm{cov}(\mathbf{f}_t)\mathbf{B}' + \mathbf{\Sigma}_u, \tag{1.3}$$

where $\mathbf{\Sigma}_u = (\sigma_{u,ij})_{p\times p}$ is the covariance matrix of $\mathbf{u}_t$. The literature on approximate factor models typically assumes that the first K eigenvalues of $\mathbf{B}\,\mathrm{cov}(\mathbf{f}_t)\mathbf{B}'$ diverge at rate $O(p)$, whereas all the eigenvalues of $\mathbf{\Sigma}_u$ are bounded as $p \to \infty$. This assumption holds easily when the factors are pervasive in the sense that a non-negligible fraction of factor loadings should be non-vanishing. The decomposition (1.3) is then asymptotically identified as $p \to \infty$. In addition to it, in this paper we assume that $\mathbf{\Sigma}_u$ is *approximately sparse* as in Bickel and Levina (2008) and Rothman *et al.* (2009): for some $q \in [0, 1)$,

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

does not grow too fast as $p \to \infty$. In particular, this includes the exact sparsity assumption ($q = 0$) under which $m_p = \max_{i \leqslant p} \sum_{j \leqslant p} I_{(\sigma_{u,ij} \neq 0)}$, the maximum number of non-zero elements in each row.

The conditional sparsity structure of form (1.2) was explored by Fan *et al.* (2011a) in estimating the covariance matrix, when the factors $\{\mathbf{f}_t\}$ are observable. This allows them to use regression analysis to estimate $\{\mathbf{u}_t\}_{t=1}^T$. This paper deals with the situation in which the factors are unobservable and must be inferred. Our approach is simple and optimization free and it uses the data only through the sample covariance matrix. Run the singular value decomposition on the sample covariance matrix $\hat{\mathbf{\Sigma}}_{\text{sam}}$ of $\mathbf{y}_t$, keep the covariance matrix that is formed by the first K principal components and apply the thresholding procedure to the remaining covariance matrix. This results in a principal orthogonal complement thresholding estimator POET. When the number of common factors K is unknown, it can be estimated from the data. See Section 2 for additional details. We shall investigate various properties of POET under the assumption that the data are serially dependent, which includes independent observations as a specific example. The rate of convergence under various norms for both estimated $\mathbf{\Sigma}$ and $\mathbf{\Sigma}_u$ and their precision (inverse) matrices will be derived. We show that the effect of estimating the unknown factors on the rate of convergence vanishes when $p\log(p) \gg T$ and, in particular, the rate of convergence for $\mathbf{\Sigma}_u$ achieves the optimal rate in Cai and Zhou (2012).

This paper focuses on the high dimensional *static factor model* (1.2), which is innately related to the principal component analysis (PCA), as clarified in Section 2. This feature makes it different from the classical factor model with fixed dimensionality (e.g. Lawley and Maxwell (1971)). In the last decade, much theory on the estimation and inference of the static factor model has been developed, e.g. Stock and Watson (1998, 2002), Bai and Ng (2002), Bai (2003) and Doz *et al.* (2011), among others. Our contribution is on the estimation of covariance matrices and their inverse in large factor models.

The *static* model that is considered in this paper is to be distinguished from the *dynamic factor model* as in Forni *et al.* (2000); the latter allows $\mathbf{y}_t$ to depend also on $\mathbf{f}_t$ with lags in time. Their approach is based on the eigenvalues and principal components of spectral density matrices, and on the frequency domain analysis. Moreover, as shown in Forni and Lippi (2001), the dynamic factor model does not really impose a restriction on the data-generating process, and the assumption of idiosyncrasy (in their terminology, a p-dimensional process is idiosyncratic if all the eigenvalues of its spectral density matrix remain bounded as $p \to \infty$) asymptotically identifies the decomposition of y_{it} into the common component and idiosyncratic error. The literature includes, for example, Forni *et al.* (2000, 2004), Forni and Lippi (2001), Hallin and Liška (2007, 2011) and many other references therein. Above all, both the static and the dynamic factor models are receiving increasing attention in applications of many fields where information usually is scattered through a (very) large number of interrelated time series.

There has been extensive literature in recent years that deals with sparse principal components, which has been widely used to enhance the convergence of the principal components in high dimensional space. d'Aspremont *et al.* (2008), Shen and Huang (2008), Witten *et al.* (2009) and Ma (2013) proposed and studied various algorithms for computations. More literature on sparse PCA is found in Johnstone and Lu (2009), Amini and Wainwright (2009), Zhang and El Ghaoui (2011) and Birnbaum *et al.* (2012), among others. In addition, there has also been a growing literature that theoretically studies the recovery from a low rank plus sparse matrix estimation problem; see, for example, Wright *et al.* (2009), Lin *et al.* (2009), Candès *et al.* (2011), Luo (2011), Agarwal *et al.* (2012) and Pati *et al.* (2012). It corresponds to the identifiability issue of our problem.

There is a big difference between our model and those considered in the aforementioned literature. In the current paper, the first K eigenvalues of $\mathbf{\Sigma}$ are spiked and grow at a rate $O(p)$, whereas the eigenvalues of the matrices that have been studied in the existing literature

on covariance estimation are usually assumed to be either bounded or slowly growing. Because of this distinctive feature, the common components and the idiosyncratic components can be identified and, in addition, PCA on the sample covariance matrix can consistently estimate the space that is spanned by the eigenvectors of $\mathbf{\Sigma}$. The existing methods of either thresholding directly or solving a constrained optimization method can fail in the presence of very spiked principal eigenvalues. However, there is a price to pay here: as the first K eigenvalues are 'too spiked', one can hardly obtain a satisfactory rate of convergence for estimating $\mathbf{\Sigma}$ in absolute terms, but it can be estimated accurately in relative terms (see Section 3.3 for details). In addition, $\mathbf{\Sigma}^{-1}$ can be estimated accurately.

We would like to note further that the low rank plus sparse representation of our model is on the population covariance matrix, whereas Candès *et al.* (2011), Wright *et al.* (2009) and Lin *et al.* (2009) considered such a representation on the data matrix. (We thank a referee for reminding us about these related works.) As there is no $\mathbf{\Sigma}$ to estimate, their goal is limited to producing a low rank plus sparse matrix decomposition of the data matrix, which corresponds to the identifiability issue of our study, and does not involve estimation and inference. In contrast, our ultimate goal is to estimate the population covariance matrices as well as the precision matrices. For this, we require the idiosyncratic components and common factors to be uncorrelated and the data-generating process to be strictly stationary. The covariances that are considered in this paper are constant over time, though slow time varying covariance matrices are applicable through localization in time (time domain smoothing). Our consistency result on $\mathbf{\Sigma}_u$ demonstrates that decomposition (1.3) is identifiable, and hence our results also shed the light of the 'surprising phenomenon' of Candès *et al.* (2011) that one can separate fully a sparse matrix from a low rank matrix when only the sum of these two components is available.

The rest of the paper is organized as follows. Section 2 gives our estimation procedures and builds the relationship between the PCA and the factor analysis in high dimensional space. Section 3 provides the asymptotic theory for various estimated quantities. Section 4 illustrates how to choose the thresholds by using cross-validation and guarantees the positive definiteness in any finite sample. Specific applications of regularized covariance matrices are given in Section 5. Numerical results are reported in Section 6. Finally, Section 7 presents a real data application on portfolio allocation. All proofs are given in Appendix A. Throughout the paper, we use $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ to denote the minimum and maximum eigenvalues of a matrix $\mathbf{A}$. We also denote by $\|\mathbf{A}\|_{\mathrm{F}}$, $\|\mathbf{A}\|$, $\|\mathbf{A}\|_1$ and $\|\mathbf{A}\|_{\max}$ the Frobenius norm, spectral norm (also called the operator norm), L_1-norm and elementwise norm of a matrix $\mathbf{A}$, defined respectively by $\|\mathbf{A}\|_{\mathrm{F}} = \mathrm{tr}^{1/2}(\mathbf{A}'\mathbf{A})$, $\|\mathbf{A}\| = \lambda_{\max}^{1/2}(\mathbf{A}'\mathbf{A})$, $\|\mathbf{A}\|_1 = \max_j \Sigma_i |a_{ij}|$ and $\|\mathbf{A}\|_{\max} = \max_{i,j}|a_{ij}|$. When $\mathbf{A}$ is a vector, both $\|\mathbf{A}\|_{\mathrm{F}}$ and $\|\mathbf{A}\|$ are equal to the Euclidean norm. Finally, for two sequences, we write $a_T \gg b_T$ if $b_T = o(a_T)$ and $a_T \asymp b_T$ if $a_T = O(b_T)$ and $b_T = O(a_T)$.

The programs that were used to analyse the data can be obtained from

`http://www.blackwellpublishing.com/rss`

2. Regularized covariance matrix via principal components analysis

There are three main objectives of this paper:

(a) to understand the relationship between PCA and high dimensional factor analysis;
(b) to estimate both covariance matrices $\mathbf{\Sigma}$ and the idiosyncratic $\mathbf{\Sigma}_u$ and their precision matrices in the presence of common factors;
(c) to investigate the effect of estimating the unknown factors on the covariance estimation.

The propositions in Section 2.1 show that the space that is spanned by the principal components in the population level $\mathbf{\Sigma}$ is close to the space that is spanned by the columns of the factor loading matrix $\mathbf{B}$.

2.1. High dimensional principal components analysis and factor model

Consider a factor model

$$y_{it} = \mathbf{b}_i' \mathbf{f}_t + u_{it}, \qquad i \leqslant p, t \leqslant T,$$

where the number of common factors, $K = \dim(\mathbf{f}_t)$, is small compared with p and T, and thus is assumed to be fixed throughout the paper. In the model, the only observable variable is the data y_{it}. One of the distinguished features of the factor model is that the principal eigenvalues of $\mathbf{\Sigma}$ are no longer bounded, but growing fast with the dimensionality. We illustrate this in the following example.

2.1.1. Example 1

Consider a single-factor model $y_{it} = b_i f_t + u_{it}$ where $b_i \in \mathbb{R}$. Suppose that the factor is pervasive in the sense that it has non-negligible effect on a non-vanishing proportion of outcomes. It is then reasonable to assume that $\Sigma_{i=1}^p b_i^2 > cp$ for some $c > 0$. Therefore, assuming that $\lambda_{\max}(\mathbf{\Sigma}_u) = o(p)$, an application of decomposition (1.3) yields

$$\lambda_{\max}(\mathbf{\Sigma}) \geqslant \operatorname{var}(f_t) \sum_{i=1}^{p} b_i^2 - \lambda_{\max}(\mathbf{\Sigma}_u) > \frac{c}{2} \operatorname{var}(f_t) p$$

for all large p, assuming that $\operatorname{var}(f_t) > 0$.

We now elucidate why PCA can be used for the factor analysis in the presence of spiked eigenvalues. Write $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_p)'$ as the $p \times K$ loading matrix. Note that the linear space that is spanned by the first K principal components of $\mathbf{B} \operatorname{cov}(\mathbf{f}_t) \mathbf{B}'$ is the same as that spanned by the columns of $\mathbf{B}$ when $\operatorname{cov}(\mathbf{f}_t)$ is non-degenerate. Thus, we can assume without loss of generality that the columns of $\mathbf{B}$ are orthogonal and $\operatorname{cov}(\mathbf{f}_t) = \mathbf{I}_K$, the identity matrix. This canonical form corresponds to the identifiability condition in decomposition (1.3). Let $\tilde{\mathbf{b}}_1, \ldots, \tilde{\mathbf{b}}_K$ be the columns of $\mathbf{B}$, ordered such that $\{\|\tilde{\mathbf{b}}_j\|\}_{j=1}^K$ is in a non-increasing order. Then, $\{\tilde{\mathbf{b}}_j / \|\tilde{\mathbf{b}}_j\|\}_{j=1}^K$ are eigenvectors of the matrix $\mathbf{BB}'$ with eigenvalues $\{\|\tilde{\mathbf{b}}_j\|^2\}_{j=1}^K$ and the rest 0. We shall impose the pervasiveness assumption that all eigenvalues of the $K \times K$ matrix $p^{-1}\mathbf{B}'\mathbf{B}$ are bounded away from 0, which holds if the factor loadings $\{\mathbf{b}_i\}_{i=1}^p$ are independent realizations from a non-degenerate population. Since the non-vanishing eigenvalues of the matrix $\mathbf{BB}'$ are the same as those of $\mathbf{B}'\mathbf{B}$, from the pervasiveness assumption it follows that $\{\|\tilde{\mathbf{b}}_j\|^2\}_{j=1}^K$ are all growing at rate $O(p)$.

Let $\{\lambda_j\}_{j=1}^p$ be the eigenvalues of $\mathbf{\Sigma}$ in a descending order and $\{\boldsymbol{\xi}_j\}_{j=1}^p$ be their corresponding eigenvectors. Then, an application of Weyl's eigenvalue theorem (see Appendix A) yields the following proposition.

Proposition 1. Assume that the eigenvalues of $p^{-1}\mathbf{B}'\mathbf{B}$ are bounded away from 0 for all large p. For the factor model (1.3) with the canonical condition

$$\operatorname{cov}(\mathbf{f}_t) = \mathbf{I}_K \text{ and } \mathbf{B}'\mathbf{B} \text{ is diagonal}, \tag{2.1}$$

we have

$$\begin{aligned} |\lambda_j - \|\tilde{\mathbf{b}}_j\|^2| &\leqslant \|\mathbf{\Sigma}_u\|, \qquad \text{for } j \leqslant K, \\ |\lambda_j| &\leqslant \|\mathbf{\Sigma}_u\|, \qquad \text{for } j > K. \end{aligned}$$

In addition, for $j \leqslant K$, $\liminf_{p\to\infty} \|\tilde{\mathbf{b}}_j\|^2/p > 0$.

Using proposition 1 and the $\sin(\theta)$ theorem of Davis and Kahn (1970) (see their appendix), we have the following proposition.

Proposition 2. Under the assumptions of proposition 1, if $\{\|\tilde{\mathbf{b}}_j\|\}_{j=1}^K$ are distinct, then

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

Propositions 1 and 2 state that PCA and factor analysis are approximately the same if $\|\boldsymbol{\Sigma}_u\| = o(p)$. This is assured through a sparsity condition on $\boldsymbol{\Sigma}_u = (\sigma_{u,ij})_{p\times p}$, which is frequently measured through

$$m_p = \max_{i\leqslant p} \sum_{j\leqslant p} |\sigma_{u,ij}|^q, \qquad \text{for some } q \in [0,1]. \tag{2.2}$$

The intuition is that, after taking out the common factors, many pairs of the cross-sectional units become weakly correlated. This generalized notion of sparsity was used in Bickel and Levina (2008) and Cai and Liu (2011). Under this generalized measure of sparsity, we have

$$\|\boldsymbol{\Sigma}_u\| \leqslant \|\boldsymbol{\Sigma}_u\|_1 \leqslant \max_i \sum_{j=1}^p |\sigma_{u,ij}|^q (\sigma_{u,ii}\sigma_{u,jj})^{(1-q)/2} = O(m_p),$$

if the noise variances $\{\sigma_{u,ii}^2\}$ are bounded. Therefore, when $m_p = o(p)$, proposition 1 implies that we have distinguished eigenvalues between the principal components $\{\lambda_j\}_{j=1}^K$ and the rest of the components $\{\lambda_j\}_{j=K+1}^p$ and proposition 2 ensures that the first K principal components are approximately the same as the columns of the factor loadings.

The aforementioned sparsity assumption appears reasonable in empirical applications. Boivin and Ng (2006) conducted an empirical study and showed that imposing zero correlation between weakly correlated idiosyncratic components improves the forecast. (We thank a referee for this interesting reference.) More recently, Phan (2012) empirically estimated the level of sparsity of the idiosyncratic covariance by using UK market data.

Recent developments on random-matrix theory, e.g. Johnstone and Lu (2009) and Paul (2007), have shown that, when p/T is not negligible, the eigenvalues and eigenvectors of $\boldsymbol{\Sigma}$ might not be consistently estimated from the sample covariance matrix. A distinguished feature of the covariance that is considered in this paper is that there are some very spiked eigenvalues. By propositions 1 and 2, in the factor model, the pervasiveness condition

$$\lambda_{\min}(p^{-1}\mathbf{B}'\mathbf{B}) > c > 0 \tag{2.3}$$

implies that the first K eigenvalues are growing at a rate p. Moreover, when p is large, the principal components $\{\boldsymbol{\xi}_j\}_{j=1}^K$ are close to the normalized vectors $\{\tilde{\mathbf{b}}_j\}_{j=1}^K$ when $m_p = o(p)$. This provides the mathematics for using the first K principal components as a proxy for the space that is spanned by the columns of the factor loading matrix $\mathbf{B}$. In addition, because of condition (2.3), the signals of the first K eigenvalues are stronger than those of the spiked covariance model that was considered by Jung and Marron (2009) and Birnbaum *et al.* (2012). Therefore, our other conditions for the consistency of principal components at the population level are much weaker than those in the spiked covariance literature. However, this also shows that, under our setting, PCA is a valid approximation to factor analysis only if $p \to \infty$. The fact that PCA on the sample covariance is inconsistent when p is bounded has also previously been demonstrated in the literature (see, for example, Bai (2003)).

With assumption (2.3), the standard literature on approximate factor models has shown that

PCA on the sample covariance matrix $\hat{\mathbf{\Sigma}}_{\text{sam}}$ can consistently estimate the space that is spanned by the factor loadings (e.g. Stock and Watson (1998) and Bai (2003)). Our contribution in propositions 1 and 2 is that we connect the high dimensional factor model to the principal components and obtain the consistency of the spectrum in the population level $\mathbf{\Sigma}$ instead of the sample level $\hat{\mathbf{\Sigma}}_{\text{sam}}$. The spectral consistency also enhances the results in Chamberlain and Rothschild (1983). This provides the rationale behind the consistency results in the factor model literature.

2.2. Principal orthogonal complement thresholding

A sparsity assumption directly on $\mathbf{\Sigma}$ is inappropriate in many applications owing to the presence of common factors. Instead, we propose a non-parametric estimator of $\mathbf{\Sigma}$ based on PCA. Let $\hat{\lambda}_1 \geqslant \hat{\lambda}_2 \geqslant \ldots \geqslant \hat{\lambda}_p$ be the ordered eigenvalues of the sample covariance matrix $\hat{\mathbf{\Sigma}}_{\text{sam}}$ and $\{\hat{\boldsymbol{\xi}}_i\}_{i=1}^p$ be their corresponding eigenvectors. Then the sample covariance has the following spectral decomposition:

$$\hat{\mathbf{\Sigma}}_{\text{sam}} = \sum_{i=1}^{K} \hat{\lambda}_i \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i' + \hat{\mathbf{R}}_K, \tag{2.4}$$

where $\hat{\mathbf{R}}_K = \Sigma_{i=K+1}^p \hat{\lambda}_i \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i' = (\hat{r}_{ij})_{p\times p}$ is the principal orthogonal complement, and K is the number of diverging eigenvalues of $\mathbf{\Sigma}$. Let us first assume that K is known.

Now we apply thresholding on $\hat{\mathbf{R}}_K$. Define

$$\hat{\mathbf{R}}_K^{\mathcal{T}} = (\hat{r}_{ij}^{\mathcal{T}})_{p\times p}, \qquad \hat{r}_{ij}^{\mathcal{T}} = \begin{cases} \hat{r}_{ii}, & i = j, \\ s_{ij}(\hat{r}_{ij})\, I(|\hat{r}_{ij}| \geqslant \tau_{ij}), & i \neq j, \end{cases} \tag{2.5}$$

where $s_{ij}(\cdot)$ is a generalized shrinkage function of Antoniadis and Fan (2001), employed by Rothman *et al.* (2009) and Cai and Liu (2011), and $\tau_{ij} > 0$ is an entry-dependent threshold. In particular, the hard thresholding rule $s_{ij}(x) = x\, I(|x| \geqslant \tau_{ij})$ (Bickel and Levina, 2008) and the constant thresholding parameter $\tau_{ij} = \delta$ are allowed. In practice, it is more desirable to have τ_{ij} entry adaptive. An example of the adaptive thresholding is

$$\tau_{ij} = \tau(\hat{r}_{ii}\hat{r}_{jj})^{1/2}, \qquad \text{for a given } \tau > 0, \tag{2.6}$$

where $\hat{r}_{ii}$ is the ith diagonal element of $\hat{\mathbf{R}}_K$. This corresponds to applying the thresholding with parameter τ to the correlation matrix of $\hat{\mathbf{R}}_K$.

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

$$\hat{\mathbf{\Sigma}}_K = \sum_{i=1}^{K} \hat{\lambda}_i \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i' + \hat{\mathbf{R}}_K^{\mathcal{T}}. \tag{2.7}$$

We shall call this estimator the principal orthogonal complement thresholding estimator POET. It is obtained by thresholding the remaining components of the sample covariance matrix, after taking out the first K principal components. One of the attractive features of POET is that it is optimization free and hence is computationally appealing. (We have written an R package for POET, which outputs the estimated $\mathbf{\Sigma}$, $\mathbf{\Sigma}_u$, K, the factors and the loadings.)

With the choice of τ_{ij} in expression (2.6) and the hard thresholding rule, our estimator encompasses many popular estimators as its specific cases. When $\tau = 0$, the estimator is the sample covariance matrix and, when $\tau = 1$, the estimator becomes that based on the strict factor model (Fan *et al.*, 2008). When $K = 0$, our estimator is the same as the thresholding estimator of Bickel and Levina (2008) and (with a more general thresholding function) Rothman *et al.* (2009) or the adaptive thresholding estimator of Cai and Liu (2011) with a proper choice of τ_{ij}.

In practice, the number of diverging eigenvalues (or common factors) can be estimated on the basis of the sample covariance matrix. Determining K in a data-driven way is an important topic and is well understood in the literature. We shall describe the estimator POET with a data-driven K in Section 2.4.

2.3. Least squares point of view

The estimator POET (2.7) has an equivalent representation using a constrained least squares method. The least squares method seeks $\hat{\mathbf{\Lambda}}_K = (\hat{\mathbf{b}}_1^K, \ldots, \hat{\mathbf{b}}_p^K)'$ and $\hat{\mathbf{F}}_K' = (\hat{\mathbf{f}}_1^K, \ldots, \hat{\mathbf{f}}_T^K)$ such that

$$(\hat{\mathbf{\Lambda}}_K, \hat{\mathbf{F}}_K) = \arg \min_{\mathbf{b}_i \in \mathbb{R}^K, \mathbf{f}_t \in \mathbb{R}^K} \sum_{i=1}^{p} \sum_{t=1}^{T} (y_{it} - \mathbf{b}_i' \mathbf{f}_t)^2, \tag{2.8}$$

subject to the normalization

$$\frac{1}{T} \sum_{t=1}^{T} \mathbf{f}_t \mathbf{f}_t' = \mathbf{I}_K, \quad \text{and} \quad \frac{1}{p} \sum_{i=1}^{p} \mathbf{b}_i \mathbf{b}_i' \text{ is diagonal.} \tag{2.9}$$

The constraints (2.9) correspond to the normalization (2.1). Here we assume that the mean of each variable $\{y_{it}\}_{t=1}^T$ has been removed, i.e. $E(y_{it}) = E(f_{jt}) = 0$ for all $i \leqslant p$, $j \leqslant K$ and $t \leqslant T$. Putting it in a matrix form, the optimization problem can be written as

$$\begin{aligned} &\arg \min_{\mathbf{B}, \mathbf{F}} \|\mathbf{Y} - \mathbf{B}\mathbf{F}'\|_{\mathrm{F}}^2, \\ &T^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_K, \quad \mathbf{B}'\mathbf{B} \text{ is diagonal,} \end{aligned} \tag{2.10}$$

where $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_T)$ and $\mathbf{F}' = (\mathbf{f}_1, \ldots, \mathbf{f}_T)$. For each given $\mathbf{F}$, the least squares estimator of $\mathbf{B}$ is $\mathbf{\Lambda} = T^{-1}\mathbf{Y}\mathbf{F}$, using the constraint (2.9) on the factors. Substituting this into problem (2.10), the objective function now becomes $\|\mathbf{Y} - T^{-1}\mathbf{Y}\mathbf{F}\mathbf{F}'\|_{\mathrm{F}}^2 = \mathrm{tr}\{(\mathbf{I}_T - T^{-1}\mathbf{F}\mathbf{F}')\mathbf{Y}'\mathbf{Y}\}$. The minimizer is now clear: the columns of $\hat{\mathbf{F}}_K/\sqrt{T}$ are the eigenvectors corresponding to the K largest eigenvalues of the $T \times T$ matrix $\mathbf{Y}'\mathbf{Y}$ and $\hat{\mathbf{\Lambda}}_K = T^{-1}\mathbf{Y}\hat{\mathbf{F}}_K$ (see, for example, Stock and Watson (2002)).

We shall show that under some mild regularity conditions, as p and $T \to \infty$, $\hat{\mathbf{b}}_i^{K'} \hat{\mathbf{f}}_t^K$ consistently estimates the true $\mathbf{b}_i'\mathbf{f}_t$ uniformly over $i \leqslant p$ and $t \leqslant T$. Since $\mathbf{\Sigma}_u$ is assumed to be sparse, we can construct an estimator of $\mathbf{\Sigma}_u$ by using the adaptive thresholding method by Cai and Liu (2011) as follows. Let $\hat{u}_{it} = y_{it} - \hat{\mathbf{b}}_i^{K'} \hat{\mathbf{f}}_t^K$, $\hat{\sigma}_{ij} = (1/T)\sum_{t=1}^T \hat{u}_{it}\hat{u}_{jt}$ and $\hat{\theta}_{ij} = (1/T)\sum_{t=1}^T (\hat{u}_{it}\hat{u}_{jt} - \hat{\sigma}_{ij})^2$. For some predetermined decreasing sequence $\omega_T > 0$, and sufficiently large $C > 0$, define the adaptive threshold parameter as $\tau_{ij} = C\omega_T\sqrt{\hat{\theta}_{ij}}$. The estimated idiosyncratic covariance estimator is then given by

$$\hat{\mathbf{\Sigma}}_{u,K}^{\mathcal{T}} = (\hat{\sigma}_{ij}^{\mathcal{T}})_{p\times p}, \qquad \hat{\sigma}_{ij}^{\mathcal{T}} = \begin{cases} \hat{\sigma}_{ii}, & i = j, \\ s_{ij}(\hat{\sigma}_{ij}), & i \neq j, \end{cases} \tag{2.11}$$

where, for all $z \in \mathbb{R}$ (see Antoniadis and Fan (2001)),

$$s_{ij}(z) = 0 \text{ when } |z| \leqslant \tau_{ij}, \qquad |s_{ij}(z) - z| \leqslant \tau_{ij}.$$

It is easy to verify that $s_{ij}(\cdot)$ includes many interesting thresholding functions such as hard thresholding ($s_{ij}(z) = zI_{(|z|\geqslant\tau_{ij})}$), soft thresholding ($s_{ij}(z) = \mathrm{sgn}(z)(|z| - \tau_{ij})_+$), smoothly clipped absolute deviation and the adaptive lasso (see Rothman *et al.* (2009)).

Analogous to the decomposition (1.3), we obtain the following substitution estimators:

$$\tilde{\mathbf{\Sigma}}_K = \hat{\mathbf{\Lambda}}_K \hat{\mathbf{\Lambda}}_K' + \hat{\mathbf{\Sigma}}_{u,K}^{\mathcal{T}}, \tag{2.12}$$

and, by the Sherman–Morrison–Woodbury formula, noting that $(1/T)\Sigma_{t=1}^{T}\hat{\mathbf{f}}_t^K\hat{\mathbf{f}}_t^{K'} = \mathbf{I}_K$,

$$(\tilde{\boldsymbol{\Sigma}}_K)^{-1} = (\hat{\boldsymbol{\Sigma}}_{u,K}^{\mathcal{T}})^{-1} - (\hat{\boldsymbol{\Sigma}}_{u,K}^{\mathcal{T}})^{-1}\hat{\boldsymbol{\Lambda}}_K\{\mathbf{I}_K + \hat{\boldsymbol{\Lambda}}_K'(\hat{\boldsymbol{\Sigma}}_{u,K}^{\mathcal{T}})^{-1}\hat{\boldsymbol{\Lambda}}_K\}^{-1}\hat{\boldsymbol{\Lambda}}_K'(\hat{\boldsymbol{\Sigma}}_{u,K}^{\mathcal{T}})^{-1}. \tag{2.13}$$

In practice, the true number of factors K might be unknown to us. However, for any determined $K_1 \leqslant p$, we can always construct either $(\hat{\boldsymbol{\Sigma}}_{K_1}, \hat{\mathbf{R}}_{K_1}^{\mathcal{T}})$ as in estimator (2.7) or $(\tilde{\Sigma}_{K_1}, \hat{\boldsymbol{\Sigma}}_{u,K_1}^{\mathcal{T}})$ as in estimator (2.12) to estimate $(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_u)$. The following theorem shows that, for each given K_1, the two estimators based on either regularized PCA or least squares substitution are equivalent. Similar results were obtained by Bai (2003) when $K_1 = K$ and no thresholding was imposed.

Theorem 1. Suppose that the entry-dependent threshold in definition (2.5) is the same as the thresholding parameter that is used in expression (2.11). Then, for any $K_1 \leqslant p$, estimator (2.7) is equivalent to the substitution estimator (2.12), i.e.

$$\hat{\boldsymbol{\Sigma}}_{K_1} = \tilde{\boldsymbol{\Sigma}}_{K_1}, \quad \text{and} \quad \hat{\boldsymbol{\Sigma}}_{u,K_1}^{\mathcal{T}} = \hat{\mathbf{R}}_{K_1}^{\mathcal{T}}.$$

In this paper, we shall use a data-driven $\hat{K}$ to construct POET (see Section 2.4), which has two equivalent representations according to theorem 1.

2.4. Principal orthogonal complement thresholding with unknown K

Determining the number of factors in a data-driven way has been an important research topic in the econometrics literature. Bai and Ng (2002) proposed a consistent estimator as both p and T diverge. Other recent criteria have been proposed by Kapetanios (2010), Onatski (2010) and Alessi *et al.* (2010), among others.

Our method also allows a data-driven $\hat{K}$ to estimate the covariance matrices. In principle, any procedure that gives a consistent estimate of K can be adopted. In this paper we apply the well-known method in Bai and Ng (2002). It estimates K by

$$\hat{K} = \arg\min_{0 \leqslant K_1 \leqslant M} \log\left(\frac{1}{pT}\|\mathbf{Y} - T^{-1}\mathbf{Y}\hat{\mathbf{F}}_{K_1}\hat{\mathbf{F}}_{K_1}'\|_{\mathrm{F}}^2\right) + K_1\, g(T, p), \tag{2.14}$$

where M is a prescribed upper bound, $\hat{\mathbf{F}}_{K_1}$ is a $T \times K_1$ matrix whose columns are $\sqrt{T}$ times the eigenvectors corresponding to the K_1 largest eigenvalues of the $T \times T$ matrix $\mathbf{Y}'\mathbf{Y}$ and $g(T, p)$ is a penalty function of (p, T) such that $g(T, p) = o(1)$ and $\min\{p, T\}\, g(T, p) \to \infty$. Two examples suggested by Bai and Ng (2002), IC1 and IC2, are respectively

$$g(T, p) = \frac{p+T}{pT}\log\left(\frac{pT}{p+T}\right),$$

$$g(T, p) = \frac{p+T}{pT}\log(\min\{p, T\}).$$

Throughout the paper, we let $\hat{K}$ be the solution to problem (2.14) by using either IC1 or IC2. The asymptotic results are not affected regardless of the specific choice of $g(T, p)$. We define the POET-estimator with unknown K as

$$\hat{\boldsymbol{\Sigma}}_{\hat{K}} = \sum_{i=1}^{\hat{K}} \hat{\lambda}_i\hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i' + \hat{\mathbf{R}}_{\hat{K}}^{\mathcal{T}}. \tag{2.15}$$

The procedure is as stated in Section 2.2 except that $\hat{K}$ is now data driven.

3. Asymptotic properties

3.1. Assumptions

This section presents the assumptions on model (1.2), in which only $\{\mathbf{y}_t\}_{t=1}^T$ are observable. Recall the identifiability condition (2.1).

The first assumption has been one of the most essential in the literature of approximate factor models. Under this assumption and other regularity conditions, the number of factors, loadings and common factors can be consistently estimated (e.g. Stock and Watson (1998, 2002), Bai and Ng (2002) and Bai (2003)).

Assumption 1. All the eigenvalues of the $K \times K$ matrix $p^{-1}\mathbf{B}'\mathbf{B}$ are bounded away from both 0 and ∞ as $p \to \infty$.

Remark 1.

(a) It is implied from proposition 1 in Section 2 that the first K eigenvalues of $\mathbf{\Sigma}$ grow at rate $O(p)$. This unique feature distinguishes our work from most of other work on low rank plus sparse covariances that has been considered in the literature, e.g. Luo (2011), Pati *et al.* (2012), Agarwal *et al.* (2012) and Birnbaum *et al.* (2012). (To our best knowledge, the only other references that estimate large covariances with diverging eigenvalues (growing at the rate of dimensionality $O(p)$) are Fan *et al.* (2008, 2011a) and Bai and Shi (2011). Whereas Fan *et al.* (2008, 2011a) assumed that the factors are observable, Bai and Shi (2011) considered the strict factor model in which $\mathbf{\Sigma}_u$ is diagonal.)

(b) Assumption 1 requires the factors to be pervasive, i.e. to impact a non-vanishing proportion of individual time series. See example 1 in Section 2.1.1 for its meaning. (It is important to distinguish the model that we consider in this paper from the 'sparse factor model' in the literature, e.g. Carvalho *et al.* (2008) and Pati *et al.* (2012), which assumes that the loading matrix $\mathbf{B}$ is sparse. The intuition of a sparse loading matrix is that each factor is related to only a relatively small number of stocks, assets, genes, etc. With $\mathbf{B}$ being sparse, all the eigenvalues of $\mathbf{B}'\mathbf{B}$ and hence those of $\mathbf{\Sigma}$ are bounded.)

(c) As to be illustrated in Section 3.3 below, owing to the fast diverging eigenvalues, we can hardly achieve a good rate of convergence for estimating $\mathbf{\Sigma}$ under either the spectral norm or Frobenius norm when $p > T$. This phenomenon arises naturally from the characteristics of the high dimensional factor model, which is another distinguished feature compared with those convergence results in the existing literature.

Assumption 2.

(a) $\{\mathbf{u}_t, \mathbf{f}_t\}_{t \geqslant 1}$ is strictly stationary. In addition, $E(u_{it}) = E(u_{it} f_{jt}) = 0$ for all $i \leqslant p$, $j \leqslant K$ and $t \leqslant T$.

(b) There are constants c_1, $c_2 > 0$ such that $\lambda_{\min}(\mathbf{\Sigma}_u) > c_1$, $\|\mathbf{\Sigma}_u\|_1 < c_2$ and
$$\min_{i \leqslant p, j \leqslant p} \operatorname{var}(u_{it} u_{jt}) > c_1.$$

(c) There are $r_1, r_2 > 0$ and $b_1, b_2 > 0$ such that, for any $s > 0$, $i \leqslant p$ and $j \leqslant K$,
$$P(|u_{it}| > s) \leqslant \exp\{-(s/b_1)^{r_1}\},$$
$$P(|f_{jt}| > s) \leqslant \exp\{-(s/b_2)^{r_2}\}.$$

Condition (a) requires strict stationarity as well as the non-correlation between $\{\mathbf{u}_t\}$ and $\{\mathbf{f}_t\}$. These conditions are slightly stronger than those in the literature, e.g. Bai (2003), but are still standard and simplify our technicalities. Condition (b) requires that $\mathbf{\Sigma}_u$ be well conditioned. The condition $\|\mathbf{\Sigma}_u\|_1 \leqslant c_2$ instead of a weaker condition $\lambda_{\max}(\mathbf{\Sigma}_u) \leqslant c_2$ is imposed here to estimate K

consistently. But it is still standard in the approximate factor model literature as in Bai and Ng (2008), Bai (2003), etc. When K is known, such a condition can be removed. Fan *et al.* (2011b) shows that the results continue to hold for a growing (known) K under the weaker condition $\lambda_{\max}(\mathbf{\Sigma}_u) \leqslant c_2$. Condition (c) requires exponential-type tails, which allow us to apply the large deviation theory to $(1/T)\Sigma_{t=1}^{T} u_{it}u_{jt} - \sigma_{u,ij}$ and $(1/T)\Sigma_{t=1}^{T} f_{jt}u_{it}$.

We impose the strong mixing condition. Let $\mathcal{F}_{-\infty}^{0}$ and $\mathcal{F}_{T}^{\infty}$ denote the σ-algebras that are generated by $\{(\mathbf{f}_t, \mathbf{u}_t) : t \leqslant 0\}$ and $\{(\mathbf{f}_t, \mathbf{u}_t) : t \geqslant T\}$ respectively. In addition, define the mixing coefficient

$$\alpha(T) = \sup_{A \in \mathcal{F}_{-\infty}^{0}, B \in \mathcal{F}_{T}^{\infty}} |P(A)\,P(B) - P(AB)|. \tag{3.1}$$

Assumption 3 (strong mixing). There exists $r_3 > 0$ such that $3r_1^{-1} + 1.5r_2^{-1} + r_3^{-1} > 1$, and $C > 0$ satisfying, for all $T \in \mathbb{Z}^+$,

$$\alpha(T) \leqslant \exp(-CT^{r_3}).$$

In addition, we impose the following regularity conditions.

Assumption 4. There exists $M > 0$ such that, for all $i \leqslant p$, $t \leqslant T$ and $s \leqslant T$,

(a) $\|\mathbf{b}_i\|_{\max} < M$,
(b) $E[p^{-1/2}\{\mathbf{u}_s'\mathbf{u}_t - E(\mathbf{u}_s'\mathbf{u}_t)\}]^4 < M$ and
(c) $E\|p^{-1/2}\Sigma_{i=1}^{p}\mathbf{b}_i u_{it}\|^4 < M$.

These conditions are needed to estimate consistently the transformed common factors as well as the factor loadings. Similar conditions were also assumed in Bai (2003) and Bai and Ng (2006). The number of factors is assumed to be fixed. Our conditions in assumption 4 are weaker than those in Bai (2003) as we focus on different aspects of the study.

3.2. Convergence of the idiosyncratic covariance

Estimating the covariance matrix $\mathbf{\Sigma}_u$ of the idiosyncratic components $\{\mathbf{u}_t\}$ is important for many statistical inferences. For example, it is needed for large sample inference of the unknown factors and their loadings, for testing the capital asset pricing model (Sentana, 2009) and large-scale hypothesis testing (Fan *et al.*, 2012). See Section 5.

We estimate $\mathbf{\Sigma}_u$ by thresholding the principal orthogonal complements after the first $\hat{K}$ principal components of the sample covariance have been taken out: $\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}} = \hat{\mathbf{R}}_{\hat{K}}^{\mathcal{T}}$. By theorem 1, it also has an equivalent expression given by estimator (2.11), with $\hat{u}_{it} = y_{it} - (\hat{\mathbf{b}}_i^{\hat{K}})'\hat{\mathbf{f}}_t^{\hat{K}}$. Throughout the paper, we apply the adaptive threshold

$$\tau_{ij} = C\omega_T\sqrt{\hat{\theta}_{ij}}, \qquad \omega_T = \frac{1}{\sqrt{p}} + \sqrt{\left\{\frac{\log(p)}{T}\right\}} \tag{3.2}$$

where $C > 0$ is a sufficiently large constant, though the results hold for other types of thresholding. As in Bickel and Levina (2008) and Cai and Liu (2011), the threshold that is chosen in the current paper is in fact obtained from the optimal uniform rate of convergence of $\max_{i \leqslant p, j \leqslant p}|\hat{\sigma}_{ij} - \sigma_{u,ij}|$. When direct observation of u_{it} is not available, the effect of estimating the unknown factors also contributes to this uniform estimation error, which is why $p^{-1/2}$ appears in the threshold.

The following theorem gives the rate of convergence of the estimated idiosyncratic covariance. Let $\gamma^{-1} = 3r_1^{-1} + 1.5r_2^{-1} + r_3^{-1} + 1$. In the convergence rate below, recall that m_p and q are defined in the measure of sparsity (2.2).

Theorem 1. Suppose that $\log(p) = o(T^{\gamma/6})$, $T = o(p^2)$ and assumptions 1–4 hold. Then, for a sufficiently large constant $C > 0$ in the threshold (3.2), the POET-estimator $\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}}$ satisfies

$$\|\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}} - \mathbf{\Sigma}_u\| = O_p(\omega_T^{1-q} m_p).$$

If further $\omega_T^{1-q} m_p = o(1)$, then the eigenvalues of $\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}}$ are all bounded away from 0 with probability approaching 1, and

$$\|(\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}})^{-1} - \mathbf{\Sigma}_u^{-1}\| = O_p(\omega_T^{1-q} m_p).$$

When estimating $\mathbf{\Sigma}_u$, p is allowed to grow exponentially fast in T, and $\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}}$ can be made consistent under the spectral norm. In addition, $\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}}$ is asymptotically invertible whereas the classical sample covariance matrix based on the residuals is not when $p > T$.

Remark 2.

(a) Consistent estimation of $\mathbf{\Sigma}_u$ indicates that $\mathbf{\Sigma}_u$ is identifiable in model (1.3), namely the sparse $\mathbf{\Sigma}_u$ can be separated perfectly from the low rank matrix there. The result here gives another proof (when assuming that $\omega_T^{1-q} m_p = o(1)$) of the 'surprising phenomenon' in Candès *et al.* (2011) under different technical conditions.

(b) Fan *et al.* (2011a) recently showed that, when $\{\mathbf{f}_t\}_{t=1}^T$ are observable and $q = 0$, the rate of convergence of the adaptive thresholding estimator is given by

$$\|\hat{\mathbf{\Sigma}}_u^{\mathcal{T}} - \mathbf{\Sigma}_u\| = O_p\left[m_p \sqrt{\left\{\frac{\log(p)}{T}\right\}}\right] = \|(\hat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1} - \mathbf{\Sigma}_u^{-1}\|.$$

Hence, when the common factors are unobservable, the rate of convergence has an additional term $m_p/\sqrt{p}$, coming from the effect of estimating the unknown factors. This effect vanishes when $p \log(p) \gg T$, in which case the minimax rate as in Cai and Zhou (2012) is achieved. As p increases, more information about the common factors is collected, which results in more accurate estimation of the common factors $\{\mathbf{f}_t\}_{t=1}^T$.

(c) When K is known and grows with p and T, with slightly weaker assumptions, Fan *et al.* (2011b) shows that, under the exactly sparse case (i.e. $q = 0$), the result continues to hold with convergence rate

$$m_p\left[K^2 \sqrt{\left\{\frac{\log(p)}{T}\right\}} + \frac{K^3}{\sqrt{p}}\right].$$

3.3. Convergence of POET

Since the first K eigenvalues of $\mathbf{\Sigma}$ grow with p, we can hardly estimate $\mathbf{\Sigma}$ with satisfactory accuracy in absolute terms. This problem does not arise from the limitation of any estimation method but is due to the nature of the high dimensional factor model. We illustrate this by using a simple example.

3.3.1. Example 2

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

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

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

In the presence of very spiked eigenvalues, although the covariance $\mathbf{\Sigma}$ cannot be consistently estimated in absolute terms, it can be well estimated in terms of the *relative error* matrix

$$\mathbf{\Sigma}^{-1/2}\hat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1/2} - \mathbf{I}_p$$

which is more relevant for many applications (see example 4 in Section 5). The relative error matrix can be measured by either its spectral norm or the normalized Frobenius norm defined by

$$p^{-1/2}\|\mathbf{\Sigma}^{-1/2}\hat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1/2} - \mathbf{I}_p\|_{\mathrm{F}} = [p^{-1}\,\mathrm{tr}\{(\mathbf{\Sigma}^{-1/2}\hat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1/2} - \mathbf{I}_p)^2\}]^{1/2}. \tag{3.3}$$

In equality (3.3), there are p terms being added in the trace operation and the factor p^{-1} plays the role of normalization. The loss (3.3) is closely related to the entropy loss, which was introduced by James and Stein (1961). Also note that

$$p^{-1/2}\|\mathbf{\Sigma}^{-1/2}\hat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1/2} - \mathbf{I}_p\|_{\mathrm{F}} = \|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\Sigma}$$

where $\|\mathbf{A}\|_{\Sigma} = p^{-1/2}\|\mathbf{\Sigma}^{-1/2}\mathbf{A}\mathbf{\Sigma}^{-1/2}\|_{\mathrm{F}}$ is the weighted quadratic norm in Fan *et al.* (2008).

Fan *et al.* (2008) showed that, in a large factor model, the sample covariance is such that $\|\hat{\mathbf{\Sigma}}_{\mathrm{sam}} - \mathbf{\Sigma}\|_{\Sigma} = O_p\{\sqrt{(p/T)}\}$, which does not converge if $p > T$. In contrast, theorem 3 below shows that $\|\hat{\mathbf{\Sigma}}_{\hat{K}} - \mathbf{\Sigma}\|_{\Sigma}$ can still be convergent as long as $p = o(T^2)$. Technically, the effect of high dimensionality on the convergence rate of $\hat{\mathbf{\Sigma}}_{\hat{K}} - \mathbf{\Sigma}$ is via the number of rows in $\mathbf{B}$. We show in Appendix A that $\mathbf{B}$ appears in $\|\hat{\mathbf{\Sigma}}_{\hat{K}} - \mathbf{\Sigma}\|_{\Sigma}$ through $\mathbf{B}'\mathbf{\Sigma}^{-1}\mathbf{B}$ whose eigenvalues are bounded. Therefore it successfully cancels out the curse of high dimensionality that is introduced by $\mathbf{B}$.

Compared with estimating $\mathbf{\Sigma}$, in a large approximate factor model, we can estimate the precision matrix with a satisfactory rate under the spectral norm. The intuition follows from the fact that $\mathbf{\Sigma}^{-1}$ has bounded eigenvalues.

The following theorem summarizes the rate of convergence under various norms.

Theorem 3. Under the assumptions of theorem 2 the POET-estimator that is defined in equation (2.15) satisfies

$$\|\hat{\mathbf{\Sigma}}_{\hat{K}} - \mathbf{\Sigma}\|_{\Sigma} = O_p\left\{\frac{\sqrt{p}\log(p)}{T} + m_p\omega_T^{1-q}\right\},$$

$$\|\hat{\mathbf{\Sigma}}_{\hat{K}} - \mathbf{\Sigma}\|_{\max} = O_p(\omega_T).$$

In addition, if $m_p\omega_T^{1-q} = o(1)$, then $\hat{\mathbf{\Sigma}}_{\hat{K}}$ is non-singular with probability approaching 1, with

$$\|\hat{\mathbf{\Sigma}}_{\hat{K}}^{-1} - \mathbf{\Sigma}^{-1}\| = O_p(m_p\omega_T^{1-q}).$$

Remark 3.

(a) When estimating $\mathbf{\Sigma}^{-1}$, p is allowed to grow exponentially fast in T, and the estimator has the same rate of convergence as that of the estimator $\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{T}$ in theorem 2. When p becomes much larger than T, the precision matrix can be estimated at the same rate as if the factors were observable.

(b) As in remark 2, when $K > 0$ is known and grows with p and T, Fan *et al.* (2011a) prove the following results (when $q = 0$):

$$\|\hat{\mathbf{\Sigma}}^{\mathcal{T}} - \mathbf{\Sigma}\|_{\Sigma} = O_p\left[\frac{K\sqrt{p}\log(p)}{T} + K^2 m_p\sqrt{\left\{\frac{\log(p)}{T}\right\}} + \frac{m_p K^3}{\sqrt{p}}\right],$$

$$\|\hat{\mathbf{\Sigma}}^{\mathcal{T}} - \mathbf{\Sigma}\|_{\max} = O_p\left[K^3\sqrt{\left\{\frac{\log(p)}{T}\right\}} + \frac{K^3}{\sqrt{p}}\right],$$

$$\|(\hat{\mathbf{\Sigma}}^{\mathcal{T}})^{-1} - \mathbf{\Sigma}^{-1}\| = O_p\left[K^2 m_p\sqrt{\left\{\frac{\log(p)}{T}\right\}} + \frac{K^3 m_p}{\sqrt{p}}\right].$$

The results state explicitly the dependence of the rate of convergence on the number of factors. (The assumptions in Fan *et al.* (2011a) are slightly weaker than those presented here, in that they required that $\lambda_{\max}(\mathbf{\Sigma}_u)$ instead of $\|\mathbf{\Sigma}_u\|_1$ be bounded.)

(c) The relative error $\|\mathbf{\Sigma}^{-1/2}\hat{\mathbf{\Sigma}}_{\hat{K}}\mathbf{\Sigma}^{-1/2} - \mathbf{I}_p\|$ in operator norm can be shown to have the same order as the maximum relative error of estimated eigenvalues. It does not converge to 0 nor diverge. It is much smaller than $\|\hat{\mathbf{\Sigma}}_{\hat{K}} - \mathbf{\Sigma}\|$, which is of order $p/\sqrt{T}$ (see example 2).

3.4. Convergence of unknown factors and factor loadings

Many applications of the factor model require estimating the unknown factors. In general, factor loadings in $\mathbf{B}$ and the common factors $\mathbf{f}_t$ are not separably identifiable, as, for any matrix $\mathbf{H}$ such that $\mathbf{H}'\mathbf{H} = \mathbf{I}_K$, $\mathbf{B}\mathbf{f}_t = \mathbf{B}\mathbf{H}'\mathbf{H}\mathbf{f}_t$. Hence $(\mathbf{B}, \mathbf{f}_t)$ cannot be identified from $(\mathbf{B}\mathbf{H}', \mathbf{H}\mathbf{f}_t)$. Note that the linear space that is spanned by the rows of $\mathbf{B}$ is the same as that by those of $\mathbf{B}\mathbf{H}'$. In practice, it often does not matter which is used.

Let $\mathbf{V}$ denote the $\hat{K} \times \hat{K}$ diagonal matrix of the first $\hat{K}$ largest eigenvalues of the sample covariance matrix in decreasing order. Recall that $\mathbf{F}' = (\mathbf{f}_1, \ldots, \mathbf{f}_T)$ and define a $\hat{K} \times \hat{K}$ matrix $\mathbf{H} = (1/T)\mathbf{V}^{-1}\hat{\mathbf{F}}'\mathbf{F}\mathbf{B}'\mathbf{B}$. Then, for $t \leqslant T$, $\mathbf{H}\mathbf{f}_t = T^{-1}\mathbf{V}^{-1}\hat{\mathbf{F}}'(\mathbf{B}\mathbf{f}_1, \ldots, \mathbf{B}\mathbf{f}_T)'\mathbf{B}\mathbf{f}_t$. Note that $\mathbf{H}\mathbf{f}_t$ depends only on the data $\mathbf{V}^{-1}\hat{\mathbf{F}}'$ and an identifiable part of parameters $\{\mathbf{B}\mathbf{f}_t\}_{t=1}^T$. Therefore, there is no identifiability issue in $\mathbf{H}\mathbf{f}_t$ regardless of the identifiability condition imposed.

Bai (2003) obtained the rate of convergence for both $\hat{\mathbf{b}}_i$ and $\hat{\mathbf{f}}_t$ for any fixed (i, t). However, the uniform rate of convergence is more relevant for many applications (see example 3 in Section 5). The following theorem extends those results in Bai (2003) in a uniformity sense. In particular, with a more refined technique, we have improved the uniform convergence rate for $\hat{\mathbf{f}}_t$.

Theorem 4. Under the assumptions of theorem 2,

$$\max_{i \leqslant p} \|\hat{\mathbf{b}}_i - \mathbf{H}\mathbf{b}_i\| = O_p(\omega_T),$$

$$\max_{t \leqslant T} \|\hat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t\| = O_p\left(\frac{1}{T^{1/2}} + \frac{T^{1/4}}{\sqrt{p}}\right).$$

As a consequence of theorem 4, we obtain the following corollary (recall that the constant r_2 is defined in assumption 2).

Corollary 1. Under the assumptions of theorem 2,

$$\max_{i \leqslant p, t \leqslant T} \|\hat{\mathbf{b}}_i'\hat{\mathbf{f}}_t - \mathbf{b}_i'\mathbf{f}_t\| = O_p\left[\log(T)^{1/r_2}\sqrt{\left\{\frac{\log(p)}{T}\right\}} + \frac{T^{1/4}}{\sqrt{p}}\right].$$

The rates of convergence that were obtained above also explain the condition $T = o(p^2)$ in theorems 2 and 3. It is needed to estimate the common factors $\{\mathbf{f}_t\}_{t=1}^T$ uniformly in $t \leqslant T$. When we do not observe $\{\mathbf{f}_t\}_{t=1}^T$, in addition to the factor loadings, there are KT factors to estimate.

Intuitively, the condition $T = o(p^2)$ requires the number of parameters that are introduced by the unknown factors to be 'not too many', so that we can consistently estimate them uniformly. Technically, as demonstrated by Bickel and Levina (2008), Cai and Liu (2011) and many others, achieving uniform accuracy is essential for large covariance estimations.

4. Choice of threshold

4.1. *Finite sample positive definiteness*

Recall that the threshold value $\tau_{ij} = C\omega_T\sqrt{\hat{\theta}_{ij}}$, where C is determined by the users. To make POET operational in practice, we must choose C to maintain the positive definiteness of the estimated covariances for any given finite sample. We write $\hat{\mathbf{\Sigma}}^{\mathcal{T}}_{u,\hat{K}}(C) = \hat{\mathbf{\Sigma}}^{\mathcal{T}}_{u,\hat{K}}$, where the covariance estimator depends on C via the threshold. We choose C in the range where $\lambda_{\min}(\hat{\mathbf{\Sigma}}^{\mathcal{T}}_{u,\hat{K}}) > 0$. Define

$$C_{\min} = \inf[C > 0 : \lambda_{\min}\{\hat{\mathbf{\Sigma}}^{\mathcal{T}}_{u,\hat{K}}(M)\} > 0, \quad \forall M > C]. \tag{4.1}$$

When C is sufficiently large, the estimator becomes diagonal, whereas its minimum eigenvalue must retain strict positivity. Thus, $C_{\min}$ is well defined and, for all $C > C_{\min}$, $\hat{\mathbf{\Sigma}}^{\mathcal{T}}_{u,\hat{K}}(C)$ is positive definite under finite samples. We can obtain $C_{\min}$ by solving $\lambda_{\min}\{\hat{\mathbf{\Sigma}}^{\mathcal{T}}_{u,\hat{K}}(C)\} = 0, C \neq 0$. We can also approximate $C_{\min}$ by plotting $\lambda_{\min}\{\hat{\mathbf{\Sigma}}^{\mathcal{T}}_{u,\hat{K}}(C)\}$ as a function of C, as illustrated in Fig. 1. In practice, we can choose C in the range $(C_{\min} + \varepsilon, M)$ for a small ε and sufficiently large M. Choosing the threshold in a range to guarantee the finite sample positive definiteness has also been previously suggested by Fryzlewicz (2012).

4.2. *Multifold cross-validation*

In practice, C can be data driven, and chosen through multifold cross-validation. After obtaining the estimated residuals $\{\hat{\mathbf{u}}_t\}_{t \leqslant T}$ by PCA, we divide them randomly into two subsets, which are, for simplicity, denoted by $\{\hat{\mathbf{u}}_t\}_{t \in J_1}$ and $\{\hat{\mathbf{u}}_t\}_{t \in J_2}$. The sizes of J_1 and J_2, which are denoted by $T(J_1)$ and $T(J_2)$, are $T(J_1) \asymp T$ and $T(J_2) + T(J_1) = T$. For example, in sparse matrix estimation, Bickel and Levina (2008) suggested the choice $T(J_1) = T\{1 - \log(T)^{-1}\}$.

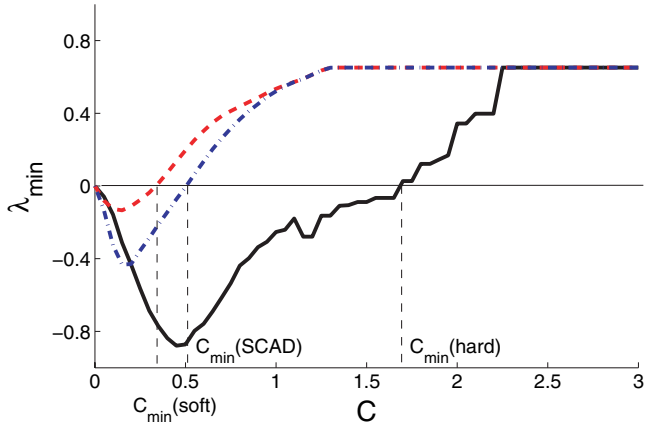

Fig. 1. Minimum eigenvalue of $\hat{\mathbf{\Sigma}}^{\mathcal{T}}_{u,\hat{K}}(C)$ as a function of C for three choices of thresholding rules (the plot is based on the simulated data set in Section 6.2): ———, hard thresholding; – – –, soft thresholding; · – · – ·, smoothly clipped absolute deviation

We repeat this procedure H times. At the jth split, we denote by $\hat{\mathbf{\Sigma}}_u^{\mathcal{T},j}(C)$ the POET-estimator with the threshold $C\omega_T\sqrt{\theta_{ij}}$ on the training data set $\{\hat{\mathbf{u}}_t\}_{t\in J_1}$. We also denote by $\hat{\mathbf{\Sigma}}_u^j$ the sample covariance based on the validation set, defined by $\hat{\mathbf{\Sigma}}_u^j = T(J_2)^{-1}\sum_{t\in J_2}\hat{\mathbf{u}}_t\hat{\mathbf{u}}_t'$. Then we choose the constant C^* by minimizing a cross-validation objective function over a compact interval

$$C^* = \arg\min_{C_{\min}+\varepsilon\leqslant C\leqslant M}\frac{1}{H}\sum_{j=1}^{H}\|\hat{\mathbf{\Sigma}}_u^{\mathcal{T},j}(C)-\hat{\mathbf{\Sigma}}_u^j\|_{\mathrm{F}}^2. \tag{4.2}$$

Here $C_{\min}$ is the minimum constant that guarantees the positive definiteness of $\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}}(C)$ for $C > C_{\min}$ as described in the previous subsection, and M is a large constant such that $\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}}(M)$ is diagonal. The resulting C^* is data driven, so it depends on $\mathbf{Y}$ as well as p and T via the data. In contrast, for each given $N\times T$ data matrix $\mathbf{Y}$, C^* is a universal constant in the threshold $\tau_{ij} = C^*\omega_T\sqrt{\hat{\theta}_{ij}}$ in the sense that it does not change with respect to the position (i, j). We also note that the cross-validation is based on the estimate of $\mathbf{\Sigma}_u$ rather than $\mathbf{\Sigma}$ because POET thresholds the error covariance matrix. Thus cross-validation improves the performance of thresholding.

It is possible to derive the rate of convergence for $\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}}(C^*)$ under the current model setting, but it ought to be much more technically involved than the regular sparse matrix estimation that was considered by Bickel and Levina (2008) and Cai and Liu (2011). To keep our presentation simple we do not pursue it in the current paper.

5. Applications of POET

We give four examples to which the results in theorems 2–4 can be applied. Detailed pursuits of these are beyond the scope of the paper.

5.1. Example 3 (large-scale hypothesis testing)

Controlling the false discovery rate in large-scale hypothesis testing based on correlated test statistics is an important and challenging problem in statistics (Leek and Storey, 2008; Efron, 2010; Fan *et al.*, 2012). Suppose that the test statistic for each of the hypotheses

$$H_{i0}:\mu_i = 0 \quad versus \quad H_{i1}:\mu_i \neq 0$$

is $Z_i \sim N(\mu_i, 1)$ and these test statistics $\mathbf{Z}$ are jointly normal $N(\boldsymbol{\mu}, \mathbf{\Sigma})$ where $\mathbf{\Sigma}$ is unknown. For a given critical value x, the false discovery proportion is then defined as $\mathrm{FDP}(x) = V(x)/R(x)$ where $V(x) = p^{-1}\sum_{\mu_i=0} I(|Z_i| > x)$ and $R(x) = p^{-1}\sum_{i=1}^{p} I(|Z_i| > x)$ are the total number of false discoveries and the total number of discoveries respectively. Our interest is to estimate $\mathrm{FDP}(x)$ for each given x. Note that $R(x)$ is an observable quantity. Only $V(x)$ needs to be estimated.

If the covariance $\mathbf{\Sigma}$ admits the approximate factor structure (1.3), then the test statistics can be stochastically decomposed as

$$\mathbf{Z} = \boldsymbol{\mu} + \mathbf{Bf} + \mathbf{u}, \tag{5.1}$$

where $\mathbf{\Sigma}_u$ is sparse. By the principal factor approximation (theorem 1, Fan *et al.* (2012))

$$V(x) = \sum_{i=1}^{p}[\Phi\{a_i(z_{x/2}+\eta_i)\}+\Phi\{a_i(z_{x/2}-\eta_i)\}]+o_P(p), \tag{5.2}$$

when $m_p = o(p)$ and the number of true significant hypotheses $\{i:\mu_i\neq 0\}$ is $o(p)$, where z_x is the upper x-quantile of the standard normal distribution, $\eta_i = (\mathbf{Bf})_i$ and $a_i = \mathrm{var}(u_i)^{-1}$.

Now suppose that we have n repeated measurements from model (5.1). Then, by corollary 1, $\{\eta_i\}$ can be uniformly consistently estimated, and hence $p^{-1}V(x)$ and $\mathrm{FDP}(x)$ can be consistently estimated. Efron (2010) obtained these repeated test statistics on the basis of the bootstrap sample from the original raw data. Our theory (theorem 4) gives a formal justification to the framework of Efron (2007, 2010).

5.2. Example 4 (risk management)

The maximum elementwise estimation error $\|\hat{\mathbf{\Sigma}}_{\hat{K}} - \mathbf{\Sigma}\|_{\max}$ appears in risk assessment as in Fan *et al.* (2012). For a fixed portfolio allocation vector $\mathbf{w}$, the true portfolio variance and the estimated variance are given by $\mathbf{w}'\mathbf{\Sigma}\mathbf{w}$ and $\mathbf{w}'\hat{\mathbf{\Sigma}}_{\hat{K}}\mathbf{w}$ respectively. The estimation error is bounded by

$$|\mathbf{w}'\hat{\mathbf{\Sigma}}_{\hat{K}}\mathbf{w} - \mathbf{w}'\mathbf{\Sigma}\mathbf{w}| \leqslant \|\hat{\mathbf{\Sigma}}_{\hat{K}} - \mathbf{\Sigma}\|_{\max}\|\mathbf{w}\|_1^2,$$

where $\|\mathbf{w}\|_1$, the L_1-norm of $\mathbf{w}$, is the gross exposure of the portfolio. Usually a constraint is placed on the total percentage of the short positions, in which case we have a restriction $\|\mathbf{w}\|_1 \leqslant c$ for some $c > 0$. In particular, $c = 1$ corresponds to a portfolio with no short positions (all weights are non-negative). Theorem 3 quantifies the maximum approximation error.

The above discussion compares the absolute error of perceived risk and true risk. The relative error is bounded by

$$|\mathbf{w}'\hat{\mathbf{\Sigma}}_{\hat{K}}\mathbf{w}/\mathbf{w}'\mathbf{\Sigma}\mathbf{w} - 1| \leqslant \|\mathbf{\Sigma}^{-1/2}\hat{\mathbf{\Sigma}}_{\hat{K}}\mathbf{\Sigma}^{-1/2} - \mathbf{I}_p\|$$

for any allocation vector $\mathbf{w}$. Theorem 3 quantifies this relative error.

5.3. Example 5 (panel regression with a factor structure in the errors)

Consider the panel regression model

$$Y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \qquad \varepsilon_{it} = \mathbf{b}_i'\mathbf{f}_t + u_{it}, \quad i \leqslant p, t \leqslant T,$$

where $\mathbf{x}_{it}$ is a vector of observable regressors with fixed dimension. The regression error ε_{it} has a factor structure and is assumed to be independent of $\mathbf{x}_{it}$, but $\mathbf{b}_i$, $\mathbf{f}_t$ and u_{it} are all unobservable. We are interested in the common regression coefficients $\boldsymbol{\beta}$. This panel regression model has been considered by many researchers, such as Ahn *et al.* (2001) and Pesaran (2006), and has broad applications in social sciences.

Although ordinary least squares produces a consistent estimator of $\boldsymbol{\beta}$, a more efficient estimation can be obtained by generalized least squares. The generalized least squares method depends, however, on an estimator of $\mathbf{\Sigma}_\varepsilon^{-1}$, which is the inverse of the covariance matrix of $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, \ldots, \varepsilon_{pt})'$. By assuming that the covariance matrix of $(u_{1t}, \ldots, u_{pt})$ is sparse, we can successfully solve this problem by applying theorem 3. Although ε_{it} is unobservable, it can be replaced by the regression residuals $\hat{\varepsilon}_{it}$, obtained via first regressing Y_{it} on $\mathbf{x}_{it}$. We then apply POET to $T^{-1}\sum_{t=1}^T \hat{\boldsymbol{\varepsilon}}_t\hat{\boldsymbol{\varepsilon}}_t'$. By theorem 3, the inverse of the resulting estimator is a consistent estimator of $\mathbf{\Sigma}_\varepsilon^{-1}$ under the spectral norm. A slight difference lies in the fact that, when we apply POET, $T^{-1}\sum_{t=1}^T \boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'$ is replaced with $T^{-1}\sum_{t=1}^T \hat{\boldsymbol{\varepsilon}}_t\hat{\boldsymbol{\varepsilon}}_t'$, which introduces an additional term $O_p[\sqrt{\{\log(p)/T\}}]$ in the estimation error.

5.4. Example 6 (validating an asset pricing theory)

A celebrated financial economic theory is the capital asset pricing model (Sharpe, 1964) that helped William Sharpe to win the Nobel prize in economics in 1990, whose extension is the

multifactor model (Ross, 1976; Chamberlain and Rothschild, 1983). It states that, in a frictionless market, the excessive return of any financial asset equals the excessive returns of the risk factors times its factor loadings plus noise. In the multiperiod model, the excess return y_{it} of firm i at time t follows model (1.1), in which $\mathbf{f}_t$ are the excess returns of the risk factors at time t. To test the null hypothesis (1.2), we embed the model into the multivariate linear model

$$\mathbf{y}_t = \boldsymbol{\alpha} + \mathbf{B}\mathbf{f}_t + \mathbf{u}_t, \qquad t = 1, \ldots, T, \tag{5.3}$$

and wish to test $H_0 : \boldsymbol{\alpha} = 0$. The F-test statistic involves the estimation of the covariance matrix $\boldsymbol{\Sigma}_u$, whose estimates are degenerate without regularization when $p \geqslant T$. Therefore, in the literature (Sentana (2009), and references therein), we focus on the case that p is relatively small. The typical choices of parameters are $T = 60$ monthly data and the number of assets $p = 5$, 10, 25. However, the capital asset pricing model should hold for all tradeable assets, not just a small fraction of assets. With our regularization technique, non-degenerate estimate $\hat{\boldsymbol{\Sigma}}_{u,\hat{K}}^{\mathcal{T}}$ can be obtained and the F-test or likelihood ratio test statistics can be employed even when $p \gg T$.

To provide some insights, let $\hat{\boldsymbol{\alpha}}$ be the least squares estimator of model (5.3). Then, when $\mathbf{u}_t \sim N(0, \boldsymbol{\Sigma}_u)$, $\hat{\boldsymbol{\alpha}} \sim N(\boldsymbol{\alpha}, \boldsymbol{\Sigma}_u / c_T)$ for a constant c_T which depends on the observed factors. When $\boldsymbol{\Sigma}_u$ is known, the Wald test statistic is $W = c_T \hat{\boldsymbol{\alpha}}' \boldsymbol{\Sigma}_u^{-1} \hat{\boldsymbol{\alpha}}$. When it is unknown and p is large, it is natural to use the F-type of test statistic $\hat{W} = c_T \hat{\boldsymbol{\alpha}}' (\hat{\boldsymbol{\Sigma}}_{u,\hat{K}}^{\mathcal{T}})^{-1} \hat{\boldsymbol{\alpha}}$. The difference between these two statistics is bounded by

$$|\hat{W} - W| \leqslant c_T \|(\hat{\boldsymbol{\Sigma}}_{u,\hat{K}}^{\mathcal{T}})^{-1} - \boldsymbol{\Sigma}_u^{-1}\| \|\hat{\boldsymbol{\alpha}}\|^2.$$

Since under the null hypothesis $\hat{\boldsymbol{\alpha}} \sim N(0, \boldsymbol{\Sigma}_u / c_T)$, we have $c_T \|\boldsymbol{\Sigma}_u^{-1/2} \hat{\boldsymbol{\alpha}}\|^2 = O(p)$. Thus, it follows from boundness of $\|\boldsymbol{\Sigma}_u\|$ that $|\hat{W} - W| = O(p) \|(\hat{\boldsymbol{\Sigma}}_{u,\hat{K}}^{\mathcal{T}})^{-1} - \boldsymbol{\Sigma}_u^{-1}\|$. Theorem 2 provides the rate of convergence for this difference. Detailed development is out of the scope of the current paper, and we shall leave it as a separate research project (see Pesaran and Yamagata (2012)).

6. Monte Carlo experiments

In this section, we shall examine the performance of POET in a finite sample. We shall also demonstrate the effect of this estimator on asset allocation and risk assessment. Similarly to Fan *et al.* (2008, 2011a), we simulated from a standard Fama–French three-factor model, assuming a sparse error covariance matrix and three factors. Throughout this section, the timespan is fixed at $T = 300$, and the dimensionality p increases from 1 to 600. We assume that the excess returns of each of p stocks over the risk-free interest rate follow the model

$$y_{it} = b_{i1} f_{1t} + b_{i2} f_{2t} + b_{i3} f_{3t} + u_{it}.$$

The factor loadings are drawn from a trivariate normal distribution $\mathbf{b} \sim N_3(\boldsymbol{\mu}_B, \boldsymbol{\Sigma}_B)$ and the idiosyncratic errors from $\mathbf{u}_t \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}_u)$, and the factor returns $\mathbf{f}_t$ follow a vector auto-regressive VAR(1) model. To make the simulation more realistic, model parameters are calibrated from the financial returns, as detailed in the following section.

6.1. Calibration

To calibrate the model, we use the data on annualized returns of 100 industrial portfolios from the Web site of Kenneth French, and the data on 3-month Treasury bill rates from the Center for Research in Security Prices database. These industrial portfolios are formed as the intersection of 10 portfolios based on size (market equity) and 10 portfolios based on the book equity to market equity ratio. Their excess returns $\tilde{\mathbf{y}}_t$ are computed for the period from

January 1st, 2009, to December 31st, 2010. Here, we present a short outline of the calibration procedure.

(a) Given $\{\tilde{\mathbf{y}}_t\}_{t=1}^{500}$ as the input data, we fit a Fama–French three-factor model and calculate a 100×3 matrix $\tilde{\mathbf{B}}$, and 500×3 matrix $\tilde{\mathbf{F}}$, using the principal components method that was described in Section 3.1.

(b) We summarize 100 factor loadings (the rows of $\tilde{\mathbf{B}}$) by their sample mean vector $\boldsymbol{\mu}_B$ and sample covariance matrix $\boldsymbol{\Sigma}_B$, which are reported in Table 1. The factor loadings $\mathbf{b}_i = (b_{i1}, b_{i2}, b_{i3})^{\mathrm{T}}$ for $i = 1, \ldots, p$ are drawn from $N_3(\boldsymbol{\mu}_B, \boldsymbol{\Sigma}_B)$.

(c) We run the stationary vector auto-regressive model $\mathbf{f}_t = \boldsymbol{\mu} + \boldsymbol{\Phi}\mathbf{f}_{t-1} + \boldsymbol{\varepsilon}_t$, which is a VAR(1) model, to the data $\tilde{\mathbf{F}}$ to obtain the multivariate least squares estimator for $\boldsymbol{\mu}$ and $\boldsymbol{\Phi}$, and we estimate $\boldsymbol{\Sigma}_\varepsilon$. Note that all eigenvalues of $\boldsymbol{\Phi}$ in Table 2 fall within the unit circle, so our model is stationary. The covariance matrix $\mathrm{cov}(\mathbf{f}_t)$ can be obtained by solving the linear equation $\mathrm{cov}(\mathbf{f}_t) = \boldsymbol{\Phi}\,\mathrm{cov}(\mathbf{f}_t)\boldsymbol{\Phi}' + \boldsymbol{\Sigma}_\varepsilon$. The estimated parameters are depicted in Table 2 and are used to generate $\mathbf{f}_t$.

(d) For each value of p, we generate a sparse covariance matrix $\boldsymbol{\Sigma}_u$ of the form

$$\boldsymbol{\Sigma}_u = \mathbf{D}\boldsymbol{\Sigma}_0\mathbf{D}.$$

Here, $\boldsymbol{\Sigma}_0$ is the error correlation matrix, and $\mathbf{D}$ is the diagonal matrix of the standard deviations of the errors. We set $\mathbf{D} = \mathrm{diag}(\sigma_1, \ldots, \sigma_p)$, where each σ_i is generated independently from a gamma distribution $G(\alpha, \beta)$, and α and β are chosen to match the sample mean and sample standard deviation of the standard deviations of the errors. A similar approach to that of Fan *et al.* (2011a) has been used in this calibration step. The off-diagonal entries of $\boldsymbol{\Sigma}_0$ are generated independently from a normal distribution, with mean and standard deviation equal to the sample mean and sample standard deviation of the sample correlations between the estimated residuals, conditional on their absolute values being no larger than 0.95. We then employ hard thresholding to make $\boldsymbol{\Sigma}_0$ sparse, where the threshold is found as the smallest constant that provides the positive definiteness of $\boldsymbol{\Sigma}_0$. More precisely, start with threshold value 1, which gives $\boldsymbol{\Sigma}_0 = \mathbf{I}_p$, and then decrease the threshold values in a grid until positive definiteness is violated.

Table 1. Mean and covariance matrix used to generate **b**

$\boldsymbol{\mu}_B$	$\boldsymbol{\Sigma}_B$		
0.0047	0.0767	−0.00004	0.0087
0.0007	−0.00004	0.0841	0.0013
−1.8078	0.0087	0.0013	0.1649

Table 2. Parameters of the $\mathbf{f}_t$-generating process

$\boldsymbol{\mu}$	$cov(\mathbf{f}_t)$		$\boldsymbol{\Phi}$			
−0.0050	1.0037	0.0011	−0.0009	−0.0712	0.0468	0.1413
0.0335	0.0011	0.9999	0.0042	−0.0764	−0.0008	0.0646
−0.0756	−0.0009	0.0042	0.9973	0.0195	−0.0071	−0.0544

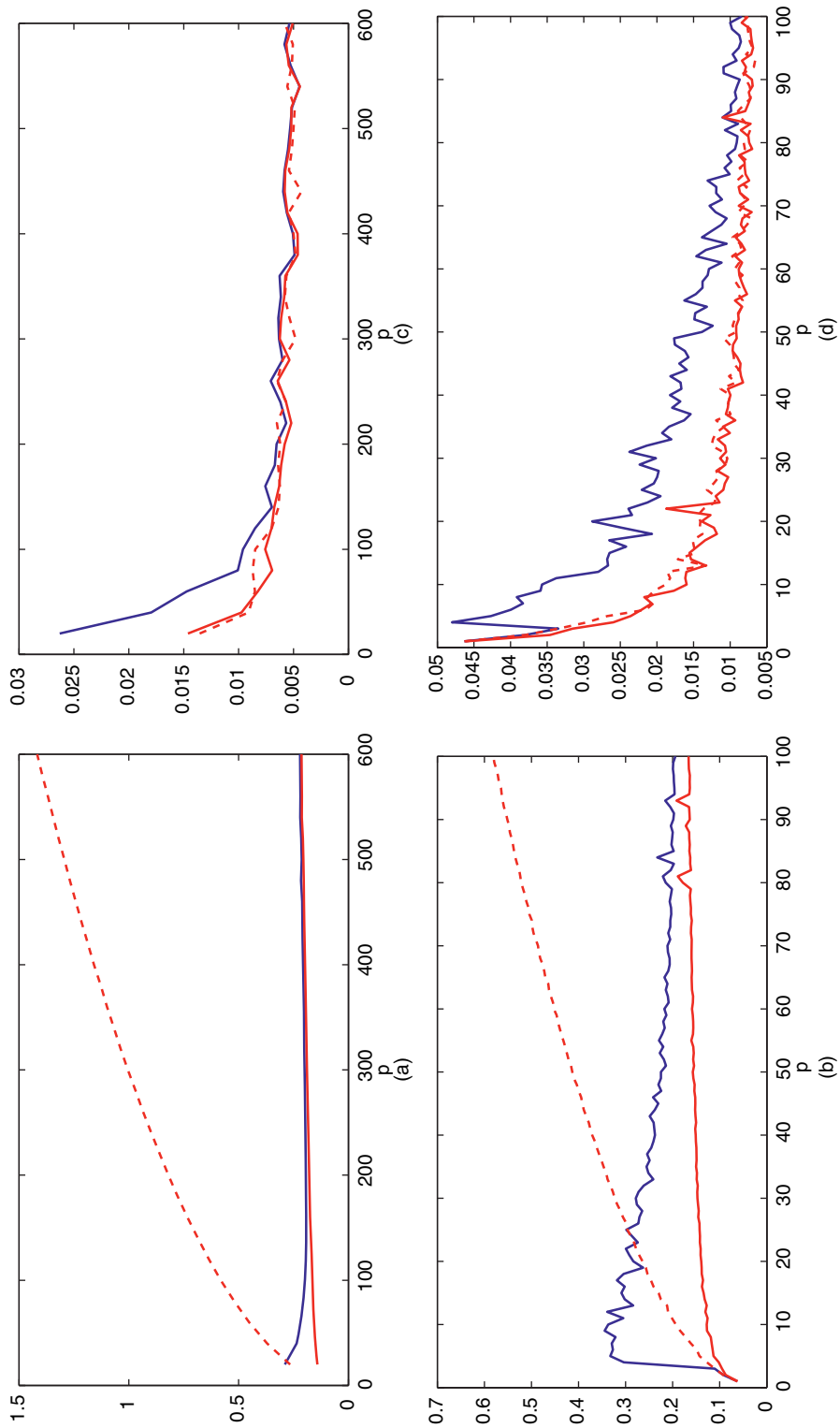

Fig. 2. (a), (b) Averages and (c), (d) standard deviations of the relative error $p^{-1/2}\|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - \mathbf{I}_p\|_F$ with known factors (——, $\hat{\Sigma} = \hat{\Sigma}_{\text{obs}}$), POET (——, $\hat{\Sigma} = \hat{\Sigma}_K$) and the sample covariance (- - - - -, $\hat{\Sigma} = \hat{\Sigma}_{\text{sam}}$) over 200 simulations, as a function of the dimensionality p: (a), (c) p ranges in 20–600 with increment 20; (b), (d) p ranges in 1–100 with increment 1

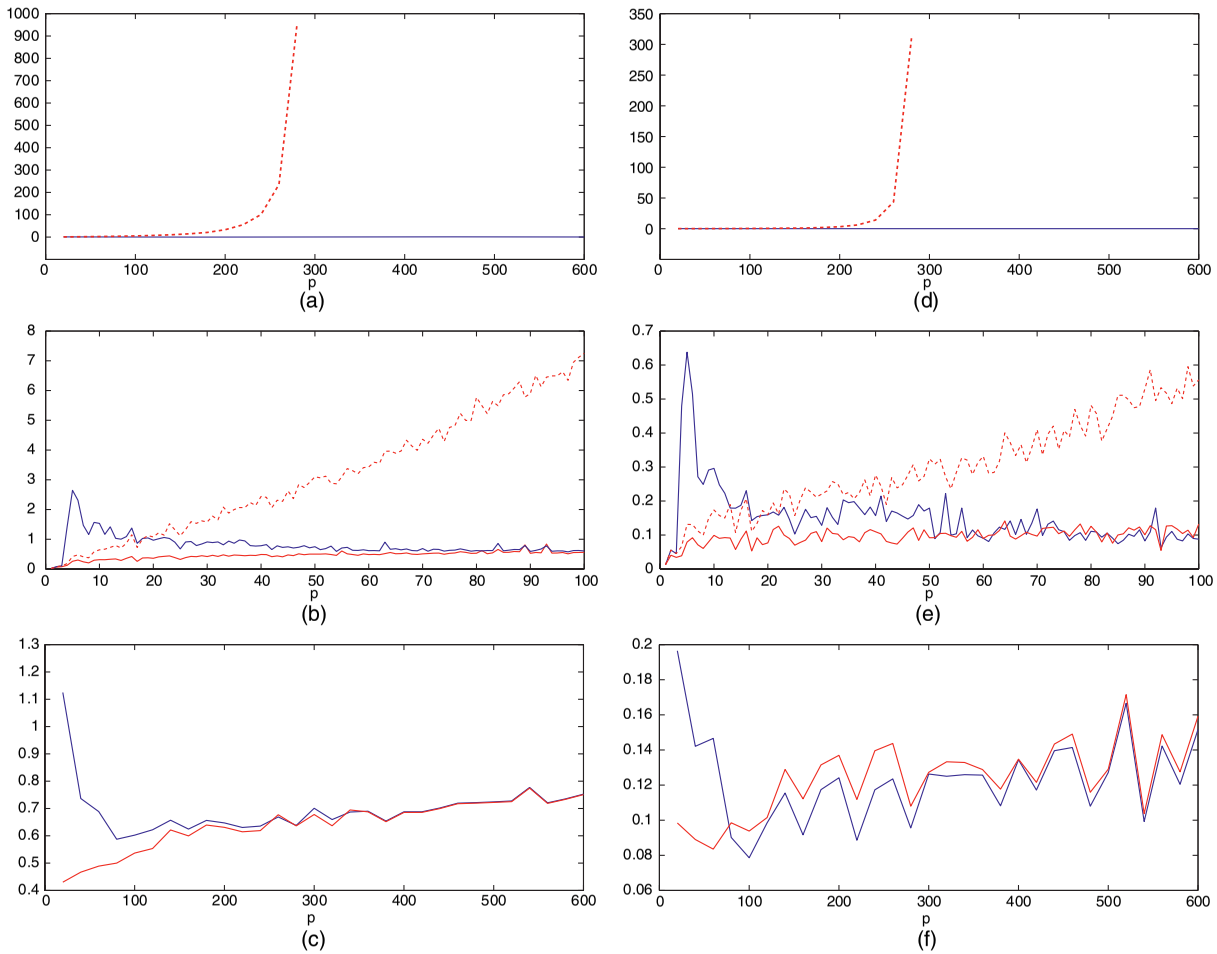

Fig. 3. (a)–(c) Averages and (d)–(f) standard deviations of $\|\hat{\Sigma}^{-1} - \Sigma^{-1}\|$ with known factors (——, $\hat{\Sigma} = \hat{\Sigma}_{\text{obs}}$), POET (——, $\hat{\Sigma} = \hat{\Sigma}_{\hat{K}}$) and sample covariance (------, $\hat{\Sigma} = \hat{\Sigma}_{\text{sam}}$) over 200 simulations, as a function of the dimensionality p: (a), (d) p ranges in 20–600 with increment 20; (b), (e) p ranges in 1–100 with increment 1; (c), (f) the same as (a) and (d) with the sample covariance curve omitted

6.2. Simulation

For the simulation, we fix $T = 300$, and let p increase from 1 to 600. For each fixed p, we repeat the following steps $N = 200$ times, and record the means and the standard deviations of each respective norm.

Step 1: generate independently $\{\mathbf{b}_i\}_{i=1}^{p} \sim N_3(\boldsymbol{\mu}_B, \boldsymbol{\Sigma}_B)$, and set $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_p)'$.
Step 2: generate independently $\{\mathbf{u}_t\}_{t=1}^{T} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}_u)$.
Step 3: generate $\{\mathbf{f}_t\}_{t=1}^{T}$ as a vector auto-regressive sequence of the form $\mathbf{f}_t = \boldsymbol{\mu} + \Phi \mathbf{f}_{t-1} + \boldsymbol{\varepsilon}_t$.
Step 4: calculate $\{\mathbf{y}_t\}_{t=1}^{T}$ from $\mathbf{y}_t = \mathbf{B}\mathbf{f}_t + \mathbf{u}_t$.
Step 5: set hard thresholding with threshold $0.5\sqrt{\hat{\theta}_{ij}}[\sqrt{\{\log(p)/T\}} + 1/\sqrt{p}]$. Estimate K by using IC1 of Bai and Ng (2002). Calculate covariance estimators by using POET. Calculate the sample covariance matrix $\hat{\boldsymbol{\Sigma}}_{\text{sam}}$.

In the graphs below, we plot the averages and standard deviations of the distance from $\hat{\boldsymbol{\Sigma}}_{\hat{K}}$ and $\hat{\boldsymbol{\Sigma}}_{\text{sam}}$ to the true covariance matrix $\boldsymbol{\Sigma}$, under norms $\|\cdot\|_{\Sigma}$, $\|\cdot\|$ and $\|\cdot\|_{\max}$. We also plot the means and standard deviations of the distances from $(\hat{\boldsymbol{\Sigma}}_{\hat{K}})^{-1}$ and $\hat{\boldsymbol{\Sigma}}_{\text{sam}}^{-1}$ to $\boldsymbol{\Sigma}^{-1}$ under the spectral norm. The dimensionality p ranges from 20 to 600 in increments of 20. Because of invertibility, the spectral norm for $\hat{\boldsymbol{\Sigma}}_{\text{sam}}^{-1}$ is plotted only up to $p = 280$. Also, we zoom into these graphs by plotting the values of p from 1 to 100, this time in increments of 1. Note that we also plot the distance from $\hat{\boldsymbol{\Sigma}}_{\text{obs}}$ to $\boldsymbol{\Sigma}$ for comparison, where $\hat{\boldsymbol{\Sigma}}_{\text{obs}}$ is the estimated covariance matrix that was proposed by Fan *et al.* (2011a), assuming that the factors are observable.

6.3. Results

In a factor model, we expect POET to perform as well as $\hat{\boldsymbol{\Sigma}}_{\text{obs}}$ when p is relatively large, since the effect of estimating the unknown factors should vanish as p increases. This is illustrated in the plots.

From the simulation results, reported in Figs 2–5, we observe that POET under the unobservable factor model performs just as well as the estimator in Fan *et al.* (2011a) if the factors are known, when p is sufficiently large. The cost of not knowing the factors is approximately of order $O_p(1/\sqrt{p})$. It can be seen in Figs 2 and 3 that this cost vanishes for $p \geqslant 200$. To give a better insight of the effect of estimating the unknown factors for small p, a separate set of simulations is conducted for $p \leqslant 100$. As we can see from Figs 2(b) and 2(d) and 3(b), 3(c), 3(e) and 3(f) the effect decreases quickly. In addition, when estimating $\boldsymbol{\Sigma}^{-1}$, it is difficult to distinguish the estimators with known and unknown factors, whose performances are quite stable compared with the sample covariance matrix. Also, the maximum absolute elementwise error (Fig. 4) of our estimator performs very similarly to that of the sample covariance matrix, which coincides with our asymptotic result. Fig. 5 shows that the performances of the three methods are indistinguishable in the spectral norm, as expected.

6.4. Robustness to the estimation of K

POET depends on the estimated number of factors. Our theory uses a consistent esimator $\hat{K}$. To assess the robustness of our procedure to $\hat{K}$ in finite samples, we calculate $\hat{\boldsymbol{\Sigma}}_{u,K}^{\mathcal{T}}$ for $K = 1, 2, \ldots, 10$. Again, the threshold is fixed to be $0.5\sqrt{\hat{\theta}_{ij}}[\sqrt{\{\log(p)/T\}} + 1/\sqrt{p}]$.

6.4.1. Design 1

The simulation set-up is the same as before where the true $K_0 = 3$. We calculate $\|\hat{\boldsymbol{\Sigma}}_{u,K}^{\mathcal{T}} - \boldsymbol{\Sigma}_u\|$, $\|(\hat{\boldsymbol{\Sigma}}_{u,K}^{\mathcal{T}})^{-1} - \boldsymbol{\Sigma}_u^{-1}\|$, $\|\hat{\boldsymbol{\Sigma}}_K^{-1} - \boldsymbol{\Sigma}^{-1}\|$ and $\|\hat{\boldsymbol{\Sigma}}_K - \boldsymbol{\Sigma}\|_{\Sigma}$ for $K = 1, 2, \ldots, 10$. Fig. 6 plots these norms

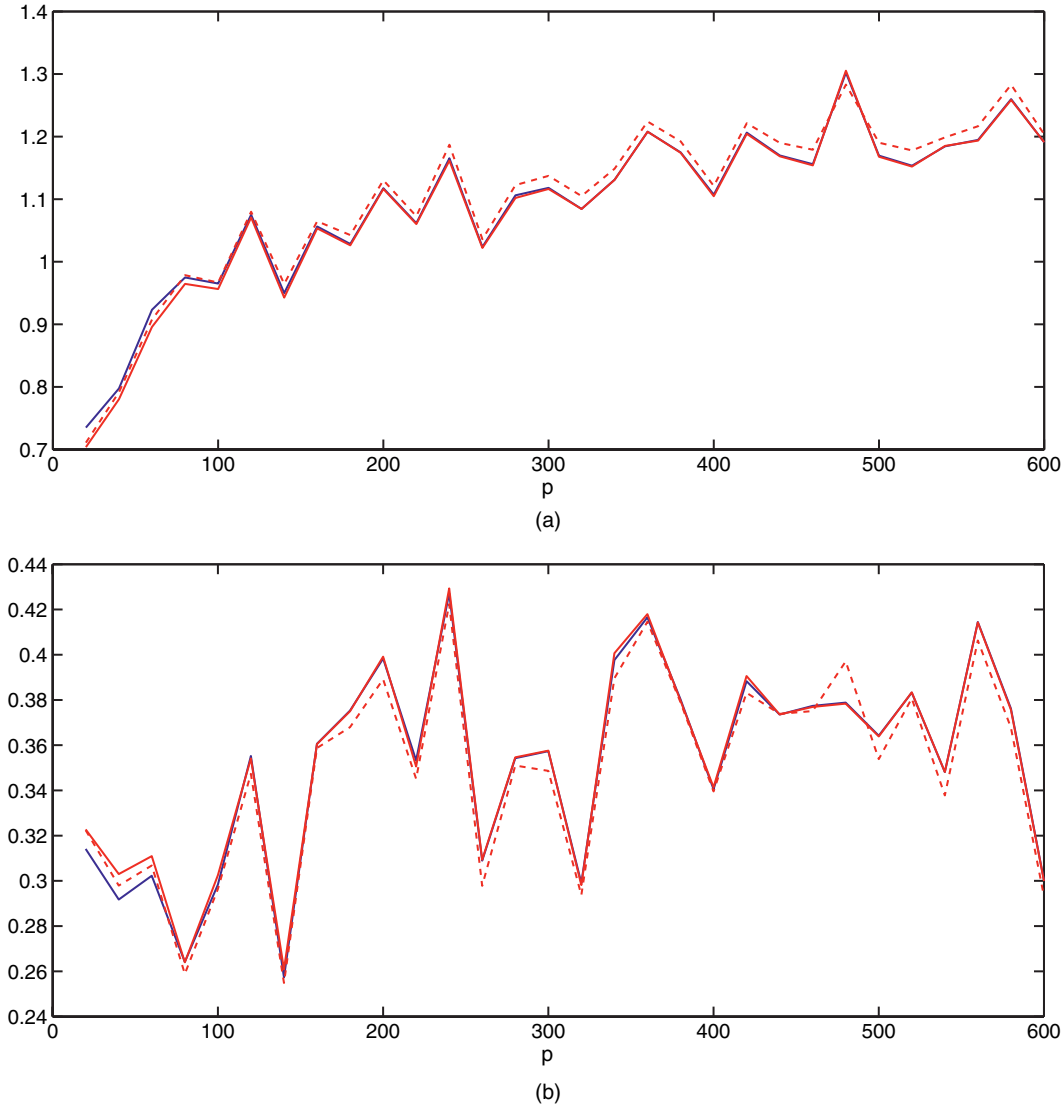

Fig. 4. (a) Averages and (b) standard deviations of $\|\hat{\Sigma} - \Sigma\|_{\max}$ with known factors (——, $\hat{\Sigma} = \hat{\Sigma}_{\text{obs}}$), POET (——, $\hat{\Sigma} = \hat{\Sigma}_{\hat{K}}$) and sample covariance (------, $\hat{\Sigma} = \hat{\Sigma}_{\text{sam}}$) over 200 simulations, as a function of the dimensionality p: they are nearly undifferentiable

as p increases but with a fixed $T = 300$. The results demonstrate a trend that is quite robust when $K \geqslant 3$; especially, the accuracy of estimation of the spectral norms for large p are close to each other. When $K = 1$ or $K = 2$, the estimators perform badly because of modelling bias. Therefore, POET is robust to overestimated K, but not to underestimation.

6.4.2. Design 2

We also simulated from a new data-generating process for the robustness assessment. Consider a banded idiosyncratic matrix

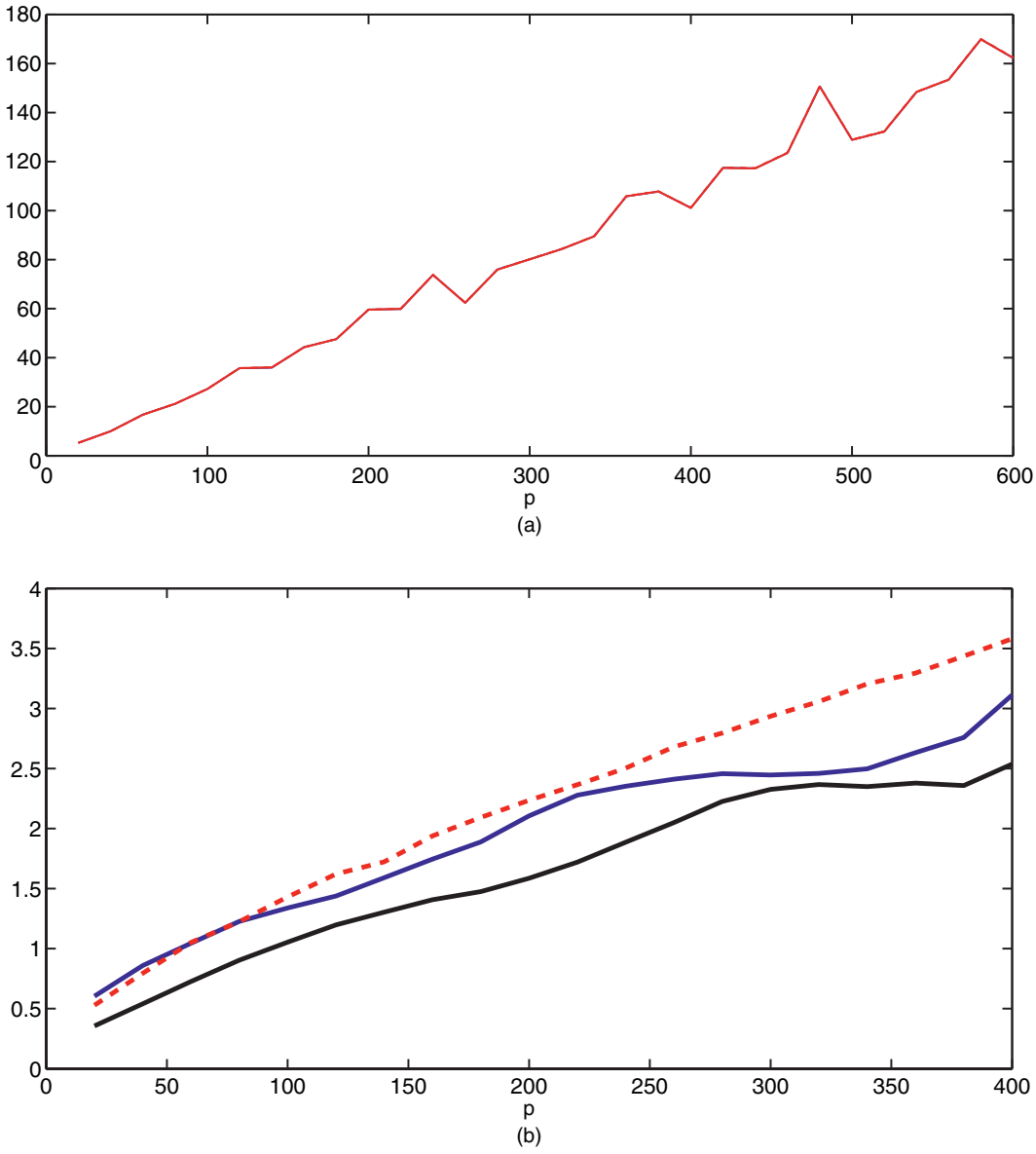

Fig. 5. (a) Averages of $\|\hat{\Sigma} - \Sigma\|$ and (b) $\|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - \mathbf{I}_p\|$ with known factors (——, $\hat{\Sigma} = \hat{\Sigma}_{\text{obs}}$), POET (——, $\hat{\Sigma} = \hat{\Sigma}_{\hat{K}}$) and sample covariance (------, $\hat{\Sigma} = \hat{\Sigma}_{\text{sam}}$) over 200 simulations, as a function of the dimensionality p: the three curves are barely distinguishable in (a)

$$\sigma_{u,ij} = \begin{cases} 0.5^{|i-j|}, & |i-j| \leqslant 9, \\ 0, & |i-j| > 9. \end{cases} \quad (\mathbf{u}_1, \ldots, \mathbf{u}_T) \overset{\text{IID}}{\sim} N_p(0, \boldsymbol{\Sigma}_u).$$

We still consider a $K_0 = 3$ factor model, where the factors are independently simulated as

$$f_{it} \sim N(0,1), \qquad b_{ji} \sim N(0,1), \qquad i \leqslant 3, j \leqslant p, t \leqslant T.$$

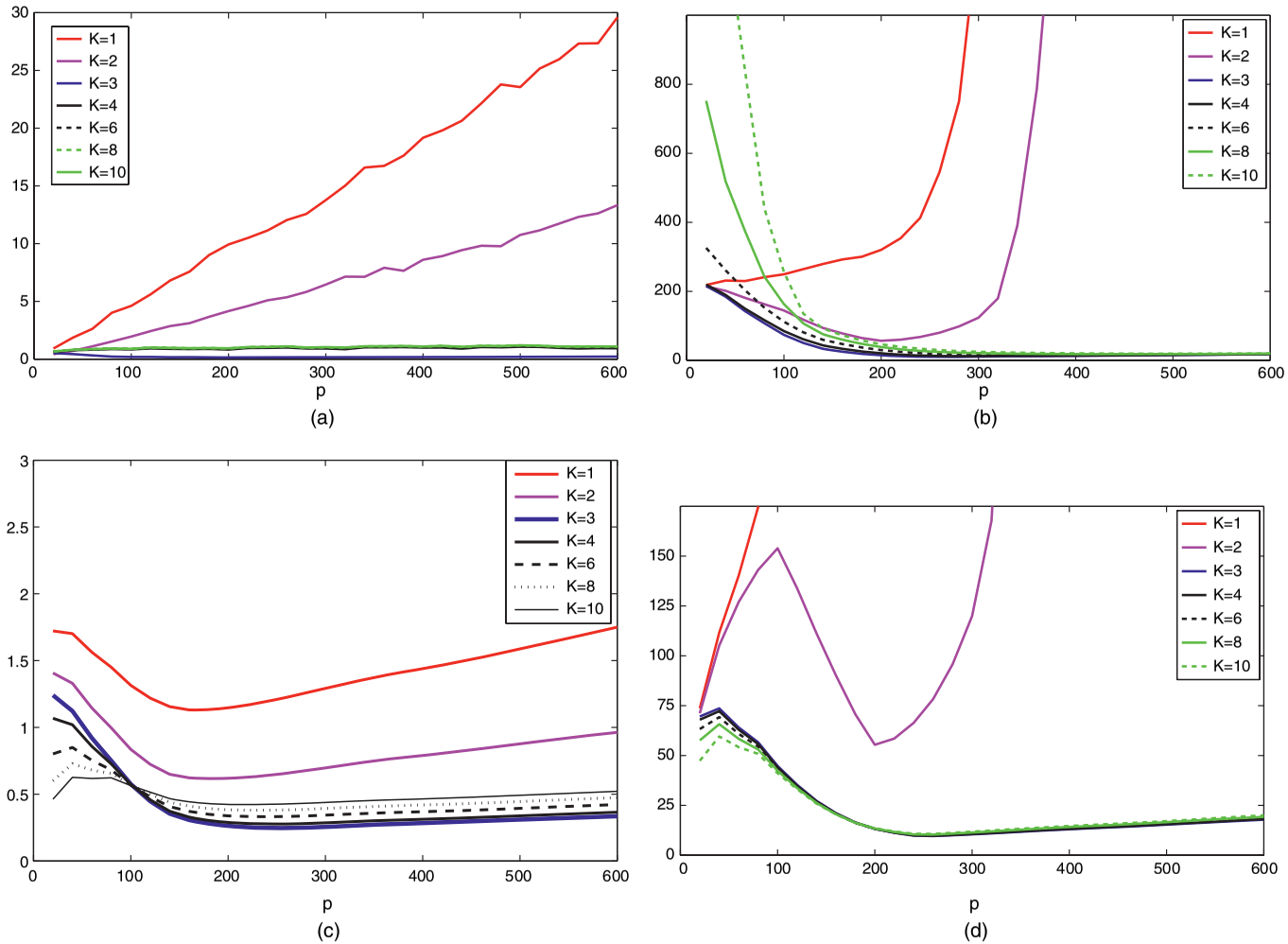

Fig. 6. Robustness of K as p increases for various choices of K (design 1, $T = 300$): (a) $\|\hat{\Sigma}_{u,K}^{\mathcal{T}} - \Sigma_u\|$; (b) $\|(\hat{\Sigma}_{u,K}^{\mathcal{T}})^{-1} - \Sigma_u^{-1}\|$; (c) $\|\hat{\Sigma}_K - \Sigma\|_u$; (d) $\|\hat{\Sigma}_K - \Sigma^{-1}\|$

Table 3. Robustness of K: design 2, estimation errors in spectral norm†

	Errors for the following values of K:						
	1	*2*	*3*	*4*	*5*	*6*	*8*
$p=100$							
$\hat{\mathbf{\Sigma}}_{u,K}^{\mathcal{T}}$	10.70	5.23	1.63	1.80	1.91	2.04	2.22
$(\hat{\mathbf{\Sigma}}_{u,K}^{\mathcal{T}})^{-1}$	2.71	2.51	1.51	1.50	1.44	1.84	2.82
$\hat{\mathbf{\Sigma}}_K^{-1}$	2.69	2.48	1.47	1.49	1.41	1.56	2.35
$\hat{\mathbf{\Sigma}}_K$	94.66	91.36	29.41	31.45	30.91	33.59	33.48
$\mathbf{\Sigma}^{-1/2}\hat{\mathbf{\Sigma}}_K\mathbf{\Sigma}^{-1/2}$	17.37	10.04	2.05	2.83	2.94	2.95	2.93
$p=200$							
$\hat{\mathbf{\Sigma}}_{u,K}^{\mathcal{T}}$	11.34	11.45	1.64	1.71	1.79	1.87	2.01
$(\hat{\mathbf{\Sigma}}_{u,K}^{\mathcal{T}})^{-1}$	2.69	3.91	1.57	1.56	1.81	2.26	3.42
$\hat{\mathbf{\Sigma}}_K^{-1}$	2.67	3.72	1.57	1.55	1.70	2.13	3.19
$\hat{\mathbf{\Sigma}}_K$	200.82	195.64	57.44	63.09	64.53	60.24	56.20
$\mathbf{\Sigma}^{-1/2}\hat{\mathbf{\Sigma}}_K\mathbf{\Sigma}^{-1/2}$	20.86	14.22	3.29	4.52	4.72	4.69	4.76
$p=300$							
$\hat{\mathbf{\Sigma}}_{u,K}^{\mathcal{T}}$	12.74	15.20	1.66	1.71	1.78	1.84	1.95
$(\hat{\mathbf{\Sigma}}_{u,K}^{\mathcal{T}})^{-1}$	7.58	7.80	1.74	2.18	2.58	3.54	5.45
$\hat{\mathbf{\Sigma}}_K^{-1}$	7.59	7.49	1.70	2.13	2.49	3.37	5.13
$\hat{\mathbf{\Sigma}}_K$	302.16	274.12	87.92	92.47	91.90	83.21	92.50
$\mathbf{\Sigma}^{-1/2}\hat{\mathbf{\Sigma}}_K\mathbf{\Sigma}^{-1/2}$	23.43	16.89	4.38	6.04	6.16	6.14	6.20

†True $K=3$.

Table 3 summarizes the average estimation error of covariance matrices across K in the spectral norm. Each simulation is replicated 50 times and $T=200$.

Table 3 illustrates some interesting patterns. First, the best accuracy of estimation is achieved when $K=K_0$. Second, the estimation is robust for $K \geqslant K_0$. As K increases from K_0, the estimation error becomes larger but is increasing slowly in general, which indicates the robustness when a slightly larger K has been used. Third, when the number of factors is underestimated, corresponding to $K=1,2$, all the estimators perform badly, which demonstrates the danger of missing any common factors. Therefore, overestimating the number of factors, while still maintaining a satisfactory accuracy of estimation of the covariance matrices, is much better than underestimating. The resulting bias caused by underestimation is more severe than the additional variance that is introduced by overestimation. Finally, estimating $\mathbf{\Sigma}$, the covariance of $\mathbf{y}_t$, does not achieve good accuracy even when $K=K_0$ in the absolute term $\|\hat{\mathbf{\Sigma}}-\mathbf{\Sigma}\|$, but the relative error $\|\mathbf{\Sigma}^{-1/2}\hat{\mathbf{\Sigma}}_K\mathbf{\Sigma}^{-1/2}-\mathbf{I}_p\|$ is much smaller. This is consistent with our discussions in Section 3.3.

6.5. Comparisons with other methods

6.5.1. Comparison with related methods

We compare POET with related methods that address low rank plus sparse covariance estimation, specifically, the low rank and sparse covariance estimator LOREC proposed by Luo (2011), the strict factor model SFM by Fan *et al.* (2008), the dual method (Dual) by Lin *et al.* (2009) and, finally, the singular value thresholding method of Cai *et al.* (2008), SVT. In particular,

Table 4. Method comparison under spectral norm for $T = 100$†

Method	$\hat{\mathbf{\Sigma}}_u$	$\hat{\mathbf{\Sigma}}_u^{-1}$	*RelE*	$\hat{\mathbf{\Sigma}}^{-1}$	$\hat{\mathbf{\Sigma}}$
p = 100					
POET	1.624	1.336	2.080	1.309	29.107
LOREC	2.274	1.880	2.564	1.511	32.365
SFM	2.084	2.039	2.707	2.022	34.949
Dual	2.306	5.654	2.707	4.674	29.000
SVT	2.59	13.64	2.806	103.1	29.670
p = 200					
POET	1.641	1.358	3.295	1.346	58.769
LOREC	2.179	1.767	3.874	1.543	62.731
SFM	2.098	2.071	3.758	2.065	60.905
Dual	2.41	6.554	4.541	5.813	56.264
SVT	2.930	362.5	4.680	47.21	63.670
p = 300					
POET	1.662	1.394	4.337	1.395	65.392
LOREC	2.364	1.635	4.909	1.742	91.618
SFM	2.091	2.064	4.874	2.061	88.852
Dual	2.475	2.602	6.190	2.234	74.059
SVT	2.681	$> 10^3$	6.247	$> 10^3$	80.954

†RelE represents the relative error $\|\mathbf{\Sigma}^{-1/2}\hat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1/2} - \mathbf{I}_p\|$.

SFM is a special case of POET which employs a large threshold that forces $\hat{\mathbf{\Sigma}}_u$ to be diagonal even when the true $\mathbf{\Sigma}_u$ might not be. Note that Dual, SVT and many others dealing with low rank plus sparseness, such as Candès *et al.* (2011) and Wright *et al.* (2009), assume a known $\mathbf{\Sigma}$ and focus on recovering the decomposition. Hence they do not estimate $\mathbf{\Sigma}$ or its inverse, but decompose the sample covariance into two components. The resulting sparse component may not be positive definite, which can lead to large estimation errors for $\hat{\mathbf{\Sigma}}_u^{-1}$ and $\hat{\mathbf{\Sigma}}^{-1}$.

Data are generated from the same set-up as design 2 in Section 6.4. Table 4 reports the averaged estimation error of the five methods being compared, calculated on the basis of 50 replications for each simulation. Dual and SVT assume that the data matrix has a low rank plus sparse representation, which is not so for the sample covariance matrix (though the population $\mathbf{\Sigma}$ has such a representation). The tuning parameters for POET, LOREC, Dual and SVT are chosen to achieve the best performance for each method. (We used the R package for LOREC that was developed by Luo (2011) and the MATLAB codes for Dual and SVT provided on Yi Ma's Web site 'Low-rank matrix recovery and completion via convex optimization' at the University of Illinois. The tuning parameters for each method have been chosen to minimize the sum of relative errors $\|\mathbf{\Sigma}^{-1/2}\hat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1/2} - \mathbf{I}_p\| + \|\mathbf{\Sigma}_u^{-1/2}\hat{\mathbf{\Sigma}}_u\mathbf{\Sigma}_u^{-1/2} - \mathbf{I}_p\|$. We have also written an R package for POET.)

6.5.2. Comparison with direct thresholding

This section compares POET with direct thresholding on the sample covariance matrix without taking out common factors (Rothman *et al.*, 2009; Cai and Liu, 2011). We denote this method by THR. We also run simulations to demonstrate the finite sample performance when $\mathbf{\Sigma}$ itself is sparse and has bounded eigenvalues, corresponding to the case $K = 0$. Three models are considered and both POET and THR use soft thresholding. We fix $T = 200$. Reported results are the average of 100 replications.

Table 5. Method comparison for $T = 200$†

Model	$\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|$		$\|\hat{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}^{-1}\|$		$\hat{K}$
	POET	*THR*	*POET*	*THR*	
$p = 200$					
1	26.20	240.18	1.31	2.67	1
2	2.04	2.04	2.07	2.07	0
3	7.73	11.24	8.48	11.40	6.2
$p = 300$					
1	32.60	314.43	2.18	2.58	1
2	2.03	2.03	2.08	2.08	0
3	9.41	11.29	8.81	11.41	5.45

†The reported numbers are the averages based on 100 replications.

(a) *Model 1, one-factor*: the factors and loadings are independently generated from $N(0, 1)$. The error covariance is the same banded matrix as design 2 in Section 6.4. Here $\mathbf{\Sigma}$ has one diverging eigenvalue.
(b) *Model 2, sparse covariance*: set $K = 0$; hence $\mathbf{\Sigma} = \mathbf{\Sigma}_u$ itself is a banded matrix with bounded eigenvalues.
(c) *Model 3, cross-sectional AR(1)*: set $K = 0$, but $\mathbf{\Sigma} = \mathbf{\Sigma}_u = (0.85^{|i-j|})_{p \times p}$. Now $\mathbf{\Sigma}$ is no longer sparse (or banded) but is not too dense either since Σ_{ij} decreases to 0 exponentially fast as $|i - j| \to \infty$. This is the correlation matrix if $\{y_{it}\}_{i=1}^{p}$ follows a cross-sectional AR(1) process: $y_{it} = 0.85 y_{i-1,t} + \varepsilon_{it}$.

For each model, POET uses an estimated $\hat{K}$ based on IC1 of Bai and Ng (2002), whereas THR thresholds the sample covariance directly. We find that, in model 1, POET performs significantly better than THR as the latter misses the common factor. For model 2, IC1 estimates $\hat{K} = 0$ precisely in each replication, and hence POET is identical to THR. For model 3, POET still outperforms THR. The results are summarized in Table 5.

6.6. Simulated portfolio allocation

We demonstrate the improvement of our method compared with the sample covariance and that based on the strict factor model, in a problem of portfolio allocation for risk minimization purposes.

Let $\hat{\mathbf{\Sigma}}$ be a generic estimator of the covariance matrix of the return vector $\mathbf{y}_t$, and $\mathbf{w}$ be the allocation vector of a portfolio consisting of the corresponding p financial securities. Then the theoretical and the empirical risk of the given portfolio are $R(\mathbf{w}) = \mathbf{w}'\mathbf{\Sigma}\mathbf{w}$ and $\hat{R}(\mathbf{w}) = \mathbf{w}'\hat{\mathbf{\Sigma}}\mathbf{w}$ respectively. Now, define

$$\hat{\mathbf{w}} = \operatorname*{arg\,min}_{\mathbf{w}'\mathbf{1}=1} \mathbf{w}'\hat{\mathbf{\Sigma}}\mathbf{w},$$

the estimated (minimum variance) portfolio. Then the actual risk of the estimated portfolio is defined as $R(\hat{\mathbf{w}}) = \hat{\mathbf{w}}'\mathbf{\Sigma}\hat{\mathbf{w}}$, and the estimated risk (which is also called the empirical risk) is equal to $\hat{R}(\hat{\mathbf{w}}) = \hat{\mathbf{w}}'\hat{\mathbf{\Sigma}}\hat{\mathbf{w}}$. In practice, the actual risk is unknown, and only the empirical risk can be calculated.

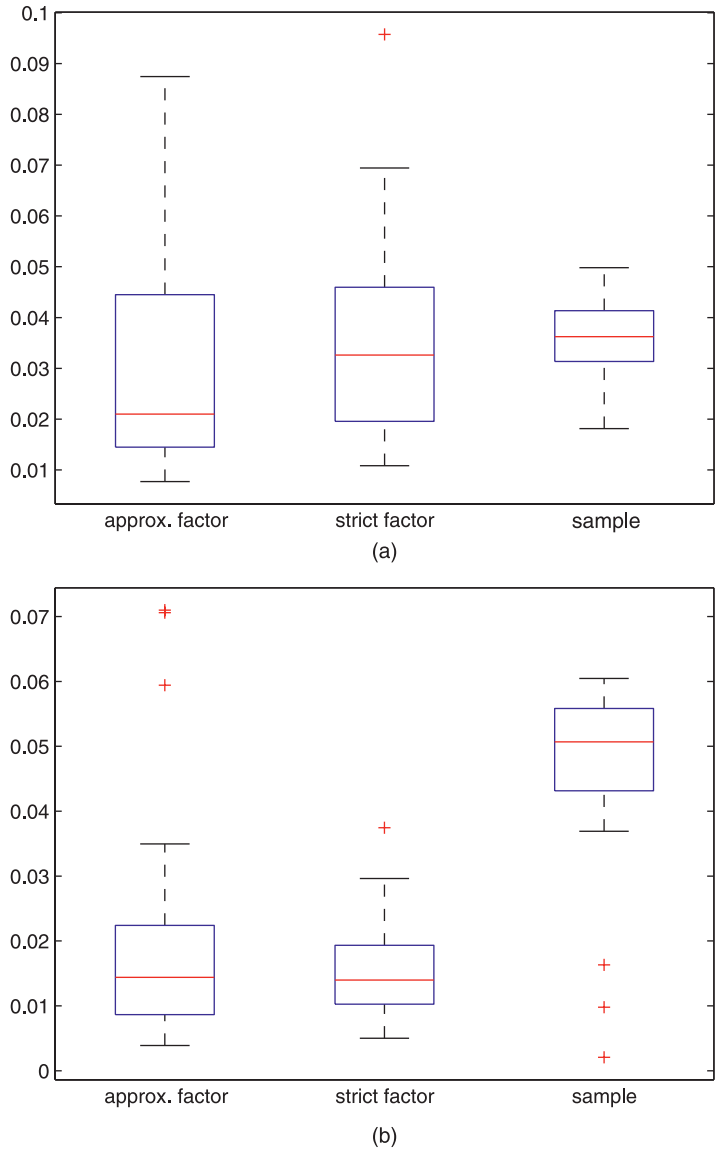

Fig. 7. Boxplots of regrets $R(\hat{\mathbf{w}}) - R^*$ for (a) $p = 80$ and (b) $p = 140$: in each panel, the boxplots from left to right correspond to $\hat{\mathbf{w}}$ obtained by using $\hat{\mathbf{\Sigma}}$ based on the approximate factor model, the SFM and the sample covariance

For each fixed p, the population $\mathbf{\Sigma}$ was generated in the same way as described in Section 6.1, with a sparse but not diagonal error covariance. We use three different methods to estimate $\mathbf{\Sigma}$ and to obtain $\hat{\mathbf{w}}$: strict factor model $\hat{\mathbf{\Sigma}}_{\text{diag}}$ (estimate $\mathbf{\Sigma}_u$ by using a diagonal matrix), our POET-estimator $\hat{\mathbf{\Sigma}}_{\text{POET}}$ (both are with unknown factors) and sample covariance $\hat{\mathbf{\Sigma}}_{\text{sam}}$. We then calculate the corresponding actual and empirical risks.

It is interesting to examine the accuracy and the performance of the actual risk of our portfolio $\hat{\mathbf{w}}$ in comparison with the oracle risk $R^* = \min_{\mathbf{w}'\mathbf{1}=1} \mathbf{w}'\mathbf{\Sigma}\mathbf{w}$, which is the theoretical

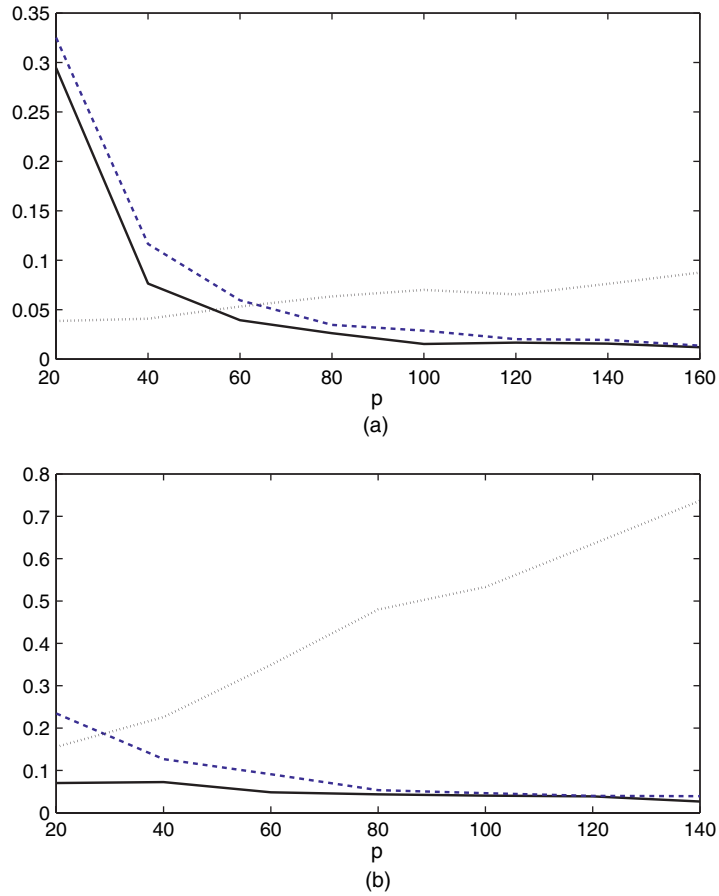

Fig. 8. Estimation errors for risk assessments as a function of the portfolio size p (———, POET; - - - - - -, SFM; · · · · · ·, sample): (a) average absolute error $|R(\hat{\mathbf{w}}) - \hat{R}(\hat{\mathbf{w}})|$; (b) average relative error $|\hat{R}(\hat{\mathbf{w}})/R(\hat{\mathbf{w}}) - 1|$ (here, $\hat{\mathbf{w}}$ and $\hat{R}$ are obtained on the basis of three estimators of $\hat{\boldsymbol{\Sigma}}$)

risk of the portfolio that we would have created if we knew the true covariance matrix $\boldsymbol{\Sigma}$. We thus compare the regret $R(\hat{\mathbf{w}}) - R^*$, which is always non-negative, for three estimators of $\hat{\boldsymbol{\Sigma}}$. They are summarized by using the boxplots over the 200 simulations. The results are reported in Fig. 7. In practice, we are also concerned about the difference between the actual and empirical risk of the chosen portfolio $\hat{\mathbf{w}}$. Hence, in Fig. 8, we also compare the average estimation error $|R(\hat{\mathbf{w}}) - \hat{R}(\hat{\mathbf{w}})|$ and the average relative estimation error $|\hat{R}(\hat{\mathbf{w}})/R(\hat{\mathbf{w}}) - 1|$ over 200 simulations. When $\hat{\mathbf{w}}$ is obtained on the basis of the strict factor model, both differences—between actual and oracle risk, and between actual and empirical risk—are persistently greater than the corresponding differences for the approximate factor estimator. Also, in terms of the relative estimation error, the factor-model-based method is negligible, whereas the sample covariance does not have such a property.

7. Real data example

We demonstrate the sparsity of the approximate factor model on real data and present the improvement of POET over the SFM in a real world application of portfolio allocation.

7.1. Sparsity of idiosyncratic errors

The data were obtained from the Center for Research in Security Prices database and consist of $p = 50$ stocks and their annualized daily returns for the period January 1st, 2010–December 31st, 2010 ($T = 252$). The stocks are chosen from five different industry sectors (more specifically, 'consumer goods—textiles and apparel clothing', 'financial—credit services', 'healthcare—hospitals', 'services—restaurants' and 'utilities—water utilities'), with 10 stocks from each sector. We made this selection to demonstrate a block diagonal trend in the sparsity. More specifically, we show that the non-zero elements are clustered mainly within companies in the same industry. We also note that these are the same groups that show predominantly positive correlation.

The largest eigenvalues of the sample covariance equal 0.0102, 0.0045 and 0.0039, whereas the rest are bounded by 0.0020. Hence $K = 0, 1, 2, 3$ are the possible values of the number of factors. Fig. 9 shows the heat map of the thresholded error correlation matrix (for simplicity, we applied hard thresholding). The threshold has been chosen by using cross-validation as described in Section 4. We compare the level of sparsity (the percentage of non-zero off-diagonal elements) for the five diagonal blocks of size 10×10, *versus* the sparsity of the rest of the matrix. For $K = 2$, our method results in 25.8% non-zero off-diagonal elements in the five diagonal blocks, as opposed to 7.3% non-zero elements in the rest of the covariance matrix. Note that, out of the non-zero elements in the central five blocks, 100% are positive, as opposed to a distribution of 60.3% positive and 39.7% negative among the non-zero elements in off-diagonal blocks. There is a strong positive correlation between the returns of companies in the same industry after the common factors have been taken out, and the thresholding has preserved them. The results for $K = 1$, 2, 3 show the same characteristics. These provide stark evidence that the strict factor model is not appropriate.

7.2. Portfolio allocation

We extend our data size by including larger industrial portfolios ($p = 100$), and a longer period (10 years): from January 1st, 2000, to December 31st, 2010, of annualized daily excess returns. Two portfolios are created at the beginning of each month, based on two different covariance estimates through approximate and strict factor models with unknown factors. At the end of each month, we compare the risks of both portfolios.

The number of factors is determined by using the penalty function that was proposed by Bai and Ng (2002), as defined in expression (2.14). For calibration, we use the last 100 consecutive business days of the above data, and both IC1 and IC2 give $\hat{K} = 3$. On the first of each month, we estimate $\hat{\mathbf{\Sigma}}_{\text{diag}}$ (method SFM) and $\hat{\mathbf{\Sigma}}_{\hat{K}}$ (POET with soft thresholding) using the historical data of excess daily returns for the preceeding 12 months ($T = 252$). The value of the threshold is determined by using the cross-validation procedure. We minimize the empirical risk of both portfolios to obtain the two respective optimal portfolio allocations $\hat{\mathbf{w}} = \hat{\mathbf{w}}_1$ and $\hat{\mathbf{w}} = \hat{\mathbf{w}}_2$ (based on $\hat{\mathbf{\Sigma}} = \hat{\mathbf{\Sigma}}_{\text{diag}}$ and $\hat{\mathbf{\Sigma}}_{\hat{K}}$): $\hat{\mathbf{w}} = \arg\min_{\hat{\mathbf{w}}'\mathbf{1}=1} \mathbf{w}'\hat{\mathbf{\Sigma}}\mathbf{w}$. At the end of the month (21 trading days), their actual risks are compared, calculated by

$$R_i = \hat{\mathbf{w}}_i' \frac{1}{21} \sum_{t=1}^{21} \mathbf{y}_t \mathbf{y}_t' \hat{\mathbf{w}}_i, \qquad \text{for } i = 1, 2.$$

We can see from Fig. 10 that the minimum risk portfolio that was created by POET performs significantly better, achieving lower variance 76% of the time. Among those months, the risk is decreased by 48.63%. In contrast, during the months that POET produces a higher risk portfolio, the risk is increased by only 17.66%.

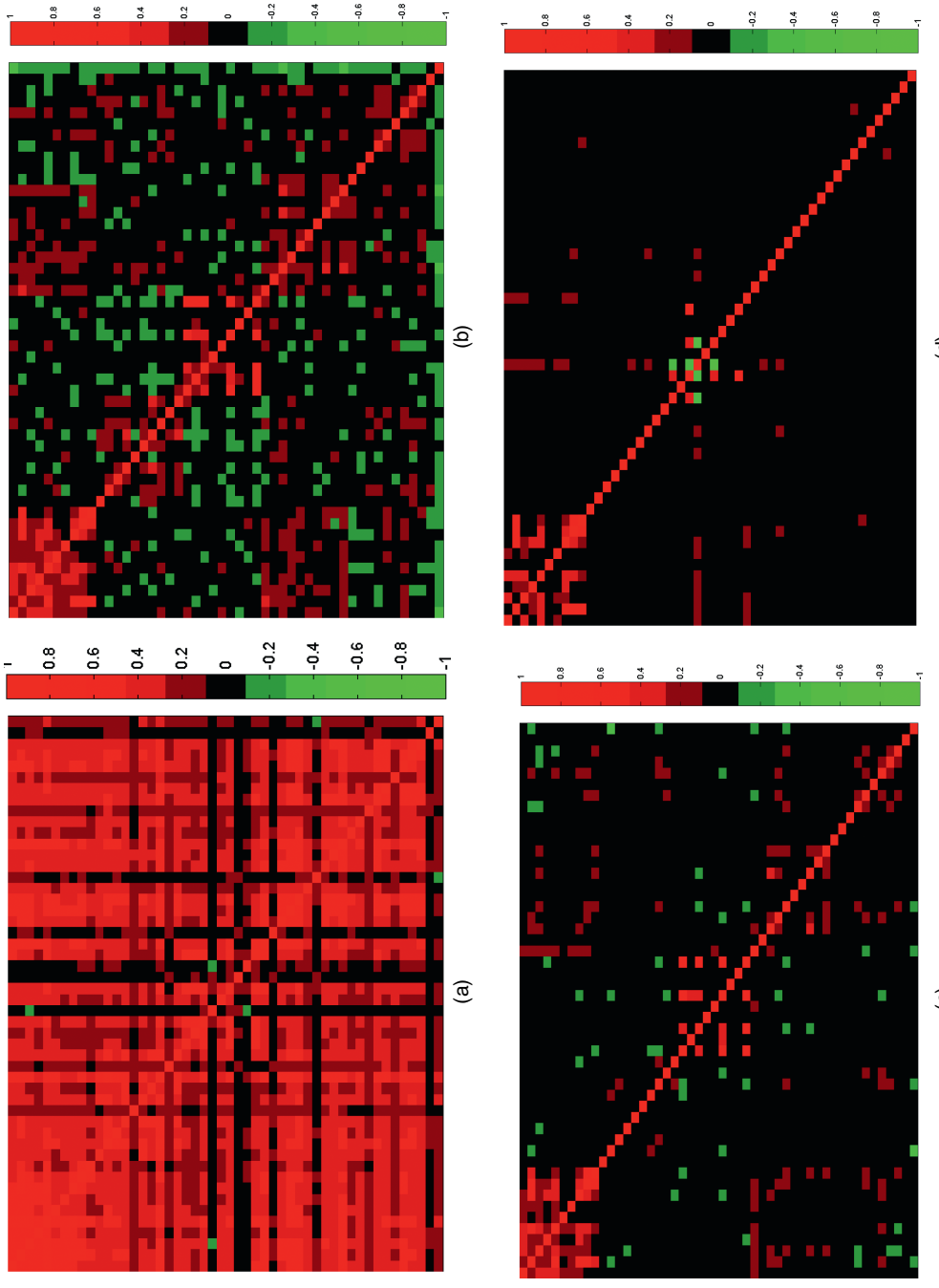

Fig. 9. Heat map of the thresholded error correlation matrix for number of factors (a) $K = 0$, (b) $K = 1$, (c) $K = 2$ and (d) $K = 3$

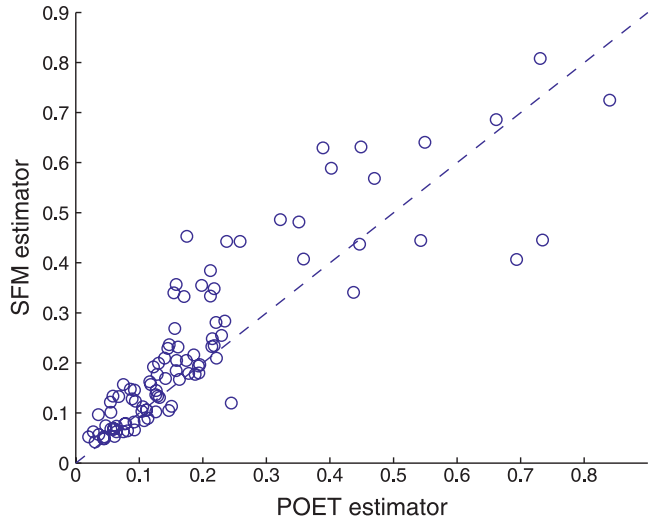

Fig. 10. Risk of portfolios created with POET and SFM

Table 6. Comparisons of the risks of portfolios by using POET and SFM†

C	*Results for the following values of* $\hat{K}$*:*		
	$\hat{K}=1$	$\hat{K}=2$	$\hat{K}=3$
0.25	0.58/29.6%	0.68/38%	0.71/33%
0.5	0.66/31.7%	0.70/38.2%	0.75/33.5%
0.75	0.68/29.3%	0.70/29.6%	0.71/25.1%
1	0.66/20.7%	0.62/19.4%	0.69/18%

†The first number is the proportion of the time that POET outperforms and the second number is the percentage of average risk improvements. C represents the constant in the threshold.

Next, we demonstrate the effect of the choice of number of factors and threshold on the performance of POET. If cross-validation seems computationally expensive, we can choose a common soft threshold throughout the whole investment process. The average constant in the cross-validation was 0.53, which is close to our suggested constant 0.5 used for simulation. We also present the results based on various choices of constant $C=0.5,0.75,1,1.25$, with soft threshold $C\omega_T\sqrt{\hat{\theta}_{ij}}$. The results are summarized in Table 6. The performance of POET seems consistent across different choices of these parameters.

8. Conclusion and discussion

We study the problem of estimating a high dimensional covariance matrix with conditional sparsity. Realizing that an unconditional sparsity assumption is inappropriate in many applications, we introduce a latent factor model that has a conditional sparsity feature and propose POET to take advantage of the structure. This expands considerably the scope of the model

based on the strict factor model, which assumes independent idiosyncratic noise and is too restrictive in practice. By assuming a sparse error covariance matrix, we allow for the presence of the cross-sectional correlation even after taking out the common factors. The sparse covariance is estimated by the adaptive thresholding technique.

It is found that the rates of convergence of the estimators have an extra term approximately $O_p(p^{-1/2})$ in addition to the results based on observable factors by Fan *et al.* (2008, 2011a), which arises from the effect of estimating the unobservable factors. As we can see, this effect vanishes as the dimensionality increases, as more information about the common factors becomes available. When p grows sufficiently large, the effect of estimating the unknown factors is negligible, and we estimate the covariance matrices as if we knew the factors.

The proposed POET also has wide applicability in statistical genomics. For example, Carvalho *et al.* (2008) applied a Bayesian sparse factor model to study breast cancer hormonal pathways. Their real data results have identified about two common factors that have highly loaded genes (about half of 250 genes). As a result, these factors should be treated as 'pervasive' (see the explanation in example 1 in Section 2.1.1), which will result in one or two very spiked eigenvalues of the gene expressions' covariance matrix. POET can be applied to estimate such a covariance matrix and its network model.

Acknowledgements

The research was partially supported by National Institutes of Health grants R01GM100474-01 and R01-GM072611, grant DMS-0704337 and the Bendheim Center for Finance at Princeton University. The bulk of the research was carried out while Yuan Liao was a postdoctoral fellow at Princeton University.

Appendix A: Estimating a sparse covariance with contaminated data

We estimate $\mathbf{\Sigma}_u$ by applying the adaptive thresholding given by expression (2.11). However, the task here is slightly different from the standard problem of estimating a sparse covariance matrix in the literature, as no direct observations for $\{\mathbf{u}_t\}_{t=1}^T$ are available. In many cases the original data are contaminated, including any type of estimate of the data when direct observations are not available. This typically happens when $\{\mathbf{u}_t\}_{t=1}^T$ represent the error terms in regression models or when data are subject to the measurement of errors. Instead, we may observe $\{\hat{\mathbf{u}}_t\}_{t=1}^T$. For instance, in the approximate factor models, $\hat{\mathbf{u}}_t = \mathbf{y}_t - \hat{\mathbf{b}}_i'\hat{\mathbf{f}}_t$.

We can estimate $\mathbf{\Sigma}_u$ by using the adaptive thresholding proposed by Cai and Liu (2011): for the threshold $\tau_{ij} = C\omega_T\sqrt{\hat{\theta}_{ij}}$, define

$$\hat{\sigma}_{ij} = \frac{1}{T}\sum_{t=1}^T \hat{u}_{it}\hat{u}_{jt},$$

$$\hat{\theta}_{ij} = \frac{1}{T}\sum_{t=1}^T (\hat{u}_{it}\hat{u}_{jt} - \hat{\sigma}_{ij})^2,$$

$$\hat{\mathbf{\Sigma}}_u^{\mathcal{T}} = (s_{ij}(\hat{\sigma}_{ij}))_{p\times p}, \tag{A.1}$$

where $s_{ij}(\cdot)$ satisfies, for all $z \in \mathbb{R}$, $s_{ij}(z) = 0$, when $|z| \leqslant \tau_{ij}$, and $|s_{ij}(z) - z| \leqslant \tau_{ij}$.

When $\{\hat{\mathbf{u}}_t\}_{t=1}^T$ is sufficiently close to $\{\mathbf{u}_t\}_{t=1}^T$, we can show that $\hat{\mathbf{\Sigma}}_u^{\mathcal{T}}$ is also consistent. The following theorem extends the standard thresholding results in Bickel and Levina (2008) and Cai and Liu (2011) to the case when no direct observations are available, or the original data are contaminated. For the tail and mixing parameters r_1 and r_3 that are defined in assumptions 2 and 3, let $\alpha = 3r_1^{-1} + r_3^{-1} + 1$.

Theorem 5. Suppose that $\log(p)^{6\alpha} = o(T)$, and that assumptions 2 and 3 hold. In addition, suppose that there is a sequence $a_T = o(1)$ so that $\max_{i\leqslant p}(1/T)\sum_{t=1}^T |u_{it} - \hat{u}_{it}|^2 = O_p(a_T^2)$, and $\max_{i\leqslant p,\, t\leqslant T} |u_{it} - \hat{u}_{it}| = o_p(1)$; then there is a constant $C > 0$ in the adaptive thresholding estimator (A.1) with

$$\omega_T = \sqrt{\left\{\frac{\log(p)}{T}\right\}} + a_T$$

such that

$$\|\hat{\mathbf{\Sigma}}_u^{\mathcal{T}} - \mathbf{\Sigma}_u\| = O_p(\omega_T^{1-q} m_p).$$

If further $\omega_T m_p = o(1)$, then $\hat{\mathbf{\Sigma}}_u^{\mathcal{T}}$ is invertible with probability approaching 1, and

$$\|(\hat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1} - \mathbf{\Sigma}_u^{-1}\| = O_p(\omega_T^{1-q} m_p).$$

Proof. By assumptions 2 and 3, the conditions of lemmas A.3 and A.4 of Fan *et al.* (2011a) are satisfied. Hence, for any $\varepsilon > 0$, there are positive constants M, θ_1 and θ_2 such that each of the events

$$A_1 = \{\max_{i \leqslant p, j \leqslant p} |\hat{\sigma}_{ij} - \sigma_{u,ij}| < M\omega_T\},$$
$$A_2 = \{\theta_1 > \sqrt{\hat{\theta}_{ij}} > \theta_2, \text{ all } i \leqslant p, j \leqslant p\}$$

occurs with probability at least $1-\varepsilon$. By the condition of threshold function, $s_{ij}(t) = s_{ij}(t) I_{|t| > C\omega_T \sqrt{\hat{\theta}_{ij}}}$. Now for $C = 2\theta_2^{-1} M$, under the event $A_1 \cap A_2$,

$$\begin{aligned}
\|\hat{\mathbf{\Sigma}}_u^{\mathcal{T}} - \mathbf{\Sigma}_u\| &\leqslant \max_{i \leqslant p} \sum_{j=1}^{p} |s_{ij}(\hat{\sigma}_{ij}) - \sigma_{u,ij}| \\
&= \max_{i \leqslant p} \sum_{j=1}^{p} |s_{ij}(\hat{\sigma}_{ij}) I_{(|\hat{\sigma}_{ij}| > C\omega_T \sqrt{\hat{\theta}_{ij}})} - \sigma_{u,ij} I_{(|\hat{\sigma}_{ij}| > C\omega_T \sqrt{\hat{\theta}_{ij}})} - \sigma_{u,ij} I_{(|\hat{\sigma}_{ij}| \leqslant C\omega_T \sqrt{\hat{\theta}_{ij}})}| \\
&\leqslant \max_{i \leqslant p} \sum_{j=1}^{p} |s_{ij}(\hat{\sigma}_{ij}) - \hat{\sigma}_{ij}| I_{(|\hat{\sigma}_{ij}| > C\omega_T \sqrt{\hat{\theta}_{ij}})} + \sum_{j=1}^{p} |\hat{\sigma}_{ij} - \sigma_{u,ij}| I_{(|\hat{\sigma}_{ij}| > C\omega_T \sqrt{\hat{\theta}_{ij}})} \\
&\quad + \sum_{j=1}^{p} |\sigma_{u,ij}| I_{(|\hat{\sigma}_{ij}| \leqslant C\omega_T \sqrt{\hat{\theta}_{ij}})} \\
&\leqslant \max_{i \leqslant p} \sum_{j=1}^{p} C\omega_T I_{(|\hat{\sigma}_{ij}| > C\omega_T \theta_2)} \sqrt{\hat{\theta}_{ij}} + M\omega_T \sum_{j=1}^{p} I_{(|\hat{\sigma}_{ij}| > C\omega_T \theta_2)} + \sum_{j=1}^{p} |\sigma_{u,ij}| I_{(|\hat{\sigma}_{ij}| \leqslant C\omega_T \theta_1)} \\
&\leqslant (C\theta_1 + M)\omega_T \max_{i \leqslant p} \sum_{j=1}^{p} I_{(|\sigma_{u,ij}| > M\omega_T)} + \max_{i \leqslant p} \sum_{j=1}^{p} |\sigma_{u,ij}| I_{(|\sigma_{u,ij}| \leqslant C\omega_T \theta_1 + M\omega_T)} \\
&\leqslant (C\theta_1 + M)\omega_T \max_{i \leqslant p} \sum_{j=1}^{p} \frac{|\sigma_{u,ij}|^q}{M^q \omega_T^q} I_{(|\sigma_{u,ij}| > M\omega_T)} \\
&\quad + \max_{i \leqslant p} \sum_{j=1}^{p} |\sigma_{u,ij}| \frac{(C\theta_1 + M)^{1-q} \omega_T^{1-q}}{|\sigma_{u,ij}|^{1-q}} I_{\{|\sigma_{u,ij}| \leqslant (C\theta_1 + M)\omega_T\}} \\
&\leqslant \frac{C\theta_1 + M}{M^q} \omega_T^{1-q} \max_{i \leqslant p} \sum_{j=1}^{p} |\sigma_{u,ij}|^q + \max_{i \leqslant p} \sum_{j=1}^{p} |\sigma_{u,ij}|^q (C\theta_1 + M)^{1-q} \omega_T^{1-q} \\
&= m_p \omega_T^{1-q} (C\theta_1 + M)\{M^{-q} + (C\theta_1 + M)^{-q}\}.
\end{aligned}$$

Let $M_1 = (C\theta_1 + M)\{M^{-q} + (C\theta_1 + M)^{-q}\}$. Then, with probability at least $1 - 2\varepsilon$, $\|\hat{\mathbf{\Sigma}}_u^{\mathcal{T}} - \mathbf{\Sigma}_u\| \leqslant m_p \omega_T^{1-q} M_1$. Since ε is arbitrary, we have $\|\hat{\mathbf{\Sigma}}_u^{\mathcal{T}} - \mathbf{\Sigma}_u\| = O_p(\omega_T^{1-q} m_p)$. If, in addition, $\omega_T m_p = o(1)$, then the minimum eigenvalue of $\hat{\mathbf{\Sigma}}_u^{\mathcal{T}}$ is bounded away from 0 with probability approaching 1 since $\lambda_{\min}(\mathbf{\Sigma}_u) > c_1$. This then implies that $\|(\hat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1} - \mathbf{\Sigma}_u^{-1}\| = O_p(\omega_T^{1-q} m_p)$.

Appendix B: Proofs for Section 2

We first cite two useful theorems, which are needed to prove propositions 1 and 2. In lemma 1 below, let $\{\lambda_i\}_{i=1}^p$ be the eigenvalues of $\mathbf{\Sigma}$ in descending order and $\{\boldsymbol{\xi}_i\}_{i=1}^p$ be their associated eigenvectors. Correspondingly, let $\{\hat{\lambda}_i\}_{i=1}^p$ be the eigenvalues of $\hat{\mathbf{\Sigma}}$ in descending order and $\{\hat{\boldsymbol{\xi}}_i\}_{i=1}^p$ be their associated eigenvectors.

Lemma 1.

(a) (*Weyl's theorem*) $|\hat{\lambda}_i - \lambda_i| \leqslant \|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|$.
(b) ($\sin(\theta)$ theorem; Davis and Kahan (1970)):

$$\|\hat{\boldsymbol{\xi}}_i - \boldsymbol{\xi}_i\| \leqslant \frac{\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|\sqrt{2}}{\min(|\hat{\lambda}_{i-1} - \lambda_i|, |\lambda_i - \hat{\lambda}_{i+1}|)}.$$

B.1. Proof of proposition 1

Since $\{\lambda_j\}_{j=1}^p$ are the eigenvalues of $\mathbf{\Sigma}$ and $\{\|\tilde{\mathbf{b}}_j\|^2\}_{j=1}^K$ are the first K eigenvalues of $\mathbf{BB}'$ (the remaining $p-K$ eigenvalues are 0), then by Weyl's theorem, for each $j \leqslant K$,

$$|\lambda_j - \|\tilde{\mathbf{b}}_j\|^2| \leqslant \|\mathbf{\Sigma} - \mathbf{BB}'\| = \|\mathbf{\Sigma}_u\|.$$

For $j > K$, $|\lambda_j| = |\lambda_j - 0| \leqslant \|\mathbf{\Sigma}_u\|$. However, the first K eigenvalues of $\mathbf{BB}$ are also the eigenvalues of $\mathbf{B}'\mathbf{B}$. By the assumption, the eigenvalues of $p^{-1}\mathbf{B}'\mathbf{B}$ are bounded away from 0. Thus, when $j \leqslant K$, $\|\tilde{\mathbf{b}}_j\|^2/p$ are bounded away from 0 for all large p.

B.2. Proof of proposition 2

Applying the $\sin(\theta)$ theorem yields

$$\left\| \xi_j - \frac{\tilde{\mathbf{b}}_j}{\|\tilde{\mathbf{b}}_j\|} \right\| \leqslant \frac{\|\mathbf{\Sigma}_u\|\sqrt{2}}{\min(|\lambda_{j-1} - \|\tilde{\mathbf{b}}_j\|^2|, |\|\tilde{\mathbf{b}}_j\|^2 - \lambda_{j+1}|)}.$$

For a generic constant $c > 0$, $|\lambda_{j-1} - \|\tilde{\mathbf{b}}_j\|^2| \geqslant |\|\tilde{\mathbf{b}}_{j-1}\|^2 - \|\tilde{\mathbf{b}}_j\|^2| - |\lambda_{j-1} - \|\tilde{\mathbf{b}}_{j-1}\|^2| \geqslant cp$ for all large p, since $|\|\tilde{\mathbf{b}}_{j-1}\|^2 - \|\tilde{\mathbf{b}}_j\|^2| \geqslant cp$ but $|\lambda_{j-1} - \|\tilde{\mathbf{b}}_{j-1}\|^2|$ is bounded by prosposition 1. However, if $j < K$, the same argument implies that $|\|\tilde{\mathbf{b}}_j\|^2 - \lambda_{j+1}| \geqslant cp$. If $j = K$, $|\|\tilde{\mathbf{b}}_j\|^2 - \lambda_{j+1}| = p|\|\tilde{\mathbf{b}}_K\|^2/p - \lambda_{K+1}/p|$, where $\|\tilde{\mathbf{b}}_K\|^2/p$ is bounded away from 0, but $\lambda_{K+1}/p = O(p^{-1})$. Hence, again, $|\|\tilde{\mathbf{b}}_j\|^2 - \lambda_{j+1}| \geqslant cp$.

B.3. Proof of theorem 1

The sample covariance matrix of the residuals by using the least squares method is given by

$$\hat{\mathbf{\Sigma}}_u \frac{1}{T}(\mathbf{Y} - \hat{\mathbf{\Lambda}}\hat{\mathbf{F}}')(\mathbf{Y}' - \hat{\mathbf{F}}\hat{\mathbf{\Lambda}}') = \frac{1}{T}\mathbf{YY}' - \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}',$$

where we used the normalization condition $\hat{\mathbf{F}}'\hat{\mathbf{F}} = T\mathbf{I}_K$ and $\hat{\mathbf{\Lambda}} = \mathbf{Y}\hat{\mathbf{F}}/T$. If we show that $\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' = \sum_{i=1}^K \hat{\lambda}_i \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i'$, then from the decompositions of the sample covariance,

$$\frac{1}{T}\mathbf{YY}' = \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' + \hat{\mathbf{\Sigma}}_u = \sum_{i=1}^K \hat{\lambda}_i \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i' + \hat{\mathbf{R}},$$

we have $\hat{\mathbf{R}} = \hat{\mathbf{\Sigma}}_u$. Consequently, applying thresholding on $\hat{\mathbf{\Sigma}}_u$ is equivalent to applying thresholding on $\hat{\mathbf{R}}$, which gives the desired result.

We now show that $\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' = \sum_{i=1}^K \hat{\lambda}_i \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i'$ indeed holds. Consider again the least squares problem (2.8) but with the following alternative normalization constraints: $(1/p)\sum_{i=1}^p \mathbf{b}_i\mathbf{b}_i' = \mathbf{I}_K$, and $(1/T)\sum_{t=1}^T \mathbf{f}_t\mathbf{f}_t'$ is diagonal. Let $(\tilde{\mathbf{\Lambda}}, \tilde{\mathbf{F}})$ be the solution to the new optimization problem. Switching the roles of $\mathbf{B}$ and $\mathbf{F}$, then the solution of problem (2.10) is $\tilde{\mathbf{\Lambda}} = (\hat{\boldsymbol{\xi}}_1, \ldots, \hat{\boldsymbol{\xi}}_K)$ and $\tilde{\mathbf{F}} = p^{-1}\mathbf{Y}'\tilde{\mathbf{\Lambda}}$. In addition, $T^{-1}\tilde{\mathbf{F}}'\tilde{\mathbf{F}} = \mathrm{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_K)$. From $\hat{\mathbf{\Lambda}}\hat{\mathbf{F}}' = \tilde{\mathbf{\Lambda}}\tilde{\mathbf{F}}'$, it follows that $\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' = (1/T)\hat{\mathbf{\Lambda}}\hat{\mathbf{F}}'\hat{\mathbf{F}}\hat{\mathbf{\Lambda}}' = (1/T)\tilde{\mathbf{\Lambda}}\tilde{\mathbf{F}}'\tilde{\mathbf{F}}\tilde{\mathbf{\Lambda}}' = \sum_{i=1}^K \hat{\lambda}_i \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i'$.

Appendix C: Proofs for Section 3

We shall proceed by subsequently showing theorems 4, 2 and 3.

C.1. Preliminary lemmas

The following results are to be used subsequently. The proofs of lemmas 2, 3 and 4 are found in Fan *et al.* (2011a).

Lemma 2. Suppose that $\mathbf{A}$ and $\mathbf{B}$ are symmetric semipositive definite matrices, and $\lambda_{\min}(\mathbf{B}) > c_T$ for a sequence $c_T > 0$. If $\|\mathbf{A}-\mathbf{B}\| = o_p(c_T)$, then $\lambda_{\min}(\mathbf{A}) > c_T/2$, and

$$\|\mathbf{A}^{-1} - \mathbf{B}^{-1}\| = O_p(c_T^{-2})\|\mathbf{A}-\mathbf{B}\|.$$

Lemma 3. Suppose that the random variables Z_1 and Z_2 both satisfy the exponential-type tail condition: there exist $r_1, r_2 \in (0,1)$ and $b_1, b_2 > 0$, such that, $\forall s > 0$,

$$P(|Z_i| > s) \leqslant \exp\{-(s/b_i)^{r_i}\}, \qquad i = 1, 2.$$

Then, for some r_3 and $b_3 > 0$, and any $s > 0$,

$$P(|Z_1 Z_2| > s) \leqslant \exp\{1 - (s/b_3)^{r_3}\}. \tag{C.1}$$

Lemma 4. Under the assumptions of theorem 2,

(a) $\max_{i,j\leqslant K} |(1/T)\Sigma_{t=1}^T f_{it} f_{jt} - E(f_{it} f_{jt})| = O_p\{\sqrt{(1/T)}\}$,
(b) $\max_{i,j\leqslant p} |(1/T)\Sigma_{t=1}^T u_{it} u_{jt} - E(u_{it} u_{jt})| = O_p[\sqrt{\{\log(p)/T\}}]$ and
(c) $\max_{i\leqslant K, j\leqslant p} |(1/T)\Sigma_{t=1}^T f_{it} u_{jt}| = O_p[\sqrt{\{\log(p)/T\}}]$.

Lemma 5. Let $\hat{\lambda}_K$ denote the Kth largest eigenvalue of $\hat{\mathbf{\Sigma}}_{\text{sam}} = (1/T)\Sigma_{t=1}^T \mathbf{y}_t \mathbf{y}_t'$; then $\hat{\lambda}_K > C_1 p$ with probability approaching 1 for some $C_1 > 0$.

Proof. First, by proposition 1, under assumption 1, the Kth largest eigenvalue λ_K of $\mathbf{\Sigma}$ satisfies, for some $c > 0$,

$$\lambda_K \geqslant \|\tilde{\mathbf{b}}_K\|^2 - |\lambda_K - \|\tilde{\mathbf{b}}_K\|^2| \geqslant cp\|\mathbf{\Sigma}_u\| \geqslant cp/2$$

for sufficiently large p. Using Weyl's theorem, we need only to prove that $\|\hat{\mathbf{\Sigma}}_{\text{sam}} - \mathbf{\Sigma}\| = o_p(p)$. Without loss of generality, we prove the result under the identifiability condition (2.1). Using model (1.2), $\hat{\mathbf{\Sigma}}_{\text{sam}} = T^{-1}\Sigma_{t=1}^T(\mathbf{B}\mathbf{f}_t + \mathbf{u}_t)(\mathbf{B}\mathbf{f}_t + \mathbf{u}_t)'$. Using this and model (1.3), $\hat{\mathbf{\Sigma}}_{\text{sam}} - \mathbf{\Sigma}$ can be decomposed as the sum of the four terms

$$\mathbf{D}_1 = (T^{-1}\mathbf{B}\sum_{t=1}^T \mathbf{f}_t\mathbf{f}_t' - \mathbf{I}_K)\mathbf{B}',$$

$$\mathbf{D}_2 = T^{-1}\sum_{t=1}^T (\mathbf{u}_t\mathbf{u}_t' - \mathbf{\Sigma}_u),$$

$$\mathbf{D}_3 = \mathbf{B}T^{-1}\sum_{t=1}^T \mathbf{f}_t\mathbf{u}_t',$$

$$\mathbf{D}_4 = \mathbf{D}_3'.$$

We now deal with them term by term. We shall repeatedly use the fact that, for a $p \times p$ matrix $\mathbf{A}$,

$$\|\mathbf{A}\| \leqslant p\|\mathbf{A}\|_{\max}.$$

First, by lemma 4, $\|T^{-1}\Sigma_{t=1}^T \mathbf{f}_t\mathbf{f}_t' - \mathbf{I}_K\| \leqslant K\|T^{-1}\Sigma_{t=1}^T \mathbf{f}_t\mathbf{f}_t' - \mathbf{I}_K\|_{\max} = O_p\{\sqrt{(1/T)}\}$, which is $o_p(p)$ if $K\log(p) = o(T)$. Consequently, by assumption 1, we have

$$\|\mathbf{D}_1\| \leqslant O_p[K\sqrt{\{\log(K)/T\}}]\|\mathbf{B}\mathbf{B}'\| = O_p\{p\sqrt{(1/T)}\}.$$

We now deal with $\mathbf{D}_2$. It follows from lemma 4 that

$$\|\mathbf{D}_2\| \leqslant p\|T^{-1}\sum_{t=1}^T (\mathbf{u}_t\mathbf{u}_t' - \mathbf{\Sigma}_u)\|_{\max} = O_p[p\sqrt{\{\log(p)/T\}}].$$

Since $\|\mathbf{D}_4\| = \|\mathbf{D}_3\|$, it remains to deal with $\mathbf{D}_3$, which is bounded by

$$\|\mathbf{D}_3\| \leqslant \|T^{-1}\sum_{t=1}^T \mathbf{f}_t\mathbf{u}_t'\|\|\mathbf{B}\| = O_p[p\sqrt{\{\log(p)/T\}}],$$

which is $o_p(p)$ since $\log(p) = o(T)$.

Lemma 6. Under assumption 3, $\max_{t\leqslant T}\Sigma_{s=1}^{T}|E(\mathbf{u}_s'\mathbf{u}_t)|/p = O(1)$.

Proof. Since $\{\mathbf{u}_t\}_{t=1}^{T}$ is weakly stationary, $\max_{t\leqslant T}\Sigma_{s=1}^{T}|E(\mathbf{u}_s'\mathbf{u}_t)|/p \leqslant 2\Sigma_{t=1}^{\infty}|E(\mathbf{u}_1'\mathbf{u}_t)|/p$. In addition, $E|u_{it}|^4 < M$ for some constant M and any i and t since u_{it} has an exponential tail. Hence by Davydov's inequality, (corollary 16.2.4 in Athreya and Lahiri (2006)), there is a constant $C>0$, for all $i\leqslant p, t\leqslant T$, $|E(u_{i1}u_{it})|\leqslant C\sqrt{\alpha(t)}$, where $\alpha(t)$ is the α-mixing coefficient. By assumption 3, $\Sigma_{t=1}^{\infty}\sqrt{\alpha(t)}<\infty$. Thus, uniformly in T,

$$\max_{t\leqslant T}\sum_{s=1}^{T}|E(\mathbf{u}_s'\mathbf{u}_t)|/p \leqslant 2\sum_{t=1}^{\infty}|E(\mathbf{u}_1'\mathbf{u}_t)|/p \leqslant 2\sum_{t=1}^{\infty}\max_{i\leqslant p}|E(u_{i1}u_{it})| \leqslant 2C\sum_{t=1}^{\infty}\sqrt{\alpha(t)}<\infty.$$

C.2. Proof of theorem 4

Our derivation below relies on a result that was obtained by Bai and Ng (2002), which showed that the estimated number of factors is consistent, in the sense that $\hat{K}$ equals the true K with probability approaching 1. Note that, under our assumptions 1–4, all the assumptions in Bai and Ng (2002) are satisfied. Thus immediately we have the following lemma.

Lemma 7 (theorem 2 in Bai and Ng (2002)). For $\hat{K}$ defined in expression (2.14),

$$P(\hat{K}=K)\to 1.$$

Proof. For a proof, see Bai and Ng (2002).

Using expression (A.1) in Bai (2003), we have the identity

$$\hat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t = \left(\frac{\mathbf{V}}{p}\right)^{-1}\left\{\frac{1}{T}\sum_{s=1}^{T}\hat{\mathbf{f}}_s\frac{E(\mathbf{u}_s'\mathbf{u}_t)}{p} + \frac{1}{T}\sum_{s=1}^{T}\hat{\mathbf{f}}_s\zeta_{st} + \frac{1}{T}\sum_{s=1}^{T}\hat{\mathbf{f}}_s\eta_{st} + \frac{1}{T}\sum_{s=1}^{T}\hat{\mathbf{f}}_s\xi_{st}\right\} \tag{C.2}$$

where $\zeta_{st} = \mathbf{u}_s'\mathbf{u}_t/p - E(\mathbf{u}_s'\mathbf{u}_t)/p$, $\eta_{st} = \mathbf{f}_s'\Sigma_{i=1}^{p}\mathbf{b}_i u_{it}/p$ and $\xi_{st} = \mathbf{f}_t'\Sigma_{i=1}^{p}\mathbf{b}_i u_{is}/p$.

We first prove some preliminary results in the following lemmas. Denote $\hat{\mathbf{f}}_t = (\hat{f}_{1t},\ldots,\hat{f}_{\hat{K}t})'$.

Lemma 8. For all $i\leqslant\hat{K}$,

(a) $(1/T)\Sigma_{t=1}^{T}\{(1/T)\Sigma_{s=1}^{T}\hat{f}_{is}E(\mathbf{u}_s'\mathbf{u}_t)/p\}^2 = O_p(T^{-1})$,
(b) $(1/T)\Sigma_{t=1}^{T}\{(1/T)\Sigma_{s=1}^{T}\hat{f}_{is}\zeta_{st}\}^2 = O_p(p^{-1})$,
(c) $(1/T)\Sigma_{t=1}^{T}\{(1/T)\Sigma_{s=1}^{T}\hat{f}_{is}\eta_{st}\}^2 = O_p(p^{-1})$ and
(d) $(1/T)\Sigma_{t=1}^{T}\{(1/T)\Sigma_{s=1}^{T}\hat{f}_{is}\xi_{st}\}^2 = O_p(p^{-1})$.

Proof.

(a) We have, $\forall i$, $\Sigma_{s=1}^{T}\hat{f}_{is}^2 = T$. By the Cauchy–Schwarz inequality,

$$\frac{1}{T}\sum_{t=1}^{T}\left\{\frac{1}{T}\sum_{s=1}^{T}\hat{f}_{is}\frac{E(\mathbf{u}_s'\mathbf{u}_t)}{p}\right\}^2 \leqslant \frac{1}{T}\sum_{t=1}^{T}\frac{1}{T}\sum_{s=1}^{T}\left\{\frac{E(\mathbf{u}_s'\mathbf{u}_t)}{p}\right\}^2$$
$$\leqslant \max_{t\leqslant T}\frac{1}{T}\sum_{s=1}^{T}\left\{\frac{E(\mathbf{u}_s'\mathbf{u}_t)}{p}\right\}^2 \leqslant \max_{s,t}\left|\frac{E(\mathbf{u}_s'\mathbf{u}_t)}{p}\right|\max_{t\leqslant T}\frac{1}{T}\sum_{s=1}^{T}\left|\frac{E(\mathbf{u}_s'\mathbf{u}_t)}{p}\right|.$$

By lemma 6, $\max_{t\leqslant T}\Sigma_{s=1}^{T}|E(\mathbf{u}_s'\mathbf{u}_t)/p| = O(1)$, which then yields the result.

(b) By the Cauchy–Schwarz inequality,

$$\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{T}\sum_{s=1}^{T}\hat{f}_{is}\zeta_{st}\right)^2 = \frac{1}{T^3}\sum_{s=1}^{T}\sum_{l=1}^{T}\hat{f}_{is}\hat{f}_{il}\left(\sum_{t=1}^{T}\zeta_{st}\zeta_{lt}\right) \leqslant \frac{1}{T^3}\left\{\sum_{sl}(\hat{f}_{is}\hat{f}_{il})^2\sum_{sl}\left(\sum_{t=1}^{T}\zeta_{st}\zeta_{lt}\right)^2\right\}^{1/2}$$
$$\leqslant \frac{1}{T^3}\sum_{s=1}^{T}\hat{f}_{is}^2\left\{\sum_{sl}\left(\sum_{t=1}^{T}\zeta_{st}\zeta_{lt}\right)^2\right\}^{1/2} = \frac{1}{T^2}\left\{\sum_{s=1}^{T}\sum_{l=1}^{T}\left(\sum_{t=1}^{T}\zeta_{st}\zeta_{lt}\right)^2\right\}^{1/2}.$$

Note that $E\{\Sigma_{s=1}^{T}\Sigma_{l=1}^{T}(\Sigma_{t=1}^{T}\zeta_{st}\zeta_{lt})^2\} = T^2E(\Sigma_{t=1}^{T}\zeta_{st}\zeta_{lt})^2 \leqslant T^4\max_{st}E|\zeta_{st}|^4$. By assumption 4, $\max_{st}E(\zeta_{st}^4) = O(p^{-2})$, which implies that $\Sigma_{s,l}(\Sigma_{t=1}^{T}\zeta_{st}\zeta_{lt})^2 = O_p(T^4/p^2)$ and yields the result.

(c) By definition, $\eta_{st} = \mathbf{f}_s' \Sigma_{i=1}^p \mathbf{b}_i u_{it}/p$. We first bound $\|\Sigma_{i=1}^p \mathbf{b}_i u_{it}\|$. Assumption 4 implies that $E\{(1/T) \times \Sigma_{t=1}^T \|\Sigma_{i=1}^p \mathbf{b}_i u_{it}\|^2\} = E\|\Sigma_{i=1}^p \mathbf{b}_i u_{it}\|^2 = O(p)$. Therefore, by the Cauchy–Schwarz inequality,

$$\frac{1}{T}\sum_{t=1}^T \left(\frac{1}{T}\sum_{s=1}^T \hat{f}_{is}\eta_{st}\right)^2 \leqslant \left\| \frac{1}{T}\sum_{s=1}^T \hat{f}_{is}\mathbf{f}_s' \right\|^2 \frac{1}{T}\sum_{t=1}^T \left\| \sum_{j=1}^p \mathbf{b}_j u_{jt} \frac{1}{p} \right\|^2$$
$$\leqslant \frac{1}{Tp^2}\sum_{t=1}^T \left\| \sum_{j=1}^p \mathbf{b}_j u_{jt} \right\|^2 \left(\frac{1}{T}\sum_{s=1}^T \hat{f}_{is}^2 \frac{1}{T}\sum_{s=1}^T \|\mathbf{f}_s\|^2\right) = O_p\left(\frac{1}{p}\right).$$

(d) Similarly to part (c), noting that ξ_{st} is a scalar, we have

$$\frac{1}{T}\sum_{t=1}^T \left(\frac{1}{T}\sum_{s=1}^T \hat{f}_{is}\xi_{st}\right)^2 = \frac{1}{T}\sum_{t=1}^T \left| \frac{1}{T}\sum_{s=1}^T \mathbf{f}_t' \sum_{j=1}^p \mathbf{b}_j u_{js} \frac{1}{p}\hat{f}_{is} \right|^2 \leqslant \left(\frac{1}{T}\sum_{t=1}^T \|\mathbf{f}_t\|^2\right) \left\| \frac{1}{T}\sum_{s=1}^T \sum_{j=1}^p \mathbf{b}_j u_{js} \frac{1}{p}\hat{f}_{is} \right\|^2$$
$$\leqslant \left\{ O_p(1) \frac{1}{T}\sum_{s=1}^T \left\| \sum_{j=1}^p \mathbf{b}_j u_{js} \frac{1}{p} \right\|^2 \right\} \frac{1}{T}\sum_{s=1}^T \hat{f}_{is}^2 \leqslant O_p\left(\frac{1}{p}\right),$$

where the last line follows from the Cauchy–Schwarz inequality.

Lemma 9.

(a) $\max_{t\leqslant T} \|\{1/(Tp)\}\Sigma_{s=1}^T \hat{\mathbf{f}}_s E(\mathbf{u}_s'\mathbf{u}_t)\| = O_p\{\sqrt{(1/T)}\}$,
(b) $\max_{t\leqslant T} \|(1/T)\Sigma_{s=1}^T \hat{\mathbf{f}}_s \zeta_{st}\| = O_p(T^{1/4}/\sqrt{p})$,
(c) $\max_{t\leqslant T} \|(1/T)\Sigma_{s=1}^T \hat{\mathbf{f}}_s \eta_{st}\| = O_p(T^{1/4}/\sqrt{p})$ and
(d) $\max_{t\leqslant T} \|(1/T)\Sigma_{s=1}^T \hat{\mathbf{f}}_s \xi_{st}\| = O_p(T^{1/4}/\sqrt{p})$.

Proof.

(a) By the Cauchy–Schwarz inequality and the fact that $(1/T)\Sigma_{t=1}^T \|\hat{\mathbf{f}}_t\|^2 = O_p(1)$,

$$\max_{t\leqslant T} \left\| \frac{1}{Tp}\sum_{s=1}^T \hat{\mathbf{f}}_s E(\mathbf{u}_s'\mathbf{u}_t) \right\| \leqslant \max_{t\leqslant T} \left[\frac{1}{T}\sum_{s=1}^T \|\hat{\mathbf{f}}_s\|^2 \frac{1}{T}\sum_{s=1}^T \left\{ \frac{E(\mathbf{u}_s'\mathbf{u}_t)}{p} \right\}^2 \right]^{1/2}$$
$$\leqslant O_p(1) \max_{t\leqslant T} \left[\frac{1}{T}\sum_{s=1}^T \left\{ \frac{E(\mathbf{u}_s'\mathbf{u}_t)}{p} \right\}^2 \right]^{1/2} \leqslant O_p(1) \max_{s,t} \sqrt{\left\{ \left| \frac{E(\mathbf{u}_s'\mathbf{u}_t)}{p} \right| \right\}} \max_{t\leqslant T} \left\{ \frac{1}{T}\sum_{s=1}^T \left| \frac{E(\mathbf{u}_s'\mathbf{u}_t)}{p} \right| \right\}^{1/2}.$$

The result then follows from assumption 3.

(b) By the Cauchy–Schwarz inequality,

$$\max_{t\leqslant T} \left\| \frac{1}{T}\sum_{s=1}^T \hat{\mathbf{f}}_s \zeta_{st} \right\| \leqslant \max_{t\leqslant T} \frac{1}{T} \left(\sum_{s=1}^T \|\hat{\mathbf{f}}_s\|^2 \sum_{s=1}^T \zeta_{st}^2 \right)^{1/2} \leqslant \left\{ O_p(1) \max_t \frac{1}{T}\sum_{s=1}^T \zeta_{st}^2 \right\}^{1/2}.$$

It follows from assumption 4 that $E\{(1/T)\Sigma_{s=1}^T \zeta_{st}^2\}^2 \leqslant \max_{s,t\leqslant T} E(\zeta_{st}^4) = O(1/p^2)$. It then follows from Chebyshev's inequality and Bonferroni's method that $\max_t (1/T)\Sigma_{s=1}^T \zeta_{st}^2 = O_p(\sqrt{T/p})$.

(c) By assumption 4, $E\|(1/\sqrt{p})\Sigma_{i=1}^p \mathbf{b}_i u_{it}\|^4 \leqslant K^2 M$. Chebyshev's inequality and Bonferroni's method yield $\max_{t\leqslant T} \|\Sigma_{i=1}^p \mathbf{b}_i u_{it}\| = O_p(T^{1/4}\sqrt{p})$ with probability 1, which then implies

$$\max_{t\leqslant T} \left\| \frac{1}{T}\sum_{s=1}^T \hat{\mathbf{f}}_s \eta_{st} \right\| \leqslant \left\| \frac{1}{T}\sum_{s=1}^T \hat{\mathbf{f}}_s \mathbf{f}_s' \right\| \max_t \left\| \frac{1}{p}\sum_{i=1}^p \mathbf{b}_i u_{it} \right\| = o_p\left(\frac{T^{1/4}}{p^{1/2}}\right).$$

(d) By the Cauchy–Schwarz inequality and assumption 4, we have demonstrated that $\|(1/T) \times \Sigma_{s=1}^T \Sigma_{i=1}^p \mathbf{b}_i u_{is}(1/p)\hat{\mathbf{f}}_s\| = O_p(p^{-1/2})$. In addition, since $E\|K^{-2}\mathbf{f}_t\|^4 < M$, $\max_{t\leqslant T} \|\mathbf{f}_t\| = O_p(T^{1/4})$. It follows that

$$\max_{t\leqslant T} \left\| \frac{1}{T}\sum_{s=1}^T \hat{\mathbf{f}}_s \xi_{st} \right\| \leqslant \max_{t\leqslant T} \|\mathbf{f}_t\| \cdot \left\| \frac{1}{T}\sum_{s=1}^T \sum_{i=1}^p \mathbf{b}_i u_{is} \frac{1}{p}\hat{\mathbf{f}}_s \right\| = O_p\left(\frac{T^{1/4}}{p^{1/2}}\right).$$

Lemma 10.

(a) $\max_{i\leqslant K} (1/T)\Sigma_{t=1}^T (\hat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t)_i^2 = O_p(1/T + 1/p)$.
(b) $(1/T)\Sigma_{t=1}^T \|\hat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t\|^2 = O_p(1/T + 1/p)$.
(c) $\max_{t\leqslant T} \|\hat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t\| = O_p\{\sqrt{(1/T)} + T^{1/4}/\sqrt{p}\}$.

Proof. We prove this lemma conditioning on the event $\hat{K} = K$. Once this has been done, because $P(\hat{K} \neq K) = o(1)$, it then implies the unconditional arguments.

(a) When $\hat{K} = K$, by lemma 5, all the eigenvalues of $\mathbf{V}/p$ are bounded away from 0. Using the inequality $(a+b+c+d)^2 \leqslant 4(a^2+b^2+c^2+d^2)$ and identity (C.2), we have, for some constant $C > 0$,

$$\max_{i \leqslant K} \frac{1}{T}\sum_{t=1}^{T}(\hat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t)_i^2 \leqslant C \max_{i \leqslant K} \frac{1}{T}\sum_{t=1}^{T}\left\{\frac{1}{T}\sum_{s=1}^{T}\hat{f}_{is}\frac{E(\mathbf{u}_s'\mathbf{u}_t)}{p}\right\}^2 + C\max_{i \leqslant K}\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{T}\sum_{s=1}^{T}\hat{f}_{is}\zeta_{st}\right)^2 + C\max_{i \leqslant K}\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{T}\sum_{s=1}^{T}\hat{f}_{is}\eta_{st}\right)^2 + C\max_{i \leqslant K}\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{T}\sum_{s=1}^{T}\hat{f}_{is}\xi_{st}\right)^2.$$

Each of the four terms on the right-hand side are bounded in lemma 8, which then yields the desired result.

(b) Part (b) follows from part (a) and

$$\frac{1}{T}\sum_{t=1}^{T}\|\hat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t\|^2 \leqslant K \max_{i \leqslant K}\frac{1}{T}\sum_{t=1}^{T}(\hat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t)_i^2.$$

Part (c) is implied by identity (C.2) and lemma 9.

Lemma 11.

(a) $\mathbf{H}\mathbf{H}' = \mathbf{I}_{\hat{K}} + O_p(1/\sqrt{T} + 1/\sqrt{p})$.
(b) $\mathbf{H}'\mathbf{H} = \mathbf{I}_K + O_p(1/\sqrt{T} + 1/\sqrt{p})$.

Proof. We first condition on $\hat{K} = K$.

(a) Lemma 5 implies that $\|\mathbf{V}^{-1}\| = O_p(p^{-1})$. Also $\|\mathbf{F}\| = \lambda_{\max}^{1/2}(\mathbf{F}\mathbf{F}') = \lambda_{\max}^{1/2}(\Sigma_{t=1}^{T}\mathbf{f}_t\mathbf{f}_t') = O_p(\sqrt{T})$. In addition, $\|\hat{\mathbf{F}}\| = \sqrt{T}$. It then follows from the definition of $\mathbf{H}$ that $\|\mathbf{H}\| = O_p(1)$. Define $\widehat{\mathrm{cov}}(\mathbf{H}\mathbf{f}_t) = (1/T)\Sigma_{t=1}^{T}\mathbf{H}\mathbf{f}_t(\mathbf{H}\mathbf{f}_t)'$. Applying the triangular inequality gives

$$\|\mathbf{H}\mathbf{H}' - \mathbf{I}_{\hat{K}}\|_{\mathrm{F}} \leqslant \|\mathbf{H}\mathbf{H}' - \widehat{\mathrm{cov}}(\mathbf{H}\mathbf{f}_t)\|_{\mathrm{F}} + \|\widehat{\mathrm{cov}}(\mathbf{H}\mathbf{f}_t) - \mathbf{I}_{\hat{K}}\|_{\mathrm{F}}. \tag{C.3}$$

By lemma 4, the first term in inequality (C.3) is $\|\mathbf{H}\mathbf{H}' - \widehat{\mathrm{cov}}(\mathbf{H}\mathbf{f}_t)\|_{\mathrm{F}} \leqslant \|\mathbf{H}\|^2\|\mathbf{I}_K - \widehat{\mathrm{cov}}(\mathbf{f}_t)\|_{\mathrm{F}} = O_p(1/\sqrt{T})$. The second term of inequality (C.3) can be bounded, by the Cauchy–Schwarz inequality and lemma 10, as follows:

$$\begin{aligned}\left\|\frac{1}{T}\sum_{t=1}^{T}\mathbf{H}\mathbf{f}_t(\mathbf{H}\mathbf{f}_t)' - \frac{1}{T}\sum_{t=1}^{T}\hat{\mathbf{f}}_t\hat{\mathbf{f}}_t'\right\|_{\mathrm{F}} &\leqslant \left\|\frac{1}{T}\sum_t(\mathbf{H}\mathbf{f}_t - \hat{\mathbf{f}}_t)(\mathbf{H}\mathbf{f}_t)'\right\|_{\mathrm{F}} + \left\|\frac{1}{T}\sum_t\hat{\mathbf{f}}_t(\hat{\mathbf{f}}_t' - (\mathbf{H}\mathbf{f}_t)')\right\|_{\mathrm{F}} \\ &\leqslant \left(\frac{1}{T}\sum_t\|\mathbf{H}\mathbf{f}_t - \hat{\mathbf{f}}_t\|^2\frac{1}{T}\sum_t\|\mathbf{H}\mathbf{f}_t\|^2\right)^{1/2} + \left(\frac{1}{T}\sum_t\|\mathbf{H}\mathbf{f}_t - \hat{\mathbf{f}}_t\|^2\frac{1}{T}\sum_t\|\hat{\mathbf{f}}_t\|^2\right)^{1/2} \\ &= O_p\left(\frac{1}{\sqrt{T}} + \frac{1}{\sqrt{p}}\right).\end{aligned}$$

(b) Still conditioning on $\hat{K} = K$, since $\mathbf{H}\mathbf{H}' = \mathbf{I}_K + O_p(1/\sqrt{T} + 1/\sqrt{p})$ and $\|\mathbf{H}\| = O_p(1)$, right multiplying $\mathbf{H}$ gives $\mathbf{H}\mathbf{H}'\mathbf{H} = \mathbf{H} + O_p(1/\sqrt{T} + 1/\sqrt{p})$. Part (a) also gives, conditioning on $\hat{K} = K$, $\|\mathbf{H}^{-1}\| = O_p(1)$. Hence further left multiplying $\mathbf{H}^{-1}$ yields $\mathbf{H}'\mathbf{H} = \mathbf{I}_K + O_p(1/\sqrt{T} + \sqrt{p})$. Because $P(\hat{K} = K) \to 1$, we reach the desired result.

C.2.1. Completion of proof of theorem 4

The second part of theorem 4 was proved in lemma 10. We now derive the convergence rate of $\max_{i \leqslant p}\|\hat{\mathbf{b}}_i - \mathbf{H}\mathbf{b}_i\|$.

Using the facts that $\hat{\mathbf{b}}_i = (1/T)\Sigma_{t=1}^{T}y_{it}\hat{\mathbf{f}}_t$, and that $(1/T)\Sigma_{t=1}^{T}\hat{\mathbf{f}}_t\hat{\mathbf{f}}_t' = I_k$, we have

$$\hat{\mathbf{b}}_i - \mathbf{H}\mathbf{b}_i = \frac{1}{T}\sum_{t=1}^{T}\mathbf{H}\mathbf{f}_t u_{it} + \frac{1}{T}\sum_{t=1}^{T}y_{it}(\hat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t) + \mathbf{H}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{f}_t\mathbf{f}_t' - \mathbf{I}_K\right)\mathbf{b}_i. \tag{C.4}$$

We bound the three terms on the right-hand side. It follows from lemmas 4 and 11 that

$$\max_{i \leqslant p}\left\|\frac{1}{T}\sum_{t=1}^{T}\mathbf{H}\mathbf{f}_t u_{it}\right\| \leqslant \|\mathbf{H}\|\max_i\sqrt{\left\{\sum_{k=1}^{K}\left(\frac{1}{T}\sum_{t=1}^{T}f_{kt}u_{it}\right)^2\right\}} = O_p\left[\left\{\frac{\log(p)}{T}\right\}\right].$$

For the second term, $E(y_{it}^2) = O(1)$. Therefore, $\max_i T^{-1}\Sigma_{t=1}^T y_{it}^2 = O_p(1)$. The Cauchy–Schwarz inequality and lemma 10 imply

$$\max_i \left\| \frac{1}{T}\sum_{t=1}^T y_{it}(\hat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t)\right\| \leqslant \max_i \left(\frac{1}{T}\sum_{t=1}^T y_{it}^2 \frac{1}{T}\sum_{t=1}^T \|\hat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t\|^2\right)^{1/2} = O_p\left(\frac{1}{\sqrt{T}} + \frac{1}{\sqrt{p}}\right).$$

Finally, $\|(1/T)\Sigma_{t=1}^T \mathbf{f}_t\mathbf{f}_t' - \mathbf{I}_K\| = O_p(T^{-1/2})$ and $\max_i \|\mathbf{b}_i\| = O(1)$ imply that the third term is $O_p(T^{-1/2})$.

C.2.2. Proof of corollary 1

Under assumption 3, it can be shown by Bonferroni's method that $\max_{t\leqslant T}\|\mathbf{f}_t\| = O_p\{\log(T)^{1/r_2}\}$. By theorem 4, uniformly in i and t,

$$\begin{aligned}\|\hat{\mathbf{b}}_i'\hat{\mathbf{f}}_t - \mathbf{b}_i'\mathbf{f}_t\| &\leqslant \|\hat{\mathbf{b}}_i - \mathbf{H}\mathbf{b}_i\|\|\hat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t\| + \|\mathbf{H}\mathbf{b}_i\|\|\hat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t\| \\ &\quad + \|\hat{\mathbf{b}}_i - \mathbf{H}\mathbf{b}_i\|\|\mathbf{H}\mathbf{f}_t\| + \|\mathbf{b}_i\|\|\mathbf{f}_t\|\|\mathbf{H}'\mathbf{H} - \mathbf{I}_K\| \\ &= O_p\left[\log(T)^{1/r_2}\sqrt{\left\{\frac{\log(p)}{T}\right\}} + \frac{T^{1/4}}{\sqrt{p}}\right].\end{aligned}$$

C.3. Proof of theorem 2

Lemma 12. $\max_{i\leqslant p}(1/T)\Sigma_{t=1}^T |u_{it} - \hat{u}_{it}|^2 = O_p(\omega_t^2)$, and $\max_{i,t}|u_{it} - \hat{u}_{it}| = o_p(1)$.

Proof. We have $u_{it} - \hat{u}_{it} = \mathbf{b}_i'\mathbf{H}'(\hat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t) + (\hat{\mathbf{b}}_i' - \mathbf{b}_i'\mathbf{H}')\hat{\mathbf{f}}_t + \mathbf{b}_i'(\mathbf{H}'\mathbf{H} - \mathbf{I}_K)\mathbf{f}_t$. Therefore, using the inequality $(a+b+c)^2 \leqslant 4a^2 + 4b^2 + 4c^2$, we have

$$\begin{aligned}\max_{i\leqslant p}\frac{1}{T}\sum_{t=1}^T (u_{it} - \hat{u}_{it})^2 &\leqslant 4\max_i \|\mathbf{b}_i'\mathbf{H}'\|^2 \frac{1}{T}\sum_{t=1}^T \|\hat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t\|^2 \\ &\quad + 4\max_i \|\hat{\mathbf{b}}_i' - \mathbf{b}_i'\mathbf{H}'\|^2 \frac{1}{T}\sum_{t=1}^T \|\hat{\mathbf{f}}_t\|^2 + 4\max_i \|\mathbf{b}_i\|^2 \frac{1}{T}\sum_{t=1}^T \|\mathbf{f}_t\|^2 \|\mathbf{H}'\mathbf{H} - \mathbf{I}_K\|_{\mathrm{F}}^2.\end{aligned}$$

The first part of lemma 12 then follows from theorem 4 and lemma 10. The second part follows from corollary 1.

C.3.1. Completion of proof of theorem 2

Theorem 2 follows immediately from theorem 5 and lemma 12.

C.4. Proof of theorem 3

Define

$$\mathbf{C}_T = \hat{\mathbf{\Lambda}} - \mathbf{B}\mathbf{H}'.$$

Lemma 13.

(a) $\|\mathbf{C}_T\|_{\mathrm{F}}^2 = O_p(\omega_T^2 p)$, and $\|\mathbf{C}_T'\mathbf{C}_T\|_\Sigma^2 = O_p(\omega_T^4 p)$.
(b) $\|\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}} - \mathbf{\Sigma}_u\|_\Sigma^2 = O_p(\omega_T^{2-2q} m_p^2)$.
(c) $\|\mathbf{B}\mathbf{H}'\mathbf{C}_T'\|_\Sigma^2 = O_p(\omega_T^2)$.
(d) $\|\mathbf{B}(\mathbf{H}'\mathbf{H} - \mathbf{I}_K)\mathbf{B}'\|_\Sigma^2 = O_p\{p^{-2} + (pT)^{-1}\}$.

Proof.

(a) We have $\|\mathbf{C}_T\|_{\mathrm{F}}^2 \leqslant \max_{i\leqslant p}\|\hat{\mathbf{b}}_i - \mathbf{H}\mathbf{b}_i\|^2 p = O_p(\omega_T^2 p)$. Moreover, since all the eigenvalues of $\mathbf{\Sigma}$ are bounded away from 0, for any matrix $\mathbf{A}$, $\|\mathbf{A}\|_\Sigma^2 = O_p(p^{-1})\|\mathbf{A}\|_{\mathrm{F}}^2$. Hence $\|\mathbf{C}_T'\mathbf{C}_T\|_\Sigma^2 = O_p(p^{-1} \times \|\mathbf{C}_T\|_{\mathrm{F}}^4) = O_p(p\omega_T^4)$.
(b) By theorem 2, $\|\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}} - \mathbf{\Sigma}_u\|_\Sigma^2 = O_p(p^{-1}\|\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}} - \mathbf{\Sigma}_u\|_{\mathrm{F}}^2) = O_p(\|\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}} - \mathbf{\Sigma}_u\|^2) = O_p(\omega_T^{2-2q} m_p^2)$.
(c) The same argument for the proof of theorem 2 in Fan *et al.* (2008) implies that $\|\mathbf{B}'\mathbf{\Sigma}^{-1}\mathbf{B}\| = O(1)$. Thus, $\|\mathbf{B}\mathbf{H}'\mathbf{C}_T'\|_\Sigma^2 = p^{-1}\mathrm{tr}(\mathbf{H}'\mathbf{C}_T'\mathbf{\Sigma}^{-1}\mathbf{C}_T\mathbf{H}\mathbf{B}'\mathbf{\Sigma}^{-1}\mathbf{B})$ is upper bounded by $p^{-1}\|\mathbf{H}\|^2\|\mathbf{B}'\mathbf{\Sigma}^{-1}\mathbf{B}\|\|\mathbf{\Sigma}^{-1}\| \times \|\mathbf{C}_T\|_{\mathrm{F}}^2 = O_p(p^{-1}\|\mathbf{C}_T\|_{\mathrm{F}}^2) = O_p(\omega_T^2)$.
(d) Again, by $\|\mathbf{B}'\mathbf{\Sigma}^{-1}\mathbf{B}\| = O(1)$, and lemma 11,

$$\begin{aligned}\|\mathbf{B}(\mathbf{H}'\mathbf{H} - \mathbf{I}_K)\mathbf{B}'\|_\Sigma^2 &= p^{-1}\mathrm{tr}\{(\mathbf{H}'\mathbf{H} - \mathbf{I}_K)\mathbf{B}'\mathbf{\Sigma}^{-1}\mathbf{B}(\mathbf{H}'\mathbf{H} - \mathbf{I}_K)\mathbf{B}'\mathbf{\Sigma}^{-1}\mathbf{B}\} \\ &\leqslant p^{-1}\|\mathbf{H}'\mathbf{H} - \mathbf{I}_K\|_{\mathrm{F}}^2\|\mathbf{B}'\mathbf{\Sigma}^{-1}\mathbf{B}\|^2 = O_p\{p^{-2} + (pT)^{-1}\}. \qquad \text{(C.5)}\end{aligned}$$

C.4.1. Proof of theorem 3, part (a)
By lemma 13, $\|\mathbf{B}(\mathbf{H}'\mathbf{H}-\mathbf{I}_K)\mathbf{B}'\|_\Sigma^2+\|\mathbf{B}\mathbf{H}'\mathbf{C}_T'\|_\Sigma^2+\|\mathbf{C}_T\mathbf{C}_T'\|_\Sigma^2=O_p\{\omega_T^2+p\log^2(p)/T^2\}$. Hence, for a generic constant $C>0$,

$$\begin{aligned}\|\hat{\mathbf{\Sigma}}_{\hat{K}}-\mathbf{\Sigma}\|_\Sigma^2 &\leqslant C\|\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}'-\mathbf{B}\mathbf{B}'\|_\Sigma^2+C\|\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}}-\mathbf{\Sigma}_u\|_\Sigma^2\\ &\leqslant C\{\|\mathbf{B}(\mathbf{H}'\mathbf{H}-\mathbf{I}_K)\mathbf{B}'\|_\Sigma^2+\|\mathbf{B}\mathbf{H}'\mathbf{C}_T{}'\|_\Sigma^2+\|\mathbf{C}_T\mathbf{C}_T{}'\|_\Sigma^2+\|\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}}-\mathbf{\Sigma}_u\|_\Sigma^2\}\\ &= O_p\left\{\omega_T^{2-2q}m_p^2+\frac{p\log^2(p)}{T^2}\right\}.\end{aligned}$$

Lemma 14. $\|\hat{\mathbf{\Lambda}}'(\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}})^{-1}\hat{\mathbf{\Lambda}}-(\mathbf{B}\mathbf{H}')'\mathbf{\Sigma}_u^{-1}\mathbf{B}\mathbf{H}'\|=O_p(p\omega_T^{1-q}m_p)$.

Proof. $\|\mathbf{C}_T\|_{\mathrm{F}}^2=O_p(\omega_T^2p)$. Hence

$$\begin{aligned}\|\hat{\mathbf{\Lambda}}'(\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}})^{-1}\hat{\mathbf{\Lambda}}-(\mathbf{B}\mathbf{H}')'\mathbf{\Sigma}_u^{-1}\mathbf{B}\mathbf{H}'\| &\leqslant \|\mathbf{C}_T'(\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}})^{-1}\mathbf{C}_T\|+2\|\mathbf{C}_T'(\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}})^{-1}\mathbf{B}\mathbf{H}'\|\\ &\quad+\|\mathbf{B}\mathbf{H}'((\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}})^{-1}-\mathbf{\Sigma}_u^{-1})\mathbf{B}\mathbf{H}'\|=O_p(p\omega_T^{1-q}m_p).\end{aligned}\tag{C.6}$$

Lemma 15. If $\omega_T^{1-q}m_p=o(1)$, then with probability approaching 1, for some $c>0$,

(a) $\lambda_{\min}\{\mathbf{I}_K+(\mathbf{B}\mathbf{H}')'\mathbf{\Sigma}_u^{-1}\mathbf{B}\mathbf{H}'\}\geqslant cp$,
(b) $\lambda_{\min}\{\mathbf{I}_K+\hat{\mathbf{\Lambda}}'(\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}})^{-1}\hat{\mathbf{\Lambda}}\}\geqslant cp$,
(c) $\lambda_{\min}(\mathbf{I}_K+\mathbf{B}'\mathbf{\Sigma}_u^{-1}\mathbf{B})\geqslant cp$ and
(d) $\lambda_{\min}\{(\mathbf{H}\mathbf{H}')^{-1}+\mathbf{B}'\mathbf{\Sigma}_u^{-1}\mathbf{B}\}\geqslant cp$.

Proof.

(a) By lemma 11, with probability approaching 1, $\lambda_{\min}(\mathbf{H}\mathbf{H}')$ is bounded away from 0. Hence,

$$\begin{aligned}\lambda_{\min}\{\mathbf{I}_K+(\mathbf{B}\mathbf{H}')'\mathbf{\Sigma}_u^{-1}\mathbf{B}\mathbf{H}'\} &\geqslant \lambda_{\min}\{(\mathbf{B}\mathbf{H}')'\mathbf{\Sigma}_u^{-1}\mathbf{B}\mathbf{H}'\}\\ &\geqslant \lambda_{\min}(\mathbf{\Sigma}_u^{-1})\,\lambda_{\min}(\mathbf{H}\mathbf{B}'\mathbf{B}\mathbf{H}')\geqslant \lambda_{\min}(\mathbf{\Sigma}_u^{-1})\,\lambda_{\min}(\mathbf{B}'\mathbf{B})\,\lambda_{\min}(\mathbf{H}\mathbf{H}')\geqslant cp.\end{aligned}$$

(b) The result follows from part (a) and lemma 14. Parts (c) and (d) follow from a similar argument to that for part (a) and lemma 11.

C.4.2. Completion of proof of theorem 3
We derive the rate for $\|\hat{\mathbf{\Sigma}}_{\hat{K}}^{-1}-\mathbf{\Sigma}^{-1}\|$. Define

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

Note that $\hat{\mathbf{\Sigma}}_{\hat{K}}=\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}'+\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}}$ and $\mathbf{\Sigma}=\mathbf{B}\mathbf{B}'+\mathbf{\Sigma}_u$. The triangular inequality gives

$$\|\hat{\mathbf{\Sigma}}_{\hat{K}}^{-1}-\mathbf{\Sigma}^{-1}\|\leqslant\|\hat{\mathbf{\Sigma}}_{\hat{K}}^{-1}-\tilde{\mathbf{\Sigma}}^{-1}\|+\|\tilde{\mathbf{\Sigma}}^{-1}-\mathbf{\Sigma}^{-1}\|.$$

Using the Sherman–Morrison–Woodbury formula, we have $\|\hat{\mathbf{\Sigma}}_{\hat{K}}^{-1}-\tilde{\mathbf{\Sigma}}^{-1}\|\leqslant\sum_{i=1}^6L_i$, where

$$\left.\begin{aligned}L_1&=\|(\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}})^{-1}-\mathbf{\Sigma}_u^{-1}\|,\\ L_2&=\|((\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}})^{-1}-\mathbf{\Sigma}_u^{-1})\hat{\mathbf{\Lambda}}\{\mathbf{I}_K+\hat{\mathbf{\Lambda}}'(\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}})^{-1}\hat{\mathbf{\Lambda}}\}^{-1}\hat{\mathbf{\Lambda}}'(\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}})^{-1}\|,\\ L_3&=\|((\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}})^{-1}-\mathbf{\Sigma}_u^{-1})\hat{\mathbf{\Lambda}}\{\mathbf{I}_K+\hat{\mathbf{\Lambda}}'(\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}})^{-1}\hat{\mathbf{\Lambda}}\}^{-1}\hat{\mathbf{\Lambda}}'\mathbf{\Sigma}_u^{-1}\|,\\ L_4&=\|\mathbf{\Sigma}_u^{-1}(\hat{\mathbf{\Lambda}}-\mathbf{B}\mathbf{H}')\{\mathbf{I}_K+\hat{\mathbf{\Lambda}}'(\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}})^{-1}\hat{\mathbf{\Lambda}}\}^{-1}\hat{\mathbf{\Lambda}}'\mathbf{\Sigma}_u^{-1}\|,\\ L_5&=\|\mathbf{\Sigma}_u^{-1}(\hat{\mathbf{\Lambda}}-\mathbf{B}\mathbf{H}')\{\mathbf{I}_K+\hat{\mathbf{\Lambda}}'(\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}})^{-1}\hat{\mathbf{\Lambda}}\}^{-1}\mathbf{H}\mathbf{B}'\mathbf{\Sigma}_u^{-1}\|,\\ L_6&=\|\mathbf{\Sigma}_u^{-1}\mathbf{B}\mathbf{H}'(\{\mathbf{I}_K+\hat{\mathbf{\Lambda}}'(\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}})^{-1}\hat{\mathbf{\Lambda}}\}^{-1}-(\mathbf{I}_K+\mathbf{H}\mathbf{B}'\mathbf{\Sigma}_u^{-1}\mathbf{B}\mathbf{H}')^{-1})\mathbf{H}\mathbf{B}'\mathbf{\Sigma}_u^{-1}\|.\end{aligned}\right\}\tag{C.7}$$

We bound each of the six terms. First, L_1 is bounded by theorem 2. Let $\mathbf{G}=\{\mathbf{I}_K+\hat{\mathbf{\Lambda}}'(\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}})^{-1}\hat{\mathbf{\Lambda}}\}^{-1}$; then

$$L_2\leqslant\|(\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}})^{-1}-\mathbf{\Sigma}_u^{-1}\|\,\|\hat{\mathbf{\Lambda}}\mathbf{G}\hat{\mathbf{\Lambda}}'\|\,\|(\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}})^{-1}\|.$$

Note that theorem 2 implies that $\|(\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}})^{-1}\|=O_p(1)$. Lemma 15 then implies that $\|\mathbf{G}\|=O_p(p^{-1})$. This shows that $L_2=O_p(L_1)$. Similarly $L_3=O_p(L_1)$. In addition, since $\|\mathbf{C}_T\|_{\mathrm{F}}^2=\|\hat{\mathbf{\Lambda}}-\mathbf{B}\mathbf{H}'\|_{\mathrm{F}}^2=O_p(\omega_T^2p)$, $L_4\leqslant\|\mathbf{\Sigma}_u^{-1}(\hat{\mathbf{\Lambda}}-\mathbf{B}\mathbf{H}')\|\,\|\mathbf{G}\|\,\|\hat{\mathbf{\Lambda}}'\mathbf{\Sigma}_u^{-1}\|=O_p(\omega_T)$. Similarly $L_5=O_p(L_4)$. Finally, let $\mathbf{G}_1=\{\mathbf{I}_K+(\mathbf{B}\mathbf{H}')'\mathbf{\Sigma}_u^{-1}\mathbf{B}\mathbf{H}'\}^{-1}$.

By lemma 15, $\|\mathbf{G}_1\| = O_p(p^{-1})$. Then, by lemma 14,

$$\begin{aligned}\|\mathbf{G}-\mathbf{G}_1\| &= \|\mathbf{G}(\mathbf{G}^{-1}-\mathbf{G}_1^{-1})\mathbf{G}_1\| \leqslant O_p(p^{-2})\|(\mathbf{BH}')'\boldsymbol{\Sigma}_u^{-1}\mathbf{BH}' - \hat{\boldsymbol{\Lambda}}'(\hat{\boldsymbol{\Sigma}}_{u,\hat{K}}^{\mathcal{T}})^{-1}\hat{\boldsymbol{\Lambda}}\| \\ &= O_p(p^{-1}\omega_T^{1-q}m_p).\end{aligned}$$

Consequently, $L_6 \leqslant \|\boldsymbol{\Sigma}_u^{-1}\mathbf{BH}'\|^2\|\mathbf{G}-\mathbf{G}_1\| = O_p(\omega_T^{1-q}m_p)$. Adding up L_1–L_6 gives

$$\|\hat{\boldsymbol{\Sigma}}_{\hat{K}}^{-1} - \tilde{\boldsymbol{\Sigma}}^{-1}\| = O_p(\omega_T^{1-q}m_p).$$

However, using the Sherman–Morrison–Woodbury formula again implies that

$$\begin{aligned}\|\tilde{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1}\| &\leqslant \|\boldsymbol{\Sigma}_u^{-1}\mathbf{B}(\{(\mathbf{H}'\mathbf{H})^{-1} + \mathbf{B}'\boldsymbol{\Sigma}_u^{-1}\mathbf{B}\}^{-1} - (\mathbf{I}_K + \mathbf{B}'\boldsymbol{\Sigma}_u^{-1}\mathbf{B})^{-1})\mathbf{B}'\boldsymbol{\Sigma}_u^{-1}\| \\ &\leqslant O(p)\|\{(\mathbf{H}'\mathbf{H})^{-1} + \mathbf{B}'\boldsymbol{\Sigma}_u^{-1}\mathbf{B}\}^{-1} - (\mathbf{I}_K + \mathbf{B}'\boldsymbol{\Sigma}_u^{-1}\mathbf{B})^{-1}\| \\ &= O_p(p^{-1})\|(\mathbf{H}'\mathbf{H})^{-1} - \mathbf{I}_K\| = o_p(\omega_T^{1-q}m_p).\end{aligned}$$

C.4.3. Completion of proof of theorem 3: $\|\hat{\boldsymbol{\Sigma}}^{\mathcal{T}} - \boldsymbol{\Sigma}\|_{\max}$

We first bound $\|\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}' - \mathbf{BB}'\|_{\max}$. Repeatedly using the triangular inequality yields

$$\begin{aligned}\|\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}' - \mathbf{BB}'\|_{\max} &= \max_{i,j\leqslant p}|\hat{\mathbf{b}}_i'\hat{\mathbf{b}}_j - \mathbf{b}_i'\mathbf{b}_j| \\ &\leqslant \max_{ij}\{|(\hat{\mathbf{b}}_i - \mathbf{H}\mathbf{b}_i)'\hat{\mathbf{b}}_j| + |\mathbf{b}_i'\mathbf{H}'(\hat{\mathbf{b}}_j - \mathbf{H}\mathbf{b}_j)| + |\mathbf{b}_i'(\mathbf{H}'\mathbf{H} - \mathbf{I}_K)\mathbf{b}_j|\} \\ &\leqslant (\max_i\|\hat{\mathbf{b}}_i - \mathbf{H}\mathbf{b}_i\|)^2 + 2\max_{ij}\|\hat{\mathbf{b}}_i - \mathbf{H}\mathbf{b}_i\|\|\mathbf{H}\mathbf{b}_j\| + \max_i\|\mathbf{b}_i\|^2\|\mathbf{H}'\mathbf{H} - \mathbf{I}_K\| \\ &= O_p(\omega_T).\end{aligned}$$

However, let $\sigma_{u,ij}$ be the (i, j) entry of $\boldsymbol{\Sigma}_u$. Then $\max_{ij}|\hat{\sigma}_{ij} - \sigma_{u,ij}| = O_p(\omega_T)$.

$$\max_{ij}|s_{ij}(\hat{\sigma}_{ij}) - \sigma_{u,ij}| \leqslant \max_{ij}|s_{ij}(\hat{\sigma}_{ij}) - \hat{\sigma}_{ij}| + |\hat{\sigma}_{ij} - \sigma_{u,ij}| \leqslant \max_{ij}\tau_{ij} + O_p(\omega_T) = O_p(\omega_T).$$

Hence $\|\hat{\boldsymbol{\Sigma}}_{u,\hat{K}}^{\mathcal{T}} - \boldsymbol{\Sigma}_u\|_{\max} = O_p(\omega_T)$. The result then follows immediately.

References

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

Ahn, S., Lee, Y. and Schmidt, P. (2001) GMM estimation of linear panel data models with time-varying individual effects. *J. Econmetr.*, **101**, 219–255.

Alessi, L., Barigozzi, M. and Capassoc, M. (2010) Improved penalization for determining the number of factors in approximate factor models. *Statist. Probab. Lett.*, **80**, 1806–1813.

Amini, A. A. and Wainwright, M. J. (2009) High-dimensional analysis of semidefinite relaxations for sparse principal components. *Ann. Statist.*, **37**, 2877–2921.

Antoniadis, A. and Fan, J. (2001) Regularized wavelet approximations. *J. Am. Statist. Ass.*, **96**, 939–967.

d'Aspremont, A., Bach, F. and El Ghaoui, L. (2008) Optimal solutions for sparse principal component analysis. *J. Mach. Learn. Res.*, **9**, 1269–1294.

Athreya, K. and Lahiri, S. (2006) *Measure Theory and Probability Theory*. New York: Springer.

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

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

Bai, J. and Ng, S. (2008) Large dimensional factor analysis. *Found. Trends Econmetr.*, **3**, 89–163.

Bai, J. and Shi, S. (2011) Estimating high dimensional covariance matrices and its applications. *Ann. Econ. Finan.*, **12**, 199–215.

Bickel, P. and Levina, E. (2008) Covariance regularization by thresholding. *Ann. Statist.*, **36**, 2577–2604.

Birnbaum, A., Johnstone, I., Nadler, B. and Paul, D. (2012) Minimax bounds for sparse PCA with noisy high-dimensional data. *Ann. Statist.*, to be published.

Boivin, J. and Ng, S. (2006) Are more data always better for factor analysis? *J. Econmetr.*, **132**, 169–194.

Cai, J., Candès, E. and Shen, Z. (2008) A singular value thresholding algorithm for matrix completion. *SIAM J. Optimizn*, **20**, 1956–1982.

Cai, T. and Liu, W. (2011) Adaptive thresholding for sparse covariance matrix estimation. *J. Am. Statist. Ass.*, **106**, 672–684.
Cai, T. and Zhou, H. (2012) Optimal rates of convergence for sparse covariance matrix estimation. *Ann. Statist.*, **40**, 2389–2420.
Candès, E., Li, X., Ma, Y. and Wright, J. (2011) Robust principal component analysis? *J. Ass. Comput. Mach.*, **58**, 3.
Carvalho, C., Chang, J., Lucas, J., Nevins, J., Wang, Q. and West, M. (2008) High-dimensional sparse factor modeling: applications in gene expression genomics. *J. Am. Statist. Ass.*, **103**, 1438–1456.
Chamberlain, G. and Rothschild, M. (1983) Arbitrage, factor structure and mean-variance analysis in large asset markets. *Econometrica*, **51**, 1305–1324.
Davis, C. and Kahan, W. (1970) The rotation of eigenvectors by a perturbation III. *SIAM J. Numer. Anal.*, **7**, 1–46.
Doz, C., Giannone, D. and Reichlin, L. (2011) A two-step estimator for large approximate dynamic factor models based on Kalman filtering. *J. Econmetr.*, **164**, 188–205.
Efron, B. (2007) Correlation and large-scale simultaneous significance testing. *J. Am. Statist. Ass.*, **102**, 93–103.
Efron, B. (2010) Correlated z-values and the accuracy of large-scale statistical estimates. *J. Am. Statist. Ass.*, **105**, 1042–1055.
Fama, E. and French, K. (1992) The cross-section of expected stock returns. *J. Finan.*, **47**, 427–465.
Fan, J., Fan, Y. and Lv, J. (2008) High dimensional covariance matrix estimation using a factor model. *J. Econmetr.*, **147**, 186–197.
Fan, J., Han, X. and Gu, W. (2012) Control of the false discovery rate under arbitrary covariance dependence (with discussion). *J. Am. Statist. Ass.*, **107**, 1019–1048.
Fan, J., Liao, Y. and Mincheva, M. (2011a) High dimensional covariance matrix estimation in approximate factor models. *Ann. Statist.*, **39**, 3320–3356.
Fan, J., Liao, Y. and Mincheva, M. (2011b) Large covariance estimation by thresholding principal orthogonal complements. *Preprint arxiv.org/pdf/1201.0175.pdf.*
Fan, J., Zhang, J. and Yu, K. (2012) Vast portfolio selection with gross-exposure constraints. *J. Am. Statist. Ass.*, **107**, 592–606.
Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2000) The generalized dynamic factor model: identification and estimation. *Rev. Econ. Statist.*, **82**, 540–554.
Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2004) The generalized dynamic factor model consistency and rates. *J. Econmetr.*, **119**, 231–255.
Forni, M. and Lippi, M. (2001) The generalized dynamic factor model: representation theory. *Econmetr. Theor.*, **17**, 1113–1141.
Fryzlewicz, P. (2012) High-dimensional volatility matrix estimation via wavelets and thresholding. *Manuscript*. London School of Economics and Political Science, London.
Hallin, M. and Liška, R. (2007) Determining the number of factors in the general dynamic factor model. *J. Am. Statist. Ass.*, **102**, 603–617.
Hallin, M. and Liška, R. (2011) Dynamic factors in the presence of blocks. *J. Econmetr.*, **163**, 29–41.
Hastie, T. J., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edn. New York: Springer.
James, W. and Stein, C. (1961) Estimation with quadratic loss. In *Proc. 4th Berkeley Symp. Mathematical Statistics and Probability*, vol. 1, pp. 361–379. Berkeley: University of California Press.
Johnstone, I. M. (2001) On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.*, **29**, 295–327.
Johnstone, I. M. and Lu, A. Y. (2009) On consistency and sparsity for principal components analysis in high dimensions. *J. Am. Statist. Ass.*, **104**, 682–693.
Jung, S. and Marron, J. S. (2009) PCA consistency in high dimension, low sample size context. *Ann. Statist.*, **37**, 4104–4130.
Kapetanios, G. (2010) A testing procedure for determining the number of factors in approximate factor models with large datasets. *J. Bus. Econ. Statist.*, **28**, 397–409.
Lam, C. and Fan, J. (2009) Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.*, **37**, 4254–4278.
Lawley, D. and Maxwell, A. (1971) *Factor Analysis as a Statistical Method*, 2nd edn. London: Butterworth.
Leek, J. and Storey, J. (2008) A general framework for multiple testing dependence. *Proc. Natn. Acad. Sci. USA*, **105**, 18718–18723.
Lin, Z., Ganesh, A., Wright, J., Wu, L., Chen, M. and Ma, Y. (2009) Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix. *Int. Wrkshp Computational Advances in Multi-Sensor Adaptive Processing, Aruba*.
Luo, X. (2011) High dimensional low rank and sparse covariance matrix estimation via convex minimization. *Manuscript*. University of Pennsylvania, Philadelphia.
Ma, Z. (2013) Sparse principal components analysis and iterative thresholding. *Ann. Statist*., to be published.
Meinshausen, N. and Bühlmann, P. (2006) High dimensional graphs and variable selection with the Lasso. *Ann. Statist.*, **34**, 1436–1462.

Onatski, A. (2010) Determining the number of factors from empirical distribution of eigenvalues. *Rev. Econ. Statist.*, **92**, 1004–1016.
Pati, D., Bhattacharya, A., Pillai, N. and Dunson, D. (2012) Posterior contraction in sparse Bayesian factor models for massive covariance matrices. *Manuscript*. Duke University, Durham.
Paul, D. (2007) Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sin.*, **17**, 1617–1642.
Pesaran, M. H. (2006) Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*, **74**, 967–1012.
Pesaran, M. H. and Yamagata, T. (2012) Testing CAPM with a large number of assets. *American Finance Association San Diego Meetings Paper*. (Available from `http://ssrn.com/abstracts`.)
Phan, Q. (2012) On the sparsity assumption of the idiosyncratic errors covariance matrix—Support from the FTSE 100 stock returns. *Manuscript*. University of Warwick, Coventry.
Ross, S. A. (1976) The arbitrage theory of capital asset pricing. *J. Econ. Theor.*, **13**, 341–360.
Rothman, A., Levina, E. and Zhu, J. (2009) Generalized thresholding of large covariance matrices. *J. Am. Statist. Ass.*, **104**, 177–186.
Sentana, E. (2009) The econometrics of mean-variance efficiency tests: a survey. *Econometr. J.*, **12**, 65–101.
Sharpe, W. (1964) Capital asset prices: a theory of market equilibrium under conditions of risks. *J. Finan.*, **19**, 425–442.
Shen, H. and Huang, J. (2008) Sparse principal component analysis via regularized low rank matrix approximation. *J. Multiv. Anal.*, **99**, 1015–1034.
Stock, J. and Watson, M. (1998) Diffusion Indexes. *Working Paper 6702*. National Bureau of Economic Research, Cambridge.
Stock, J. and Watson, M. (2002) Forecasting using principal components from a large number of predictors. *J. Am. Statist. Ass.*, **97**, 1167–1179.
Witten, D. M., Tibshirani, R. and Hastie, T. (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, **10**, 515–534.
Wright, J., Peng, Y., Ma, Y., Ganesh, A. and Rao, S. (2009) Robust principal component analysis: exact recovery of corrupted low-rank matrices by convex optimization. *New. Inform. Process. Syst*.
Xiong, H., Goulding, E. H., Carlson, E. J., Tecott, L. H., McCulloch, C. E. and Sen, S. (2011) A flexible estimating equations approach for mapping function-valued traits. *Genetics*, **189**, 305–316.
Yap, J. S., Fan, J. and Wu, R. (2009) Nonparametric modeling of longitudinal covariance structure in functional mapping of quantitative trait loci. *Biometrics*, **65**, 1068–1077.
Zhang, Y. and El Ghaoui, L. (2011) Large-scale sparse principal component analysis with application to text data. *New. Inform. Process. Syst*.

Discussion on the paper by Fan, Liao and Mincheva

Marc Hallin (*Université Libre de Bruxelles and Princeton University*)

Fan and his colleagues are dealing with finite realizations

$$\begin{matrix} Y_{11}, & Y_{12}, & \ldots, & Y_{1T} \\ \vdots & \vdots & & \vdots \\ Y_{p1}, & Y_{p2}, & \ldots, & Y_{pT} \end{matrix}$$

of double-indexed stochastic processes of the form $\{Y_{it} | i \in \mathbb{N},\ t \in \mathbb{Z}\}$. Such realizations can be seen either as a collection of p one-dimensional time series, related to p individuals or cross-sectional items, or as one observed time series $\{\mathbf{Y}_t^{(p)} := (Y_{1t}, \ldots, Y_{pt})'\}$ in dimension p. The objective is the estimation of the covariance matrix $\mathbf{\Sigma}$ of $\mathbf{Y}_t^{(p)}$. As both p and T are 'large', (p, T) asymptotics are appropriate. Owing to the effect of a common environment (unobserved 'factors' or 'common shocks' generating a large number of cross-covariances), the traditional assumption of a sparse $\mathbf{\Sigma}$ is unlikely to hold. However, the same assumption becomes reasonable once the effect of unobserved factors or common shocks has been removed. The main challenge, thus, consists in removing that effect—which is highly non-trivial, as those factors or common shocks are not observed and are not even well defined.

Factor model methods in such context are tailor-made solutions. The general idea behind factor models consists in decomposing each Y_{it} into $Y_{it} = \chi_{it} + \xi_{it}$ where $\{\chi_{it}\}$ (**Y**'s *common component*, accounting for the effect of the *factors*) and $\{\xi_{it}\}$ (**Y**'s *idiosyncratic component*, on which a sparsity assumption is to be made) are unobserved mutually orthogonal (at all leads and lags) processes. Further identification constraints on $\{\chi_{it}\}$ and $\{\xi_{it}\}$ of course are needed, yielding a variety of factor models. The better $\{\chi_{it}\}$ is at accounting for the effect of common factors (cross-correlations), the more plausible the sparsity assumption on the covariance of the ξ_{it}.

The identification assumption that was chosen by the authors requires χ_{it} to be of the form $\chi_{it} = \mathbf{b}_i'\mathbf{F}_t$, where $\{\mathbf{F}_t =: (F_{1t}, \ldots, F_{Kt})'\}$ is some latent K-dimensional vector process (the factors)—yielding a *static* (approximate) factor model of the type studied by Chamberlain and Rothschild (1983), Bai and Ng (2002), Stock and Watson (2002a, b), and many others. The authors then successfully propose a principal-component-based method for reconstructing $\{\chi_{it}\}$ and $\{\xi_{it}\}$, followed by a thresholding procedure for the estimation of $\mathbf{\Sigma}_\xi^{(p)}$, and derive powerful consistency results, with rates that depend on the sparsity of $\mathbf{\Sigma}_\xi^{(p)}$.

This is a path breaking contribution to the literature on high dimensional covariance estimation. The authors should be congratulated for this, and I have no hesitation in proposing a vote of thanks. However, 'Sans la liberté de blâmer, il n'est point déloge flatteur' ('Praising has no value in the absence of free criticism' (Beaumarchais, *Le Mariage de Figaro*)), and I would like to enhance my praising of the authors' work with a couple of friendly critical comments.

Principal components—the traditional *static* ones, computed from the 'instantaneous' covariances of the Y_{it}s—are a fundamental tool in the authors' approach. Since Brillinger (1981), however, it is widely admitted that static principal components are not the adequate concept of principal components in a time series context. By maximizing normed linear combinations of the form $\Sigma_{i=1}^{p} a_i Y_{it}$, indeed, they completely overlook serial dependences. A static principal component with a small eigenvalue may have a negligible contemporaneous effect on $\mathbf{Y}_t$, but a significant effect on $\mathbf{Y}_{t+1}$, and hence a high predictive value: discarding it, as does the static principal component method, shifts its contribution to the idiosyncratic component, possibly jeopardizing assumed idiosyncratic sparsity.

Static principal components of course are fine under the assumptions of the static factor model, which in turn are pertinent in the presence of independent observations (the type that the authors probably have in mind despite the time series setting of their paper: gene expressions, financial returns, etc.). They are no longer adequate in the presence of serial dependence, and this is an indication that the *static* factor model assumptions are unlikely to hold in a genuine time series context. The weak point of static factor assumptions is the fact that all factors are to be loaded contemporaneously at time t whereas, in most practical situations, factors are loaded with lags. A general form for the common component is $\chi_{it} = \mathbf{b}_i'(L)\mathbf{F}_t$ instead of $\chi_{it} = \mathbf{b}_i'\mathbf{F}_t$ with $K \times 1$ loading filters $\mathbf{b}_i(L)$ instead of the $K \times 1$ loading matrices $\mathbf{b}_i$ (equivalently, $\chi_{it} = \mathbf{b}_i'(L)\mathbf{U}_t$, where $\mathbf{U}_t = (U_{1t}, \ldots, U_{Kt})'$ is a K-tuple of mutually orthogonal white noises, the *common shocks*). Adopting this *dynamic* characterization of χ_{it} leads to the *general dynamic factor* that was studied in Forni *et al*. (2000) or Forni and Lippi (2001). An important advantage of that dynamic model is that, in contrast with the static one, it holds, basically, without any assumption (but second-order stationarity) on $\{Y_{it}\}$.

With this dynamic specification replacing the static model, the idiosyncratic covariance matrices, as well as the lagged idiosyncratic cross-covariance matrices, are much more likely to satisfy sparsity assumptions. Brillinger's *dynamic principal components* can be used (Forni *et al*., 2000), very much in the same way as the static principal components in the static model, to reconstruct the decomposition $Y_{it} = \chi_{it} + \xi_{it}$, then applying the thresholding technique that is recommended by the authors. (Dynamic principal components are based on the maximization of linear combinations of the form $\Sigma_{i=1}^{p} \Sigma_{k=-\infty}^{\infty} a_{ik} Y_{i,t-k}$ involving the past, present and future values of Y_{it} and can be computed from the eigenvalues and eigenvectors of the spectral density matrices of $\{Y_{it}\}$.) The resulting consistency rates will involve the sparsity of the dynamic factor idiosyncratic covariance matrix rather than that of its static factor counterpart. And, the same methods as proposed by the authors naturally apply with the more ambitious objective of estimating the full (high dimensional) autocovariance structure of the observed series.

Piotr Fryzlewicz and Na Huang (*London School of Economics and Political Science*)

We would like to start by congratulating Professor Fan, Dr Liao and Ms Mincheva for the stimulating and thought-provoking paper.

The POET-estimator is the sum of two parts: the non-sparse, low rank part resulting from the factor model, and the sparse part arising as a result of thresholding the 'principal orthogonal complement'. The estimator has been designed with a particular factor model in mind, and therefore it is natural to ask, firstly, whether and how one could verify this model assumption and, secondly, whether POET offers acceptable performance if the assumption does not hold.

We may be wrong here, but we are unaware of a reliable technique for estimating the number of factors K which works well except in the most 'textbook' cases of the first few eigenvalues being 'visibly' larger than others. Even if a factor structure is present, the presence of both stronger and less strong factors may lead to the cut-off in the eigenvalues being less obvious, in which case any inference for the number of factors may not be reliable. However, it is important to choose K correctly from the point of view of

the usability of POET: the authors warn us that POET may perform poorly if K is underestimated. It is therefore tempting to ask whether POET may benefit from averaging over K as a possible guard against picking one 'wrong' (e.g. underestimated) value of K. Averaging may also be beneficial in cases when the factor model assumption is not satisfied.

An appealing aspect of the construction of POET is the inclusion of the non-sparse part (which is done in case the target matrix Σ is not sparse) and the sparse part (to ensure the invertibility of the estimator). It is tempting to consider other possible estimators along these lines. Motivated by POET, we propose an estimator of Σ of the form

$$\hat{\Sigma}^N = \delta\hat{\Sigma}_{\text{sam}} + (1-\delta)\, t(\hat{\Sigma}_{\text{sam}}, \lambda),$$

where $\hat{\Sigma}_{\text{sam}}$ is the $p \times p$ sample covariance matrix, δ is a constant in [0,1], λ is a $p \times p$ matrix with non-negative entries and $t(\cdot, \cdot)$ is a function that applies soft, hard or other thresholding to each non-diagonal entry of its first argument, with the threshold value equal to the corresponding entry of its second argument. λ will typically be parameterized by one scalar parameter. Obviously, $\delta\hat{\Sigma}_{\text{sam}}$ and $(1-\delta)\, t(\hat{\Sigma}_{\text{sam}}, \lambda)$ are the non-sparse and sparse components respectively.

$\hat{\Sigma}^N$ performs 'shrinkage of the sample covariance towards a sparse target'. To the best of our knowledge, $\hat{\Sigma}^N$ is a new proposal, although shrinkage towards some other targets has been studied extensively before, notably by Ledoit and Wolf (2003), who proposed shrinkage towards a one-factor target and Schaefer and Strimmer (2005), who reviewed and discussed six commonly used targets. Some ideas for the 'optimal' choice of δ are proposed by Ledoit and Wolf (2003) and Schaefer and Strimmer (2005) and can be adopted in the context of $\hat{\Sigma}^N$, thereby reducing the number of 'free' parameters of the procedure to the single scalar parameter of the threshold matrix λ. If all new covariance estimators were required to have 'literary' names (such as POET), we would name ours 'NOVELIST', for 'novel integration of the sample and thresholded covariance estimators'. The benefits of NOVELIST include simplicity, ease of implementation and the fact that its application avoids eigenanalysis, which is unfamiliar to many practitioners.

We now briefly exhibit the performance of POET *versus* NOVELIST on a simulated covariance matrix Σ of size 100×100, available from `http://stats.lse.ac.uk/fryzlewicz/testcov.RData` (use `load("testcov.RData")` in R; the variable name is `testcov`). Σ was not generated from a factor model and is not sparse. The range of its diagonal elements is [3.32, 7.09], and only 56 of the non-diagonal entries are larger than 1 in absolute value. The sample size is $n = 100$, so $\hat{\Sigma}_{\text{sam}}$ itself is not invertible. In NOVELIST, we use both $\delta = 0$ and $\delta = \frac{1}{2}$, and the constant matrix $\lambda \equiv 1$. In POET, we use $K = 7$, following the authors' advice, given in the R package `POET`, to choose a large K (to avoid issues related to K being underestimated), but preferably smaller than 8. Both POET and NOVELIST use soft thresholding. Other POET parameters are set to default values. The data are Gaussian, and there are $N = 100$ repetitions. Table 7 shows the results. POET performs poorly for Σ: it is the worst in all norms by a large margin. NOVELIST with $\delta = 0$ (which reduces to the simple soft thresholding estimator) and with $\delta = \frac{1}{2}$ are difficult to tell apart in terms of their performance. However, as far Σ^{-1} is concerned, NOVELIST with $\delta = \frac{1}{2}$ is the best, followed by POET and then by the simple soft thresholding. The overall clear 'winner' in this example is NOVELIST with $\delta = \frac{1}{2}$.

Table 7. Averaged (and rounded except max-) distances to Σ (left-hand section) and Σ^{-1} (right-hand section) for $\hat{\Sigma}^N$ with $\delta = 0$, with $\delta = \frac{1}{2}$ and for the POET-estimator, in the L_∞-, Frobenius, max- and spectral norms†

Norm	Σ			Σ^{-1}		
	$\delta = 0$	$\delta = \frac{1}{2}$	*POET*	$\delta = 0$	$\delta = \frac{1}{2}$	*POET*
L_∞	*34*	*34*	61	42	*38*	41
Frobenius	*30*	32	50	34	*30*	33
max	2.09	*2.03*	2.29	4.88	*3.47*	3.92
L_2	10	*9*	19	17	*15*	17

†Distances to Σ^{-1} were multiplied by 10 before averaging. The best results are in italics.

By way of summary, POET is an elegant construction which combines parsimony of representation in the low rank component with sparsity in the thresholded part. This brief discussion

(a) attempts to list some research questions regarding POET which we believe are worth exploring further and
(b) proposes a simple competitor.

We found the paper a pleasure to read and thought it was written in a clear and pedagogical way. We are convinced that POET will stimulate further research in the important field of large covariance estimation. It therefore gives us great pleasure to second the vote of thanks for this paper.

The vote of thanks was passed by acclamation.

Wenyang Zhang (*University of York*) **and Heng Peng** (*Hong Kong Baptist University*)
We congratulate Professor Fan, Dr Liao and Ms Mincheva for such a brilliant paper. We believe that this paper will have huge influence on the estimation of covariance matrices of large size and will stimulate many further researches in this topic.

The condition of sparsity is often imposed when it comes to large matrix estimation. As the authors rightly point out sparsity may not be appropriate in some circumstances; the covariance matrix of asset returns used in portfolio allocation is an example. It is better to impose some kind of structure on the covariance matrix on the basis of the data concerned. The matrices with the structure stated in this paper are very general and appear in many research areas, such as portfolio allocation, risk management and image analysis. The estimation that is proposed in this paper is intuitive and easy to implement; it will become very popular in the areas where large matrix estimation is needed.

Among the many clever and stimulating ideas in this paper, we particularly appreciate the connection between principal component analysis and factor models. This connection leads to a brand new estimation of factor loadings in factor analysis and of the covariance matrix of idiosyncratic components.

The estimation of the covariance matrix of idiosyncratic components is based on

$$\hat{\mathbf{R}}_K = \sum_{i=K+1} \hat{\lambda}_i \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i'.$$

When p is larger than T, the sample covariance matrix would be singular, and some $\hat{\lambda}_i$ will be 0 when $i > K+1$. This problem will become more acute when $p \gg T$. How would this affect the estimator of the covariance matrix of idiosyncratic components? Would some kind of iteration improve the accuracy of the estimator? For example: treat $\hat{\mathbf{R}}_K^{\mathcal{T}}$ as an initial estimator of Σ_u, decompose $\hat{\Sigma}_{\text{sam}} - \hat{\mathbf{R}}_k^{\mathcal{T}}$ by its principal components and denote the sum of the first K terms in the decomposition by $\mathcal{F}$ and apply the thresholding rule on $\hat{\Sigma}_{\text{sam}} - \mathcal{F}$ to obtain improved $\hat{R}_K^{\mathcal{T}}$, and continue this iterative procedure until convergence.

The presence of spiked eigenvalues is clearly formulated in this paper; however, when the factor loadings are not available, which is the case in reality, how do we check whether there are spiked eigenvalues? Would it work simply by checking whether there is a jump among the eigenvalues of the sample covariance matrix?

As far as the estimation of the covariance matrix Σ is concerned, is it really necessary to have the condition of the presence of spiked eigenvalues? Would $\hat{\Sigma}_K$ always work regardless of whether the spiked eigenvalues exist or not? We may have missed some important points on this issue.

The selection of K on the basis of the Akaike or Bayesian information criterion does not seem to make use of the information about the presence of spiked eigenvalues; is there any room to improve the selection by incorporating this information in the selection procedure?

Alexei Onatski (*University of Cambridge*)
My comments on this interesting paper are confined to its central assumption that the first K eigenvalues of $\mathbf{B}\text{cov}(\mathbf{f}_t)\mathbf{B}'$ diverge at rate $O(p)$, whereas all the eigenvalues of Σ_u are bounded as $p \to \infty$. This *factor pervasiveness* assumption implies that $\Sigma_{i=1}^K \hat{\lambda}_i \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i'$ and $\Sigma_{i=k+1}^p \hat{\lambda}_i \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i'$ are good approximations to the sample covariances of the systematic and idiosyncratic components of the data respectively, which is a key to the success of POET. Unfortunately, it may be misleading in many economic and financial applications.

For example, as Fig. 11 shows, except for $i = 1$, there are no large gaps between eigenvalues i and $i+1$ of the sample covariance matrix of the excess return data that were used in Section 6. However, since, as is commonly believed, such data contain at least three factors, the factor pervasiveness assumption suggests the existence of a large gap for $i \geqslant 3$.

The absence of the gap between 'systematic' and 'idiosyncratic' eigenvalues may have a negative effect

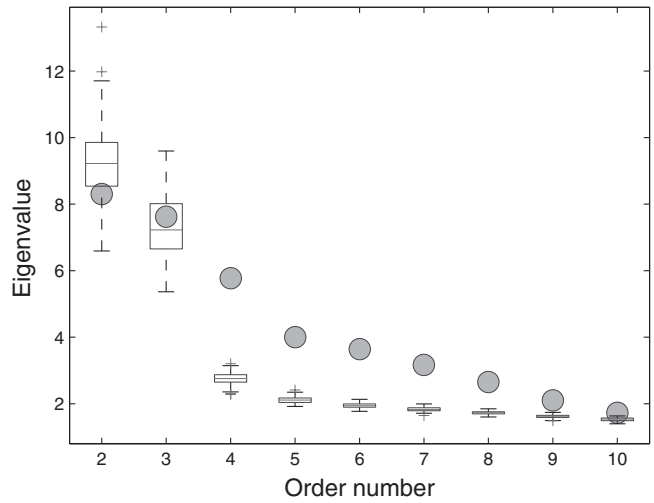

Fig. 11. Sample covariance eigenvalues of 100 industrial portfolio data (○) and of data simulated from the factor model calibrated as in Section 6 (boxplots based on 100 replications)

Table 8. Performance of POET and shrinkage when the systematic–idiosyncratic eigenvalue gap becomes small

	Estimator	*Results for the following values of* ρ *and* $\hat{\lambda}_3/\hat{\lambda}_4$*:*					
		$\rho=0.4$, $\lambda_3/\lambda_4=3.1$	$\rho=0.5$, $\lambda_3/\lambda_4=2.7$	$\rho=0.6$, $\lambda_3/\lambda_4=2.6$	$\rho=0.7$, $\lambda_3/\lambda_4=2.2$	$\rho=0.8$, $\lambda_3/\lambda_4=1.5$	$\rho=0.9$, $\lambda_3/\lambda_4=1.1$
$\|\hat{\Sigma}_{u3}-\Sigma_u\|$	POET	0.74	1.04	1.57	2.10	3.89	8.98
	Shrinkage	1.21	1.63	2.18	2.96	4.68	9.24
$\|\hat{\Sigma}_{u3}^{-1}-\Sigma_u^{-1}\|$	POET	7.38	10.1	18.7	43.2	$>10^3$	$>10^3$
	Shrinkage	6.16	9.63	14.1	22.2	39.2	90.7
$\|\hat{\Sigma}_3-\Sigma\|_\Sigma$	POET	0.85	0.90	1.07	1.33	2.14	6.65
	Shrinkage	0.97	1.05	1.11	1.16	1.20	1.27
$\|\hat{\Sigma}_3^{-1}-\Sigma^{-1}\|$	POET	7.09	9.64	17.8	39.0	$>10^3$	$>10^3$
	Shrinkage	4.35	5.67	7.47	10.5	16.9	41.9
$\|\hat{\Sigma}_3-\Sigma\|$	POET	20.2	20.4	20.4	21.3	19.7	21.2
	Shrinkage	20.3	20.4	20.4	21.3	19.7	21.1

on the performance of POET. Table 8 reports the mean quality of POET over 1000 replications of data simulated as in Section 6.2, but with $\sigma_{u,ij}=\rho^{|i-j|}$. As ρ increases from 0.4 to 0.9, the systematic–idiosyncratic gap measured by $\hat{\lambda}_3/\hat{\lambda}_4$ decreases from 3.7 to 1.1. For the 100 industrial portfolios data that were used in Section 6, $\hat{\lambda}_3/\hat{\lambda}_4=1.32$, which is best matched by the simulated data with $\rho=0.8$. The quality of POET dramatically deteriorates in the neighbourhood of $\rho=0.8$.

For comparison, in rows of Table 8 marked 'shrinkage', I report the quality of the estimator that replaces POET's thresholding step by Ledoit and Wolf's (2004) linear shrinkage procedure applied to the principal orthogonal complement. The deterioration of the quality of parameter shrinkage is not as dramatic as that of POET.

Continuing the comparison, I computed the risk of portfolios created as in Section 7 with parameter shrinkage. The minimum risk portfolio that was created with shrinkage had lower variance than that created with POET 51% of the time. Among those months, the risk was decreased by 19%. During the

months that POET produced a lower risk portfolio, the risk was decreased by 15%. These results indicate that POET does not dominate other simple covariance matrix estimation methods in applications where there is no clear gap between systematic and idiosyncratic eigenvalues. Developing covariance estimation methods that would work well in such situations is an important task for future research.

Clifford Lam and Charlie Hu (*London School of Economics and Political Science*)

We congratulate Fan and his colleagues for this insightful paper. Here we suggest a method to address two concerns:

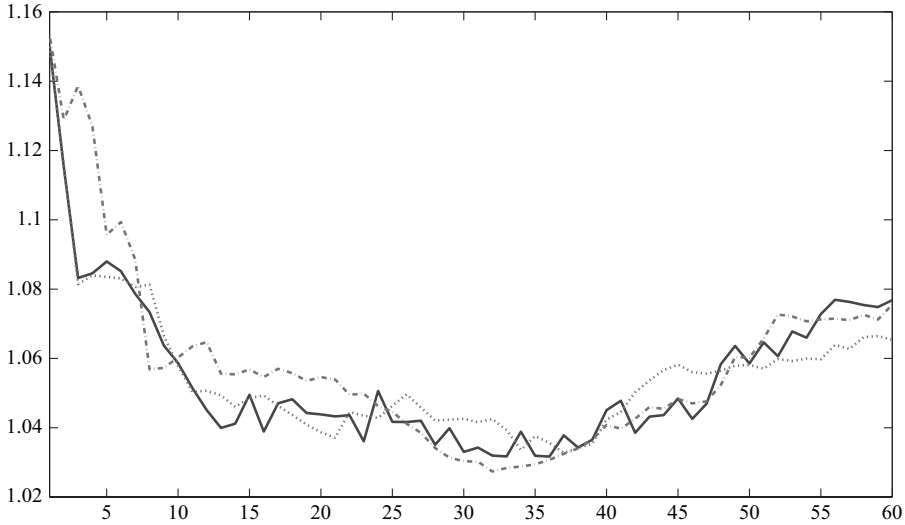

Fig. 12. Average forecast errors for various numbers of factors r: ———, autocovariance-based factor modelling; · · · · · ·, Lam *et al.* (2011); ·-·-·-·, principal component analysis

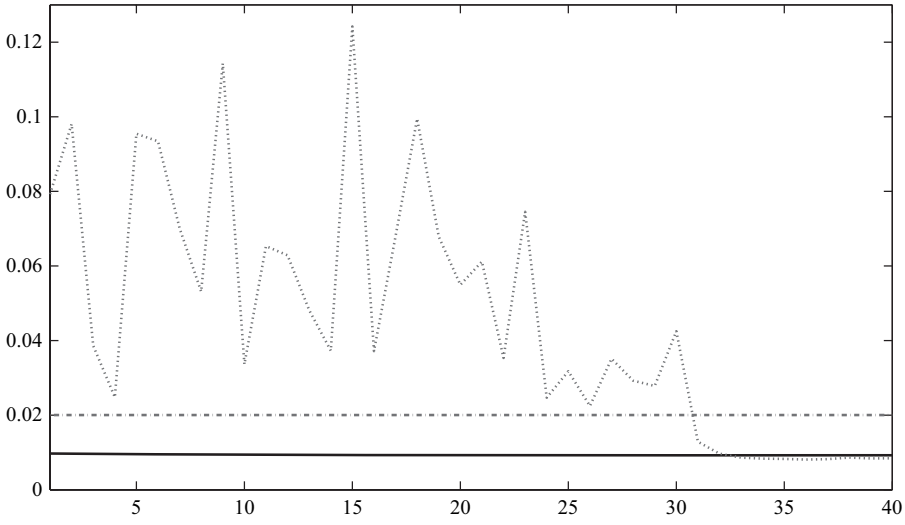

Fig. 13. Sum of absolute bias (averaged over 100 simulations) for estimating β by using generalized least squares (·-·-·-·) against the number of factors r used in POET ($C = 0.5$) (· · · · · ·) and the condition-number-regularized estimator (———) (the bias for the least square method is constant throughout)

(a) the potential underestimation of the number of factors K;
(b) the potential non-sparseness of the estimated principal orthogonal complement.

Point (a) is addressed by using a larger K. With pervasive factors assumed in the paper, it is relatively easy to find such K. However, in an analysis of macroeconomic data for example, there can be a mix of pervasive factors and many weaker ones; see Chudik *et al.* (2011), Lam *et al.* (2011) and Lam and Yao (2012) for a general definition of weak factors.

In Stock and Watson (2005), monthly data of $p = 132$ US macroeconomic time series from 1959 to 2003 ($n = 526$) were analysed. Using principal component analysis (Bai and Ng, 2002) the method in Lam *et al.* (2011) and a modified version called autocovariance-based factor modelling, we compute the average forecast errors of 30 monthly forecasts by using a vector auto-regressive model VAR(3) on the estimated factors from these methods with different number of factors r (Fig. 12). Whereas three pervasive factors decrease forecast errors sharply, including more factors, up to $r = 35$, decreases forecast errors more slowly, showing the existence of many 'weaker' factors.

Hence it is not always possible to have 'enough' factors for accurate thresholding of the principal orthogonal complement, which can still include contributions from many weak factors and is not sparse. Points (a) and (b) can therefore be closely related, and can be addressed if we regularize the condition number of the orthogonal complement instead of thresholding. Whereas Won *et al.* (2013)

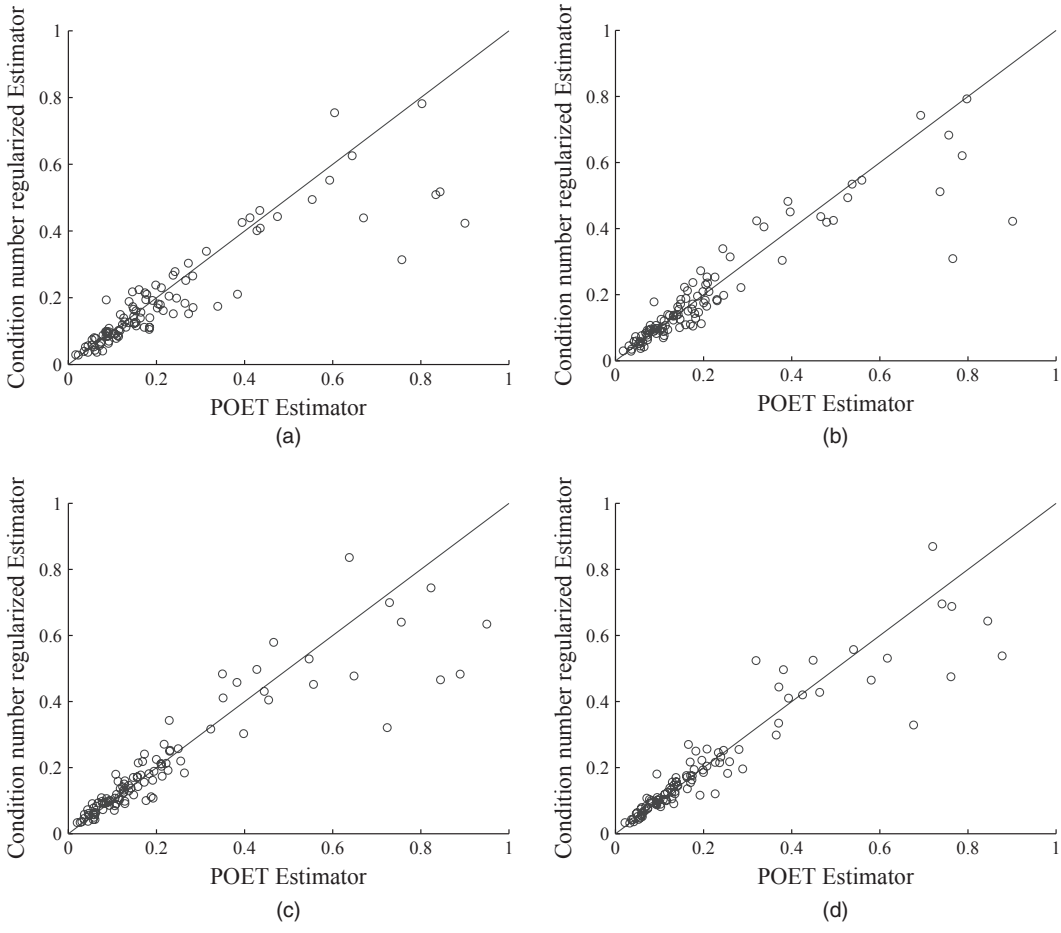

Fig. 14. Risk of portfolios created with POET ($C = 0.5$) and the condition-number-regularized estimator: (a) $K = 1$; (b) $K = 2$; (c) $K = 3$; (d) $K = 4$

Table 9. Comparisons of the risks of portfolios by using POET ($C = 0.5$ and the condition-number-regularized estimator

K	*Proportion of time POET outperforms*	*% of average risk improvements*
1	0.40	−4.07
2	0.46	−2.50
3	0.56	0.66
4	0.56	0.71

restrict the extreme eigenvalues with a tuning parameter to be chosen, we use the idea of Abadir *et al.* (2010) (the properties have not been investigated sufficiently unfortunately). We are studying its theoretical properties.

We simulate 100 times from the panel regression model

$$\mathbf{y}_t = \mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\beta} = (-0.5, 0.5, 0.3, -0.6)^{\mathrm{T}},$$

with x_{it} being independent AR(1) processes and $\boldsymbol{\varepsilon}_t$ the standardized macroeconomic data in Stock and Watson (2005) plus independent $N(0, 0.2)$ noise. Following example 5 of the paper, we estimate $\boldsymbol{\Sigma}_{\varepsilon}^{-1}$ by using different methods and plot the sum of absolute bias for estimating $\boldsymbol{\beta}$ by using generalized least squares against the number of factors r used in Fig. 13. Clearly regularizing on condition number leads to stabler estimators.

Parallel to Section 7.2, we compare the risk of portfolios created by using POET and the method above. Again Fig. 14 and Table 9 show stabler performance of regularization on condition number.

Natalia Bailey (*University of Cambridge*), **M. Hashem Pesaran** (*University of Southern California, Los Angeles, and Trinity College, Cambridge*) **and Takashi Yamagata** (*University of York*)

The paper's key contribution lies in tackling the problem of estimation of a large covariance matrix, $\boldsymbol{\Sigma} = (\sigma_{ij})_{p\times p}$, when the data set examined is 'contaminated' with strong unobserved factors. Both cross-sectional and time dimensions (p, T) are assumed large. Fan and his colleagues extend existing literature on regularization of such a matrix via thresholding when this is strictly sparse—e.g. Bickel and Levina (2008) and Cai and Liu (2011). Their proposed estimator, POET, is applied to the covariance matrix of residuals extracted from a regression of the original data $\mathbf{y}_t$ on K estimated factors. The errors $\mathbf{u}_t$ are considered to be weakly cross-sectionally dependent (i.e. $\boldsymbol{\Sigma}_u = (\sigma_{u,ij})$ is sparse or $\|\boldsymbol{\Sigma}_u\|_1 = O(1)$).

The main result of the paper is the order condition obtained for the norm of the deviation of POET from the true value of $\boldsymbol{\Sigma}_u$:

$$\|\hat{\boldsymbol{\Sigma}}_{u,\hat{K}}^{\mathrm{T}} - \boldsymbol{\Sigma}_u\| = O_p\left(m_p\left[\frac{1}{\sqrt{p}} + \sqrt{\left\{\frac{\log(p)}{T}\right\}}\right]^{1-q}\right), \tag{1}$$

where $q \in [0, 1)$. $\hat{\boldsymbol{\Sigma}}_{u,\hat{K}}^{\mathrm{T}}$ is the thresholded version of $\boldsymbol{\Sigma}_u$, and $m_p = \max_{i\leqslant p}\sum_{j\leqslant p} I_{(\sigma_{u,ij}\neq 0)}$ when $q = 0$. They find the same rate (l) for $\|(\hat{\boldsymbol{\Sigma}}_{u,\hat{K}}^{\mathrm{T}})^{-1} - \boldsymbol{\Sigma}_u^{-1}\|$, ensuring positive definiteness of $\hat{\boldsymbol{\Sigma}}_{u,\hat{K}}^{\mathrm{T}}$ by setting a lower bound on their thresholding parameter. Using this result they then suggest that $(\hat{\boldsymbol{\Sigma}}_{u,\hat{K}}^{\mathrm{T}})^{-1}$ can be used in various applications in finance, in particular in their example 6 where they consider testing $\boldsymbol{\alpha} = \mathbf{0}$, in the linear asset pricing model

$$\mathbf{y}_t = \boldsymbol{\alpha} + \mathbf{B}\mathbf{f}_t + \mathbf{u}_t. \tag{2}$$

It is claimed that use of $(\hat{\boldsymbol{\Sigma}}_{u,\hat{K}}^{\mathrm{T}})^{-1}$ in the development of such a test will be valid even when $p \gg T$. However, as shown in Pesaran and Yamagata (2012) this cannot be so even if the rate condition (1) holds. Pesaran and Yamagata (2012) conduct a test of $H_0 : \boldsymbol{\alpha} = \mathbf{0}$ using as generating process (2) and propose a test statistic which, under normality of $\mathbf{u}_t$, can be written as

$$J(\boldsymbol{\Sigma}_u) = \frac{(\boldsymbol{\tau}_T^{\mathrm{T}}\mathbf{M}_F\boldsymbol{\tau}_T)\hat{\boldsymbol{\alpha}}^{\mathrm{T}}\boldsymbol{\Sigma}_u^{-1}\hat{\alpha} - p}{\sqrt{(2p)}} \xrightarrow[\mathrm{d}]{} N(0, 1) \tag{3}$$

Table 10. Size and power of the $J(\Sigma_u)$ test in the case of models with three factors†

	T	*Results for* $p=50$		*Results for* $p=100$		*Results for* $p=200$	
		Size	*Power*	*Size*	*Power*	*Size*	*Power*
$J(\hat{\Sigma}_u^{\mathrm{T}})$	60	0.14	0.78	0.19	0.91	0.25	0.98
	100	0.08	0.93	0.11	0.98	0.14	1.00
$J(\hat{\Sigma}_u^{\mathrm{LW}})$	60	0.14	0.62	0.18	0.78	0.25	0.92
	100	0.11	0.85	0.13	0.92	0.17	0.98
J^{PY}	60	0.05	0.58	0.04	0.74	0.04	0.89
	100	0.05	0.87	0.05	0.95	0.05	1.00

†Errors are weakly cross-sectionally dependent and normally distributed. Sparseness of $\mathbf{\Sigma}_u$ is defined as in Table 3 of Pesaran and Yamagata (2012) with $\delta_b = \frac{1}{4}$. Size: $\alpha_i = 0$ for all $i = 1, \ldots, p$. Power: $\alpha_i \sim \mathrm{IIDN}(0,1)$ for $i = 1, 2, \ldots, p_\alpha$, $p_\alpha = [p\ 0.8]$; otherwise $\alpha_i = 0$. The number of replications is set to 2000. 'Hard' thresholding in $\hat{\mathbf{\Sigma}}_u^{\mathrm{T}}$ is conducted by using cross-validation.

as $p \to \infty$ for any fixed $T \geqslant K+1$, where $\boldsymbol{\tau}_T$ is a $T \times 1$ vector of 1s and $\mathbf{M}_{\mathrm{F}} = \mathbf{I}_T - \mathbf{F}(\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}'$, $\mathbf{F} = (\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_T)'$. When $\mathbf{\Sigma}_u$ is known the test is valid for any $T > K+1$ but, if an estimator of $\mathbf{\Sigma}_u^{-1}$ is inserted in expression (3), then $J(\hat{\mathbf{\Sigma}}_u) \to_{\mathrm{d}} N(0,1)$ only if $p \log(p)/T \to 0$, which requires $p < T$.

To illustrate this point we conducted a small Monte Carlo simulation following the set-up in Pesaran and Yamagata (2012) where we plugged in $(\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathrm{T}})^{-1}$ and the Ledoit and Wolf (2004) shrinkage estimator, $(\hat{\mathbf{\Sigma}}_u^{\mathrm{LW}})^{-1}$, as estimates of $\mathbf{\Sigma}_u^{-1}$ in expression (3) for a set of (p, T) combinations. As shown in Table 10, considerable size distortions are visible when either estimator is used for $p > T$. Size improves only when T increases.

To overcome this problem Pesaran and Yamagata (2012) propose the following simple test statistic that ignores the off-diagonal elements of $\mathbf{\Sigma}_u$:

$$J^{\mathrm{PY}} = \frac{p^{-1/2} \sum_{i=1}^{p} \left(t_i^2 - \dfrac{\upsilon}{\upsilon - 2} \right)}{\left(\dfrac{\upsilon}{\upsilon - 2} \right) \sqrt{\left[\dfrac{2(\upsilon - 1)}{\upsilon - 4} \{1 + (p-1)\hat{\bar{\rho}}^2\} \right]}}, \tag{4}$$

where $\upsilon = T - K - 1$, and t_i denotes the standard t-ratio of α_i in the ordinary least squares regression of individual asset returns, and

$$\hat{\bar{\rho}}^2 = \frac{2}{p(p-1)} \sum_{i=2}^{p} \sum_{j=1}^{i-1} \hat{\rho}_{ij}^2 I(\upsilon \hat{\rho}_{ij}^2 \geqslant \theta_p), \tag{5}$$

$\hat{\rho}_{ij} = \hat{\mathbf{u}}_{i.}' \hat{\mathbf{u}}_{j.} / \sqrt{\{(\hat{\mathbf{u}}_{i.}' \hat{\mathbf{u}}_{i.})(\hat{\mathbf{u}}_{j.}' \hat{\mathbf{u}}_{j.})\}}$, $I(\cdot)$ is an indicator function and the threshold value θ_p is chosen such that $\Pr(\rho_{ij} \neq 0)$ declines steadily with p. Size and power for this test are also summarized in Table 10 and show the ability of the test to control the size well with high power even if $p \gg T$.

Cinzia Viroli (*University of Bologna*)

This paper is very interesting and rich with thought-provoking themes. I would like to comment on the double formulation of the estimation problem for large covariance matrices.

Fan and his colleagues assume that the data have been generated according to an approximate K-factor model and suggest recovery of the covariance matrix via its decomposition into a low rank matrix and a sparse error matrix. They first address this issue by resorting to the first K principal components to estimate the factor loadings and to the thresholded principal orthogonal complement for the estimation of the idiosyncratic variance.

They then show that the estimator has an equivalent representation by using a constrained least squares method. For each given vector F of factor scores they derive the least squares estimator of the factor

loadings (as a function of F) and hence estimate the factor scores. Given both the factor scores and the factor loadings they recover the idiosyncratic factors, derive their covariance matrix and suitably threshold it.

One of the advantages of the least squares formulation is that it generalizes and extends, to the unknown factor case, the results that the authors have obtained for known factors, both assuming a strict (Fan *et al*., 2008) and an approximate (Fan *et al*., 2011) factor model. It is also coherent with a large part of the literature on the topic and allows us to use consolidated proof strategies and model selection tools.

However, by pursuing the least squares approach the authors seem somewhat to neglect the principal components analysis results. The way that they solve the least squares problem (first obtain the loadings and then the factor scores) is natural when the factors are known and the loadings unknown, but when both of them are unknown the dual problem (first scores and then loadings) could be as meaningful, as in Stock and Watson (2002b). This would allow us to obtain the factor loadings as the eigenvectors of the sample covariance matrix directly, thus leading to the POET-estimator, making theorem 1 in Section 2 no longer needed. Also, the principal components analysis approach would allow us to obtain the idiosyncratic errors simply as the scores in the principal components orthogonal complement with no need to estimate the common factor scores first. I have really appreciated the paper, but I cannot help noting that along the paper the poesy of POET somewhat fades. I am wondering whether there is some drawback of the principal components analysis approach that justifies the least squares prevalence.

Angela Montanari (*University of Bologna*)

This is a very interesting paper, full of stimuli in all its parts.

One of the most relevant aspects is the identification of sparsness of the idiosyncratic covariance matrix as a sufficient condition (together with factor pervasiveness) for the identifiability of an approximate factor model and for its estimation through principal components analysis (PCA).

This result, besides further justifying Chamberlain and Rothschild's (1983) result (which required limited idiosyncratic eigenvalues only) also provides a clear indication on the kind of dependence structure which the model can capture. And this is very important from an empirical point of view, as the examples in Section 5 clearly show.

Sparsness of the idiosyncratic covariance (together with diverging p) also offers PCA a sort of ground for revenge as an estimation method for a factor model. For finite p, and under the strict factor model assumption, it is well known that PCA performs poorly as an estimation method, since it generates correlated errors; but for diverging p, and under the approximate factor model assumption, this paper shows that PCA represents a natural and theoretically grounded estimation instrument. I would speak, for PCA, of a blessing of dimensionality.

Within this coherent framework, anyway, I feel a little uneasy with the empirical application in Section 7.

50 series, related to as many stocks (chosen from five different industry sectors), and their annualized daily returns for $T = 252$ days are considered. Fan and his colleagues identify three relevant factors, they estimate the factor loadings through the first three PCs of the sample covariance matrix and finally obtain the thresholded error correlation matrix. This matrix shows that a strong positive correlation between the returns of companies in the same industry is still present after taking out the common factors and from this the authors conclude that it provides strong evidence that the strict factor model is inappropriate.

In this case p is not very large with respect to T. My feeling is that we are still dealing with the finite p situation in which, as already said, PCA returns correlated idiosyncratic errors, even when they are actually uncorrelated. In other words, I am wondering whether the residual correlation is evidence of the inappropriateness of a strict factor model or, on the contrary, it is simply induced by an inappropriate use of PCA. If the factor loadings had been estimated by any of the estimation methods ordinarily used in classical factor analysis, would the residual correlations still be non-vanishing?

Yi Yu and Richard J. Samworth (*University of Cambridge*)

We congratulate the authors on their paper. POET elegantly tackles low rank plus sparse matrix estimation, provided that the eigenvalues of the low rank matrix grow at rate $O(p)$ (see assumption 1). Suppose now that this assumption does not hold, and instead, we have the following condition.

Assumption 5. All the eigenvalues of the $K \times K$ matrix $p^{-\alpha}\mathbf{B}'\mathbf{B}$ are bounded away from both 0 and ∞ as $p \to \infty$, where $0 < \alpha < 1$.

Similar conditions are widely used in sparse principal components analysis and low rank plus sparse matrix estimation problems; see, for example, Amini and Wainwright (2009) and Agarwal *et al*. (2012). In

Table 11. Perfomance of IC, AIC and BIC†

Methods	*Results for the following values of C:*			
	$C_1 = 1$	$C_1 = \frac{1}{3}$	$C_1 = \frac{1}{10}$	$C_1 = 10$
IC	6.00 (0.00)	1.08 (0.27)	1.00 (0.00)	6.00 (0.00)
AIC	20.00 (0.00)	20.00 (0.00)	20.00 (0.00)	20.00 (0.00)
BIC	6.00 (0.00)	2.00 (0.00)	1.00 (0.00)	6.00 (0.00)

†For the same $\mathbf{u}$ and $\boldsymbol{\mu}_B$ as in Section 6.2, define $\tilde{\boldsymbol{\mu}}_B' = (\boldsymbol{\mu}_B', \boldsymbol{\mu}_B')'$ and expand $\boldsymbol{\Sigma}_B$ to a block diagonal matrix, $\tilde{\boldsymbol{\Sigma}}_B$ by making $\boldsymbol{\Sigma}_B$ the diagonal block of $\tilde{\boldsymbol{\Sigma}}_B$. The rows of $\mathbf{B}_1$ are generated from an $\mathcal{N}_6(\tilde{\boldsymbol{\mu}}_B, \tilde{\boldsymbol{\Sigma}}_B)$ distribution. Expand the generating process of $\mathbf{F}$ similarly to match $\mathbf{B}_1$ and generate $\mathbf{F}_1$ accordingly, and then let $\mathbf{Y} = C_1\mathbf{B}_1\mathbf{F}_1' + \mathbf{u}$. Here, $K = 6$ and $K_{\max} = 20$. The means of the estimated K are reported over 100 repetitions, with standard error in parentheses.

what follows, we consider the three main objectives in Section 2. The notation and model are the same as those in the paper.

Proposition 3. Assume assumption 1. For the factor model with condition (2.1), we have

$$|\lambda_j - \|\tilde{\mathbf{b}}_j\|^2| \leqslant \|\boldsymbol{\Sigma}_u\|, \qquad \text{for } j \leqslant K,$$
$$|\lambda_j| \leqslant \|\boldsymbol{\Sigma}_u\|, \qquad \text{for } j > K.$$

Moreover, if $\{\|\tilde{\mathbf{b}}_j\|\}_{j=1}^K$ are distinct, then

$$\|\boldsymbol{\xi}_j - \tilde{\mathbf{b}}_j/\|\tilde{\mathbf{b}}_j\|\| = O(p^{-\alpha}\|\boldsymbol{\Sigma}_u\|), \qquad \text{for } j \leqslant K.$$

From this we see that, under a suitable sparsity condition on $\boldsymbol{\Sigma}_u$, the first K principal components are still approximately the same as the columns of the factor loadings, even if the eigenvalues are not as spiked as $O(p)$.

However, for POET to control the relative error of the matrix estimate, assumption 1 is necessary, as can be seen from a close inspection of the proof of theorem 2 of Bai and Ng (2002). In fact, if assumption 1 is replaced with assumption 5, we have, for $K' < K$ that

$$\lim_{p,T\to\infty} \mathbb{P}\{\mathrm{IC}(K') < \mathrm{IC}(K)\} > 0.$$

The other half of this theorem still holds, however, so the less spiked structure will not asymptotically increase the risk of overestimation in the selection of K.

The performances of IC, the Akaike information criterion AIC and the Bayesian information criterion BIC are compared in Table 11, with the corresponding largest eigenvalues $\mathbf{YY}'$ in Fig. 15. If the spectrum structure satisfies assumption 1 ($C_1 \geqslant 1$), both IC and BIC select the correct value of K. However, if we shrink the spiked eigenvalues, IC and BIC tend to underestimate, whereas AIC overestimates, the true K.

To examine the effect of missing the Kth common factor, assume condition (2.1) and that rank $(\mathbf{B}'\mathbf{B}) = K$, but the estimator is

$$\hat{\boldsymbol{\Sigma}}_{K-1} = \sum_{i=1}^{K-1} \hat{\lambda}_i \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i' + \hat{\mathbf{R}}_{K-1}^{\mathcal{T}},$$

where $\hat{\mathbf{R}}_{K-1}^{\mathcal{T}}$ is the entrywise shrunk estimator of $\mathbf{R}_{K-1} = \mathbf{b}_K\mathbf{f}_K\mathbf{f}_K'\mathbf{b}_K' + \boldsymbol{\Sigma}_u$. In this case, owing to the common factor, most of the pairs of cross-sectional units in $\mathbf{R}_{K-1}$ are no longer 'weakly correlated'. Note that the $\hat{\theta}_{ij}$s in Appendix A are still the same, i.e. no extra shrinkage is introduced. However, m_p used in theorems 2 and 3 is not $o(p)$, so the error bound does not converge to 0, but, when K is correctly estimated or overestimated, even substituting assumption 5 for assumption 1, the corresponding results in theorems 2 and 3 still hold. Thus, if there is doubt about the validity of assumption 1, a less severe penalty (e.g. AIC) may be preferable, to avoid the more serious error of underestimation of K.

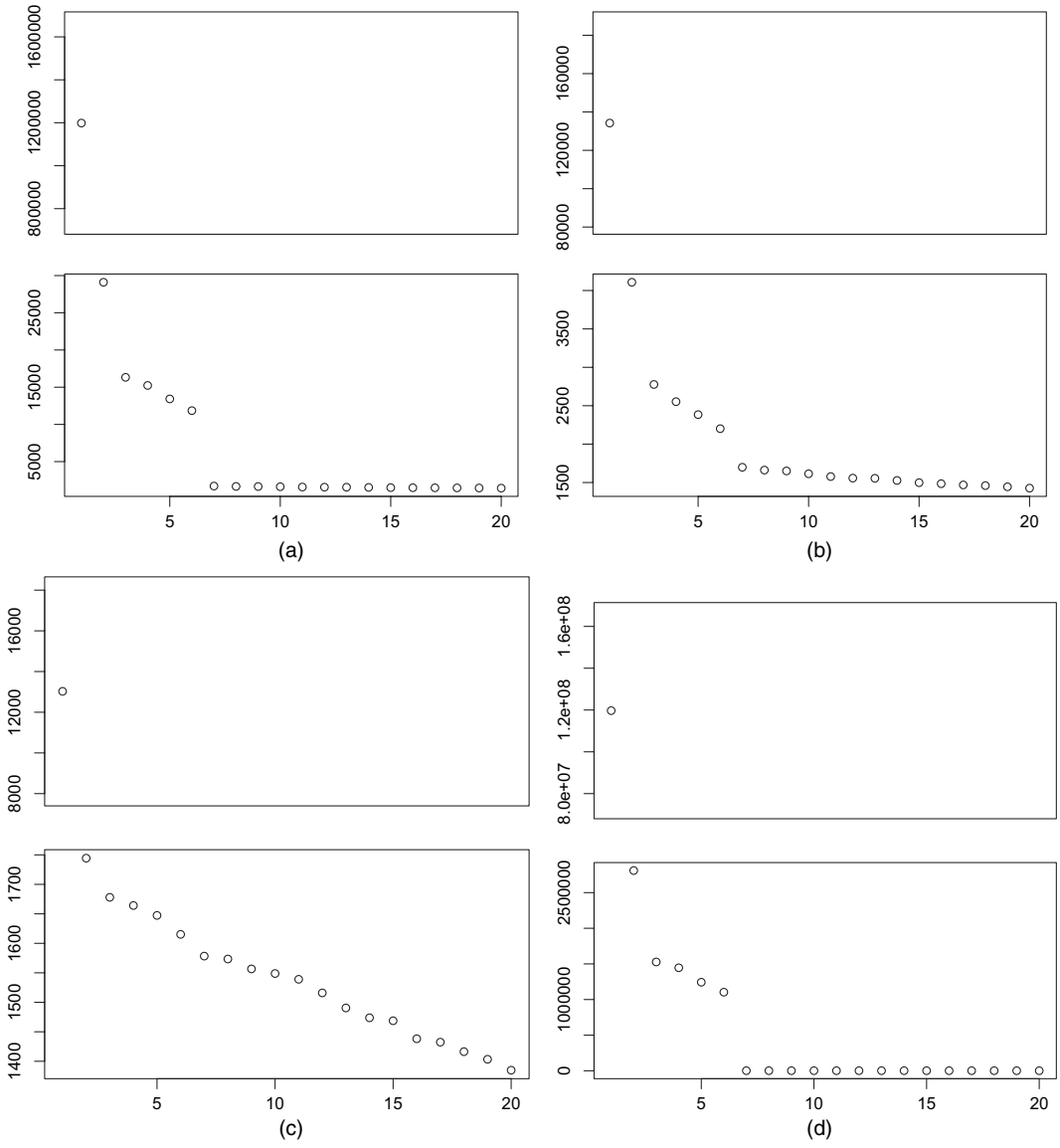

Fig. 15. Largest 20 eigenvalues of $\mathbf{YY}'$ in cases (a) $C_1 = 1$, (b) $C_1 = \frac{1}{3}$, (c) $C_1 = \frac{1}{10}$ and (d) $C_1 = 10$

Frank Critchley (*The Open University, Milton Keynes*)

In warmly welcoming tonight's paper, I offer three sets of comments, adding the aside that there could be value in using measures of discrepancy specifically adapted to non-negative definite symmetric matrices.

(a) Consistent estimation of $\mathbf{\Sigma}$, or its inverse, relies on key *assumptions*—typically, a static model with pervasiveness factors and (approximately) sparse errors. Two key questions arise.
   (i) How far can they be *checked?* Given POET's fast (singular value decomposition) nature, deletion diagnostics seem entirely feasible. Again, individuals and/or time points can be deleted. Further, the static model can be tested within broader—e.g. first-order auto-regressive—models.

(ii) What *effects* do these assumptions have on subsequent inference? Indeed, could such context-specific considerations help to guide the choices to be made within a POET-analysis? In short, is there scope for a 'POET for purpose'?

(b) POET having an equivalent least squares formulation, potential lack of robustness to outliers merits consideration, with the usual array of possible solutions. In particular, there is a variety of robust versions of both principal component analysis and factor analysis.

(c) Might there be a role for the invariant co-ordinate selection (ICS) methodology introduced in Tyler *et al.* (2009)? If so, there would seem to be several possible advantages. ICS requiring two affine equivariant scatter functionals, a robust choice could, for example, be used alongside the regular covariance. Of particular relevance here, I have recently shown that one of these functionals can be singular, without essential loss. ICS could be used as a complement to POET, the former using a generalized form of the principal component analysis asymptotically determining the latter. More radically, it could replace POET, as is perhaps natural on invariance grounds: subsuming centring, ICS's invariance to linear transformation of the data combines principal component analysis's invariance to orthogonal transformation with that of factor analysis to separate scaling of the variables. This would have the additional advantage of potentially extending POET to a wider range of data types. In particular, incommensurable variables could be accommodated. Finally, however ICS is implemented, it retains POET's computational speed, while providing visual displays. These offer a range of diagnostic and other potential benefits, notably, multivariate outlier detection or, again, group detection (via implicit estimation of Fisher's discriminant subspace).

Once again, it is a pleasure to congratulate the authors on a very stimulating paper.

Jian Zhang (*University of Kent, Canterbury*)

I congratulate Fan and his colleagues on their ground breaking and innovative contribution to high dimensional covariance estimation. I would like to contribute to the discussion by the following comments.

Their method is applicable to studying the source localization problem in magnetoencephalography-based neuroimaging. Suppose that we observe a multivariate time course $\mathbf{Y}(t)$ from n sensors, which can be modelled by

$$\mathbf{Y}(t_j) = \sum_{k=1}^{p} \mathbf{x}\{r_k, \eta(r_k)\}\,\beta(r_k, t_j) + \varepsilon(t_j), \qquad j = 1, \ldots, J, \tag{6}$$

where $\Omega = \{r_1, \ldots, r_p\}$ is a grid approximation to the brain, $\beta(r, t)$ is a latent univariate time source of interest at location r, $\mathbf{x}\{r, \eta(r)\}$ is a design vector determined by the so-called Maxwell equations with orientation $\eta(r)$ and $\varepsilon(t)$ is noise. Assume that $\beta(r, t)$ is sparse, i.e. the temporal variability (called power or the marginal variance) $\mathrm{var}\{\beta(r, \cdot)\} = 0$ for all $r \in \Omega$ except a few locations (i.e. non-null sources). We want to localize these non-null sources among an infinite number of candidates. A spatial filtering theory has been developed by Zhang *et al.* (2012) and Zhang (2012) for searching for a non-null source. Under certain conditions, the covariance matrix of $\mathbf{Y}(t)$ can be expressed as $\Sigma_{k=1}^{p} \gamma_k \mathbf{x}_k \mathbf{x}_k^{\mathrm{T}} + \Sigma_0$, where $\mathbf{x}_k$ is the output vector of these sensors that would be induced by a unit magnitude source located at r_k along orientation η_k and γ_k is the power at r_k. So it is not surprising that our theory is relying on an appropriate estimate of the covariance matrix when p is much larger than n and J. In this sense, I expect that POET can significantly improve our spatial filters.

There are two different asymptotics: expanding time domain asymptotics, the time window is expanding when J increases and infill asymptotics, where t_j, $1 \leqslant j \leqslant J$, are restricted to a certain time window when J increases. Under the infill setting, the current strong mixing condition in Section 3 will not hold. I am curious about the performance of POET in the infill setting.

John T. Kent (*University of Leeds*)

It is interesting to compare the methodology of this paper with conventional multivariate analysis in a fixed dimension p, where p is small or moderate and the relevant asymptotics involve the sample size n growing large. The simplest methodology is either invariant or equivariant under affine transformations, e.g. Hotelling's T^2-statistic. Thus if there are two variables, height and weight, say, it does not matter what units we use to measure them; further the original two variables can be replaced by any two linearly independent linear combinations.

However, even in this simple setting, there is typically less invariance for dimension reduction methods. For example, principal component analysis is equivariant only under orthogonal transformations of the data. Factor analysis is equivariant only under diagonal transformations, i.e. rescaling the variables.

For large p, especially when p exceeds the sample size n, some sort of regularization is needed; the price is stronger assumptions and less invariance. In the current paper it is assumed that

(a) the factor loadings are pervasive and
(b) the idiosyncratic covariance matrix Σ_u is sparse.

My impression is that these assumptions imply that the methodology of the paper is not equivariant under either orthogonal or diagonal transformations. If so, there are several limitations on the types of data for which this paper might be useful:

(a) the variables themselves (rather than linear combinations) are important;
(b) the choice of variables is important (for example we are not in a situation where half the underlying variables measure the same feature of the data);
(c) the choice of measurement units is important (ideally the variables should be commensurate, so that all the variables are measured in the same units with comparable variances).

I would be interested in the authors' comments on these thoughts.

The following contributions were received in writing after the meeting.

Amir Ahmad, Sarosh Hashmi and Sami M. Halawani (*King Abdulaziz University, Rabigh*)
We congratulate Fan and his colleagues for this interesting paper. The paper proposed a model for the estimation of high dimensional covariance. The proofs are detailed and the experiments are extensive. The discussion provides a good insight into the problem.

Gene expression data sets and protein expression data sets (e.g. Golub *et al*. (1999) and Alon *et al*. (1999)) provide a challenge because of their high dimensions and small number of data points. The authors have talked about statistical genomics as one of the fields of application of the methods proposed in the paper. Hence, it would be interesting if they could show some results obtained by the proposed method on these data sets and if they could comment on future extensions of the model proposed.

Charles Bouveyron (*University Paris 1 Panthéon-Sorbonne*)
Before I go further, I would like to thank the authors greatly for this very interesting and painstaking work. I found this paper made with a real care. I particularly appreciated the fair balance between theory and experiments.

The subject of the paper, large covariance matrix estimation, has become a central problem in modern statistics. Indeed, the technological advances of the last two decades have significantly modified the nature of data, and consequently of statistics. In particular, modern data are often high dimensional (large numbers of variables), big (large numbers of observations) or available as a stream (the observations pass and cannot be stored).

The paper focuses on the factor model and discusses solutions for estimating the covariance matrix. The POET-method that is introduced has the advantage of including existing regularization strategies for large covariance matrix estimation. Among those strategies, one consists in thresholding the principal directions associated with the smallest eigenvalues. For this, POET completes the eigendecomposition with a thresholded matrix, let us say R. This allows us in particular to perform the inversion of the covariance matrix efficiently. An alternative would be to use the covariance matrix approximation that was used in Bouveyron *et al*. (2007) which leads to an explicit inversion for the covariance matrix. Furthermore, recent strategies in estimating sparse covariance matrices include l_1-type penalties. A theoretical and experimental comparison with these approaches would be interesting.

D. S. Coad and H. Maruri-Aguilar (*Queen Mary University of London*)
We congratulate Fan and his colleagues on this beautiful paper, which provides an elegant method for estimating a high dimensional covariance matrix with a conditional sparsity structure. The simplicity of the approach and its wide applicability make it very appealing. Asymptotic properties and simulation results convincingly demonstrate the superiority of the method. We feel that the estimator proposed has a multitude of other potential uses in practice.

The problem of controlling the false discovery rate in example 3 often presents itself in gene association analysis, but limited numbers of observations are available. Since there is only a small number of observations for each hypothesis, a one-stage design can lead to tests with poor power. However, Zehetmayer *et al*. (2005) have shown that a two-stage design based on combining the p-values from a screening stage and a testing stage can significantly improve the power. A generalization to multistage designs is provided by

Zehetmayer *et al*. (2008). A natural question is whether the principal factor approximation can be applied to these designs.

In the multiperiod asset pricing model that is outlined in example 6, to test the null hypothesis (1.2), the model is embedded in the multivariate linear model (5.3). When $p < T$, the usual test statistic has either a χ^2- or an F-distribution under the null hypothesis, according to whether the covariance matrix $\mathbf{\Sigma}_u$ is used or an estimate. However, when $p \gg T$, the estimate of $\mathbf{\Sigma}_u$ is degenerate and the non-degenerate estimate $\hat{\mathbf{\Sigma}}^{\mathcal{T}}_{u,\hat{K}}$ can be employed instead. It would be interesting to know what the distributions are of the test statistics W and $\hat{W}$. In particular, it is unclear what the corresponding degrees of freedom would be.

A problem with large data sets in computer experiments is the intractability of the usual Gaussian process model. The main obstacle is the evaluation and inversion of large covariance matrices. Kaufman *et al*. (2008, 2011) used respectively tapering to produce sparse correlation matrices and correlation functions with compact support. The thresholding methods that are described could be used for the analysis of computer experiments, by devising a special form for the entry-adaptive thresholding rule $s_{ij}(x)$. This would allow fast covariance computations and tractability of the problem.

Wei Dang (*Shihezi University*) **and Keming Yu** (*Brunel University, London*)

The principal component analysis method for large covariance matrix estimation is a novel idea for a challenging issue. By assuming a sparse error covariance matrix in a multifactor model, the proposed principal orthogonal complement thresholding estimator POET does have a proper rate of convergence.

Whereas principal component analysis can apply to the analysis of non-stationary time series (Lansangan and Barrios, 2009), POET may lose those good properties presented in theorems 1 and 3 for non-stationary and non-ergodic time series. Because POET relies on the stationary and ergodic assumptions of underlying time series, it may exclude many important application examples, including financial time series analysis and health science data analysis. For example, modern mathematical models largely focus on martingale models, including Brownian motion. But a multi-dimensional Brownian motion may not be ergodic. In health sciences the data under analysis may be the yearly heights and weights of a large group of children recorded from their early ages to the end of their high school studies. It is often observed that the heights and weights of these children rise much quicker in certain years than in some other years, so the difference between the sample means and variance of the period of quick growth are statistically different from some of the other years; then the data under analysis would be non-stationary. One way to apply POET in these problems may use transformation first, such as detrended non-stationary processes transformed into stationary ones, and Laplace transform non-ergodic processes into ergodic processes.

The other issue with the proposed POET is to incorporate it for the analysis of data with outliers. Many empirical studies find that the distribution of stock returns departs from normality, including the stock return from the Center for Research in Security Prices database and used in the paper. Like principal component analysis, POET may become unreliable if outliers are present in the data. The same type of data may occur in health science. As Jolliffe (2002) pointed out, for a sample of healthy children of various ages between 5 and 15 years old, an observation with height and weight 175 cm and 25 kg respectively is not particularly extreme on either the height or the weight, individually, but the combination (17 cm, 25 kg) is an outlier. In such cases, it is desirable to employ a statistical estimation procedure that may be more efficient and robust than ordinary least squares for a robust POET-estimator.

Matteo Farnè (*University of Bologna*)

I thank the authors for this very challenging paper. While reading it and listening to the presentation, I have learnt much. My comment is on possible extensions of the method proposed.

In Farnè and Montanari (2013) I have done some work on a different approach to the estimation of large covariance matrices, namely the approach based on shrinkage, under assumptions which differ from those considered in this paper, Ledoit and Wolf (2003, 2004) suggested to obtain a well-behaved covariance matrix by shrinking the sample covariance matrix either towards a scaled identity matrix or, to impose some structure on the estimator, towards a single index model covariance matrix. Boehm and Von Sachs (2008, 2009) have successfully extended shrinkage approaches to the estimation of the spectral matrix of a multivariate time series.

My feeling is that the POET-method could be profitably extended to the estimation of large spectral matrices also. Of course, owing to the particular nature of spectral matrices the extension is not straightforward; for instance the effect of smoothing must also be taken into account. I am wondering whether Fan and his co-authors would suggest that we employ their method in the frequency domain also or on the contrary whether they see any reason why such an extension is not advisable.

Marco A. R. Ferreira (*University of Missouri, Columbia*)

I congratulate Professor Fan and his colleagues for their valuable contribution to the area of large covariance matrix estimation.

Professor Fan and colleagues have developed a method for estimating large covariance matrices when there are common unobservable factors and additional cross-sectional correlation. They consider the case when, as the number of individuals p and the number of time points T grow to infinite, the number of common unobservable factors K remains fixed. In addition, in their set-up the eigenvalues corresponding to the common factors are divergent as $p \to \infty$. Finally, they assume that the covariance matrix of the idiosyncratic component is approximately sparse.

With these assumptions, the authors develop a method based on principal component analysis for covariance matrix estimation. Specifically, first they estimate the contribution of the common factors to the covariance matrix by the sum of the K first terms of the sample covariance matrix spectral decomposition. Then, they subtract the estimated common factors contribution from the sample covariance matrix to obtain the principal orthogonal complement. Further, they apply thresholding to the principal orthogonal complement to obtain an estimator of the idiosyncratic covariance matrix. Finally, their covariance matrix estimator is the sum of the estimated common factors contribution and the estimated idiosyncratic covariance matrix.

I have two main comments or questions on the paper.

(a) As the number of individuals p increases, it seems intuitive to assume that the underlying process generating the data should grow in complexity, i.e. it seems intuitive that K should grow with p. What would be the potential technical issues that would arise if one decides to extend the current work to the case when K grows with p?
(b) For the application of thresholding, there are a number of constants that must be chosen such as τ in equation (2.6) and C in equation (3.2). There seems to be an opportunity for the use of empirical Bayes methodology for the estimation of those threshold parameters.

Florian Frommlet (*Medical University Vienna*)

I congratulate the authors on this impressive paper concerned with estimating high dimensional covariance matrices under conditional sparsity. Their approach is surprisingly simple: first compute the principal components of the sample covariance matrix, then estimate the number of relevant components and finally apply a thresholding procedure on the remaining covariance matrix. In spite of this simplicity extensive simulation studies in their paper show that POET, the implementation of the approach presented, outperforms competing algorithms in various scenarios.

It is not too surprising that POET performs well in those scenarios based on factor models with few factors, which mimic the situation under which the authors have derived asymptotic results for their method, i.e. when the covariance matrix has a small (fixed) number K of very large eigenvalues. It is quite intuitive that in this situation the first K principal components will simply represent the corresponding factors of the factor model. Also it appears to be clear that the procedure works well when no factors are present, as long as the number of components is then correctly estimated to be 0.

For me the most astonishing result is that POET appears to do relatively well in model 3 of Section 6.5.2, where data were simulated according to an auto-regressive AR(1) model. This is the only presented simulation scenario where data were not simulated either from a factor model with a small number of strong factors, or alternatively from a model without factors and sparse covariance matrix. The covariance matrix of the suggested AR(1) model does not have particularly spiked eigenvalues, but the eigenvalues smoothly decrease from their maximum. In fact for $p = 200$ and $p = 300$ there are 36 and 53 eigenvalues larger than 1 respectively. According to the simulations presented POET picks for this scenario (both for $p = 200$ and $p = 300$) on average roughly six factors to model the covariance structure, outperforming direct thresholding of the sample covariance matrix. This result indicates that POET might work well even in situations which are not covered by the asymptotic analysis presented. However, further work seems to be necessary to explain why that would be so.

I. Gijbels and K. Herrmann (*KU Leuven*) **and A. Verhasselt** (*Universiteit Hasselt and Universiteit Antwerpen*)

Fan and his colleagues present a very nice estimation technique (POET) for high dimensional covariance estimation, based on principal component estimation and thresholding the orthogonal complement of the principal components. They show that POET is equivalent to constrained least squares (CLS) estimation.

We wonder how robust POET is when the data matrix is corrupted, since it is well known that least-

squares-based methods are not robust. The equivalence of POET to a CLS estimation problem seems to open the way for a more robust procedure. The use of robust principal component methods (e.g. Engelen *et al*. (2005)) could also offer a possibility.

As pointed out in the literature (see for example Antoniadis (2007)), the qualitative properties of a thresholding rule turn out to be important. For example, the hard thresholding rule is discontinuous, whereas the soft thresholding rule is continuous. In CLS regression hard thresholding leads to a larger variance of the estimates, whereas soft thresholding shifts the estimates, creating a bias. What is the effect of such qualitative properties of the thresholding rule on the POET estimator?

The authors use a computationally expensive cross-validation criterion to choose C. It might be worth the effort to exploit the equivalent CLS problem and to use criteria based on this equivalence, such as an Akaike type of criterion.

Portfolio allocation in the Markowitz (1952) framework is chosen as a numerical illustration of POET. In the simulation studies and empirical application, the emphasis is on estimating the weights of the minimum variance (MV) portfolio as the solution to $\mathbf{w}_{\mathrm{MV}}(\boldsymbol{\Sigma}) = \arg\min_{\mathbf{w}'\mathbf{1}=1} \mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}$. This is in line with current literature (Kourtis *et al*. (2012) and references therein) where the MV portfolio is preferred because it alleviates the necessity of estimating the stock returns. As the MV weights admit the expression $\mathbf{w}_{\mathrm{MV}}(\boldsymbol{\Sigma}) = \boldsymbol{\Sigma}^{-1}\mathbf{1}/(\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1})$, the comparison of POET, the strict factor model (Fan *et al*., 2008) and the sample covariance (SC) matrix estimator amounts to a comparison of the estimated precision matrix in the models considered. It is known that the SC precision matrix performs poorly (Fan *et al.*, 2008; Kourtis *et al.*, 2012) and measures to counterbalance estimation errors must be taken. In a $p < T$ framework shrinkage methods for example are applied to the SC matrix before (Ledoit and Wolf, 2003) or after inversion (Kourtis *et al.*, 2012), significantly enhancing results. Shrinkage methods have also been applied to the $p \gg T$ framework (see Ledoit and Wolf (2004)), establishing a possible competitor in this scenario as well. Owing to the known shortcomings of the SC precision matrix deeper insights can be expected from a comparison with such refined methods.

Wally Gilks (*University of Leeds*)

Fan and his colleagues state that the low rank plus sparse representation of their model is for the *population* covariance matrix. The most obvious interpretation of this assertion is that the model is intended to describe the population, not the specific individuals sampled. This interpretation is somewhat at odds with the design of the sparse component of the model $\boldsymbol{\Sigma}_u$, which accounts for idiosyncratic correlations between *specific* individuals.

At a population level, such idiosyncratic components can only be represented in terms of *probabilities* of idiosyncratic correlation. For example, suppose that two individuals i and j, randomly and independently sampled from the population, have a probability π of interacting idiosyncratically. Suppose further that the covariance $\sigma_{u,ij}$ of their idiosyncratic errors is ρ if i and j interact, and 0 otherwise. In a sample of size p, the probability that individual i idiosyncratically interacts with k other individuals is distributed as binomial$(\pi, p-1)$. The authors require that their measure of sparsity, $m_p = \max_{i \leqslant p} \Sigma_{j \leqslant p} |\sigma_{u,ij}|^q$, grows with sample size as $o(p)$. Letting $\sigma_{u,ij} = \tau$, we have

$$\begin{aligned} m_p &= \tau^q + \rho^q K_p^{(1)} \\ &> \tau^q + \rho^q \bar{K}_p \\ &\approx \tau^q + \rho^q (p-1)\pi \\ &\geqslant \rho^q \pi p \end{aligned}$$

where $K_p^{(1)}$ and $\bar{K}_p$ denote the maximum and mean values in a sample of size p from a binomial$(\pi, p-1)$ distribution. Thus, $m_p \neq o(p)$ unless $\rho = 0$.

Hajo Holzmann and Anna Leister (*Philipps-Universität Marburg*)

We congratulate Fan and his colleagues for an inspiring paper on estimating the factor structure in high dimensional, approximate factor models, and its consequences for estimating the underlying covariance matrix.

Let us consider implications for the time series structure of $(\mathbf{y}_t)$, specifically its lagged covariance matrix, and convergence in the $\|\cdot\|_{\max}$-norm.

When estimating $\Sigma = \mathrm{cov}(\mathbf{y}_t)$, Fan *et al.* (2011), theorem 3.2, obtain the rate $O_p[\sqrt{\{\log(p)/T\}}]$ in $\|\cdot\|_{\max}$ for an estimate based on an observed factor structure, whereas in the present paper, utilizing estimated factors, the authors obtain the rate $O_p[1/\sqrt{p} + \sqrt{\{\log(p)/T\}}]$. Now, for the sample covariance, writing

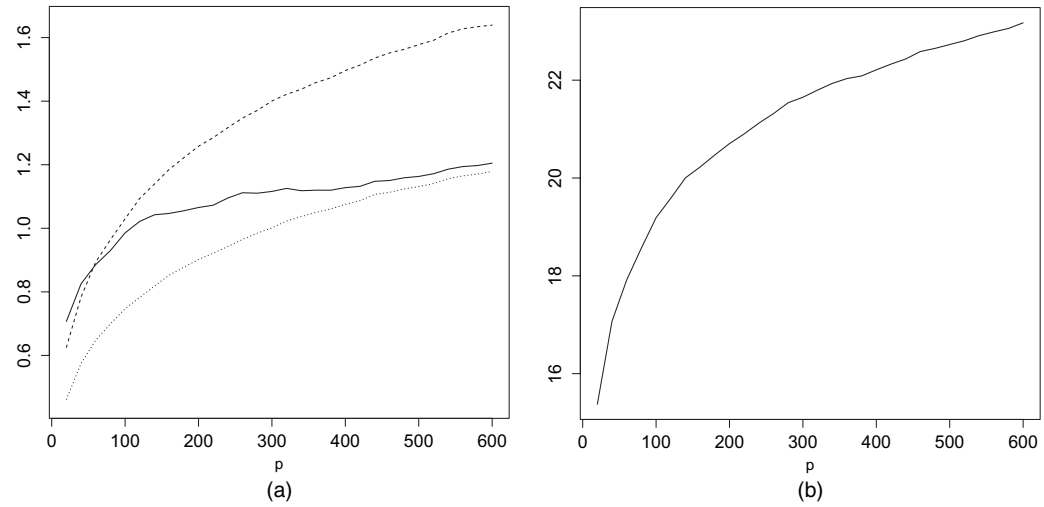

Fig. 16. Time lag $h = 1$: (a) averages of error (max-norm) over 500 simulations against p (------, sample; ———, factor; · · · · · ·, observed factor); (b) averages of lagged covariance (max-norm) over 500 simulations against p

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{y}_t\mathbf{y}_t' - \Sigma = \mathbf{B}\frac{1}{T}\sum_{t=1}^{T}\{\mathbf{f}_t\mathbf{f}_t' - \mathrm{cov}(\mathbf{f}_t)\}\mathbf{B}' + \frac{1}{T}\sum_{t=1}^{T}\{\mathbf{u}_t\mathbf{u}_t' - \mathrm{cov}(\mathbf{u}_t)\} + \mathbf{B}\frac{1}{T}\sum_{t=1}^{T}\mathbf{f}_t\mathbf{u}_t' + \frac{1}{T}\sum_{t=1}^{T}(\mathbf{u}_t\mathbf{f}_t')\mathbf{B}',$$

and using lemma 4 (in the present paper) as well as $\max_{1\leqslant j\leqslant p}\|\mathbf{b}_j\| = O_p(1)$ we obtain $O_p[\sqrt{\{\log(p)/T\}}]$. Thus, using an estimated factor structure may not be beneficial for moderate values of p.

For the lagged covariance $\mathrm{cov}(\mathbf{y}_t, \mathbf{y}_{t+h}) = E(\mathbf{y}_t\mathbf{y}_{t+h}')$, $h \geqslant 1$, assume for distinction that the errors $(\mathbf{u}_t)$ are known to be serially uncorrelated: $\mathrm{cov}(\mathbf{u}_t, \mathbf{u}_{t+h}) = \mathbf{0}$.

Then using either an observed factor structure or the sample autocovariance gives the rate $O_p[\sqrt{\{\log(p)/T\}} + (h/T)\|\mathrm{cov}(\mathbf{f}_t, \mathbf{f}_{t+h})\|_{\max}]$. In contrast, for the estimate $(1/T)\sum_{t=1}^{T-h}(\hat{\mathbf{b}}_i'\hat{\mathbf{f}}_t\hat{\mathbf{f}}_{t+h}'\hat{\mathbf{b}}_j)$, $i, j = 1, \ldots, p$, we obtain from corollary 1 (in the present paper)

$$O_P[\log(T)^{2/r2}\sqrt{\{\log(p)/T\}} + \log(T)^{1/r_2}T^{1/4}/\sqrt{p} + (h/T)\|\mathrm{cov}(\mathbf{f}_t, \mathbf{f}_{t+h})\|_{\max}].$$

We give a finite sample illustration, similar in setting to Fan *et al.* (2011), but with factors following a more strongly dependent AR(1) process. The results are plotted in Fig. 16. Further details for the above statements and the simulation can be found at `http://www.uni-marburg.de/fb12/stoch/files/holzmann/fandiscuss.pdf`.

Hanwen Huang, Yufeng Liu, J. S. Marron, Dan Shen and Haipeng Shen (*University of North Carolina at Chapel Hill*)

We congratulate Fan and his colleagues on a very interesting contribution, which takes the fundamentally important field of covariance matrix estimation in some important new directions. We agree that now is a good time to be studying asymptotic contexts, where the first K eigenvalues of Σ grow quickly. The asymptotic mode of the sample size tending to ∞, with an exponentially growing dimension, can be improved by taking the dimension as the asymptotic driver, with the sample size growing at a logarithmic rate. This makes it clear that this setting is very close to the high dimension, low sample size setting with fixed sample size (Hall *et al.*, 2005). Shen *et al.* (2012) studied another notion of principal component analysis consistency, in a wide range of such settings. Shen *et al.* (2013) studied another approach to sparsity under a growing eigenvalue assumption, establishing a new characterization of the boundary between regions of consistency and strong inconsistency for sparse principal component analysis in high dimension, low sample size settings. Can similar results be established for POET?

Another reason why we are excited about these results is that covariance estimation is a critical component of SigClust, which is very useful for testing statistical significance of clusters in high dimensional contexts (Liu *et al.*, 2008; Huang *et al.*, 2013). This motivated us to compare POET with the approaches used in SigClust. A key step of the SigClust analysis is to estimate the eigenvalues of the covariance matrix

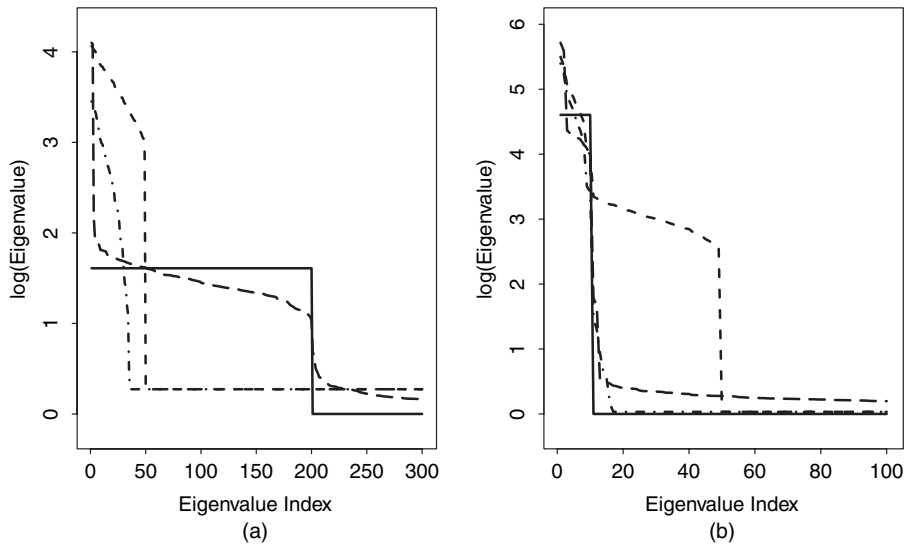

Fig. 17. Estimated covariance matrix eigenvalues based on the hard (------), soft (·-·-·-) and POET (– – –) methods (———, true) for two simulated data sets with $d = 1000$ and $n = 50$: (a) results for spike size $\lambda = 5$ and $w = 200$ spike entries (POET works best); (b) results for $\lambda = 100$ and $w = 10$ (soft thresholding works best)

of the null multivariate Gaussian distribution. Huang *et al.* (2012) proposed a likelihood-based soft thresholding approach for estimating the covariance eigenvalues which gave a large improvement relative to the hard thresholding approach of the former paper. Fig. 17 shows estimates of the eigenvalue spectrum for two simulated high dimension, low sample size examples with sample size $n = 50$ and dimension $d = 1000$. Gaussian data are simulated with mean 0 and covariance matrix

$$\Lambda = \mathrm{diag}(\underbrace{\lambda, \ldots, \lambda}_{w}, 1, \ldots, 1).$$

Fig. 17(a) shows that, in situations where the number of the spikes is larger than the sample size, the POET-method gives a major improvement. Fig. 17(b) shows that, in situations with few large spikes, the soft method works better than POET owing to better background noise estimation. Ultimately, a combination of POET with existing SigClust methods may work better.

Jian Huang (*University of Iowa, Iowa City, and Shanghai University of Finance and Economics*) **and Yong Zhou** (*Chinese Academy of Sciences, Beijing, and Shanghai University of Finance and Economics*)

We congratulate Fan and his colleagues on presenting a wonderful and thought-provoking paper dealing with an important topic in high dimensional data analysis. They introduce the POET-methodology for covariance matrix estimation and study its properties. The theoretical results obtained by them are highly original, notably the aspects concerning the relative magnitude of the sample size and the number of variables. They also describe several *poetic* and important applications. We focus our discussion on an application of POET in the linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Here $\mathbf{y} \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of regression coefficients and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ consists of independent errors.

An attractive approach to selection and estimation in high dimensional regression is based on the penalized least squares criterion $(1/2n)\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \Sigma_{j=1}^p \rho(\beta_j; \lambda)$, where $\rho(\cdot; \lambda)$ is a suitable penalty function with a tuning parameter $\lambda \geqslant 0$. The success of this approach depends on the behaviour of the restricted eigenvalues and related quantities of $X'X/n$ (see, for example, Bickel *et al.* (2009)). We discuss a way to repair the degeneracy of $X'X/n$ based on POET.

Let $\boldsymbol{\Sigma}$ be the covariance matrix of the row vectors in X. Assume factor model (1.1) in the paper for the predictors and denote the POET-estimator by $\hat{\boldsymbol{\Sigma}}$. Consider the spectral decomposition $\hat{\boldsymbol{\Sigma}} = \mathbf{V}\mathbf{D}^2\mathbf{V}'$, where $\mathbf{V}$ is a $p \times p$ orthonormal matrix of the eigenvectors and $\mathbf{D}^2$ is a diagonal matrix of the eigenvalues of $\hat{\boldsymbol{\Sigma}}$. Let $\mathbf{U} = \mathbf{X}\mathbf{V}\mathbf{D}^{-1}$. Then $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}'$. This expression is reminiscent of singular value decomposition. But here

$\mathbf{U}$ is only approximately orthogonal, since $\mathbf{V}$ is from $\hat{\mathbf{\Sigma}}$. However, it can be viewed as a POET-regularized singular value decomposition.

The least squares loss equals $(\mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta})/(2n)$. Replacing $\mathbf{X}'\mathbf{X}/n$ by $\hat{\mathbf{\Sigma}}$ in this expression and noting that $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}'$, we obtain $\|n^{-1}\mathbf{U}'\mathbf{y} - \mathbf{D}\mathbf{V}'\boldsymbol{\beta}\|^2/2$ plus a term independent of $\boldsymbol{\beta}$. Let $\tilde{\mathbf{y}} = \mathbf{U}'\mathbf{y}/n$ and $\tilde{\mathbf{X}} = \mathbf{D}\mathbf{V}'$. It is natural to consider the penalized criterion $\frac{1}{2}\|\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\beta\|^2 + \Sigma_{j=1}^{p}\rho(\beta_j; \lambda)$. The loss function here can be considered a regularized version of the least squares loss, in which the rank deficiency of $X'X/n$ is repaired by making use of $\hat{\mathbf{\Sigma}} = \tilde{\mathbf{X}}'\tilde{\mathbf{X}}$.

In particular, the least squares estimator based on $(\tilde{\mathbf{y}}, \tilde{\mathbf{X}})$ is $\hat{\boldsymbol{\beta}}_{\mathrm{LS}} = (\tilde{\mathbf{X}}'\mathbf{X})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}} = \hat{\mathbf{\Sigma}}^{-1}\mathbf{X}'\mathbf{y}/n$, which is well defined because $\hat{\mathbf{\Sigma}}$ is invertible. With standardized predictors, $\mathbf{X}'\mathbf{y}/n$ was used for screening variables by Fan and Lv (2008). The $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$ can be considered a corrected version of $\mathbf{X}'\mathbf{y}/n$ based on $\hat{\mathbf{\Sigma}}$. It also can be used for screening.

The validity of the above proposal rests on the properties of $\hat{\mathbf{\Sigma}}$. Simulation studies are needed to evaluate their finite sample performance. Also, much work is required to analyse their theoretical properties. The results obtained by the authors provide a solid basis for such analyses.

Sujngkyu Jung (*University of Pittsburgh*) **and Jason P. Fine** (*University of North Carolina at Chapel Hill*)

In this very stimulating paper, Fan and colleagues show the gain of conditional sparsity assumptions in covariance matrix estimation when principal components are pervasive. It is striking that the estimation procedure uses the first few principal components from the sample covariance matrix without any thresholding. The sample principal components are known to be inconsistent without strong assumptions (Johnstone and Lu, 2009). The present paper shows that under the conditional sparsity assumption the covariance estimator based on these principal components is consistent.

It seems that the gains in the methodology proposed arise in part from the sparsity assumptions and in part from the pervasive factor assumptions. The latter assumption requires that the K largest eigenvalues of the $p \times p$ covariance matrix Σ are of order p. It is well known that the magnitude of eigenvalues is a critical condition for the consistency of principal component directions when dimension p increases with the sample size n. As an example, suppose that the largest eigenvalue $\lambda_{1,p}$ of Σ is of magnitude $\delta(p)$ with the rest being simply 1. The corresponding sample eigenvector $\hat{\xi}_{1,p}$ is consistent in the sense that $\|\hat{\xi}_{1,p} - \xi_{1,p}\| \to 0$ as $p \to \infty$ when $\delta(p)/p \to \infty$. In contrast, such a strong result does not hold whenever $\delta(p) = O(p)$ or $o(p)$ (Jung and Marron, 2009; Jung *et al.*, 2012). This gives the insight that the sparsity assumption on the error covariance matrix is critical for the proposed estimator under the pervasive factor assumption (i.e. $\delta(p) = O(p)$).

Should we be tied to the pervasive factor assumption? Many other conditions have been considered in the literature. For example, in random-matrix theory where n and p increase at the same rate, it is customary to assume fixed eigenvalues for all p (see, for example, Paul (2007)), which corresponds to $\delta(p) \equiv \delta = o(p)$. Meanwhile, implicit assumptions in sparse estimation of principal components are that the number of non-zero loadings in population eigenvectors grows at a slower rate than p (Shen *et al.*, 2013). The case $\delta(p)/p \to \infty$ yields a trivial solution, with easy separation of the leading eigenvectors from the error covariance matrix. In contrast, the case $\delta(p)/p \to 0$ makes estimation of the covariance matrix much more difficult. What can be said about the proposed estimator when $\lambda_{1,p} = o(p)$? It would be worthwhile to investigate more carefully the interaction between the two key assumptions on the magnitudes of the eigenvalues and the sparsity of the orthogonal complement matrix.

Although the theoretical results are quite elegant, concerns arise about the practical implications of the two key assumptions. This is particularly true since there may not be information in the data to detect violations of the assumptions. In what types of applications are such assumptions reasonable? Are there situations where one type of assumption is more realistic than the other type? Are there diagnostics that might be employed in real data analysis? Additional practical guidance would be welcomed.

Oliver Linton and Michael Vogt (*University of Cambridge*)

Fan and his colleagues address the important issue of estimating large structured covariance matrices. They restrict the large $p \times p$ covariance matrix $\mathbf{\Sigma}$ to have the form $\mathbf{\Sigma} = \mathbf{\Sigma}_s + \mathbf{\Sigma}_u$, where

(a) the systematic part $\mathbf{\Sigma}_s$ has K large ($O(p)$) eigenvalues and $p - K$ zero eigenvalues and
(b) the residual part $\mathbf{\Sigma}_u$ is a sparse matrix with bounded eigenvalues.

In financial applications, $\mathbf{\Sigma}_u$ represents idiosyncratic risk that can be diversified away, and so makes a smaller order contribution to portfolio risk, but in practice it can be important. The authors are to be congratulated on their comprehensive and useful method for taking full account of this structure.

The assumption that all of the non-zero eigenvalues of $\mathbf{\Sigma}_s$ dominate the largest eigenvalue of $\mathbf{\Sigma}_u$ by the magnitude p is likely to be a little strong in practice. If K is moderately large, the Kth eigenvalue of $\mathbf{\Sigma}_s$ can be expected to be much closer to the largest eigenvalue of $\mathbf{\Sigma}_u$ than the first eigenvalue. This may affect the quality of the estimation procedure and make the problem of selecting K difficult. Fig. 18 illustrates this point with the help of the data from Section 7. We wonder whether the main theoretical results continue to hold under weaker assumptions on the growth rate of the smallest positive eigenvalue of $\mathbf{\Sigma}_s$.

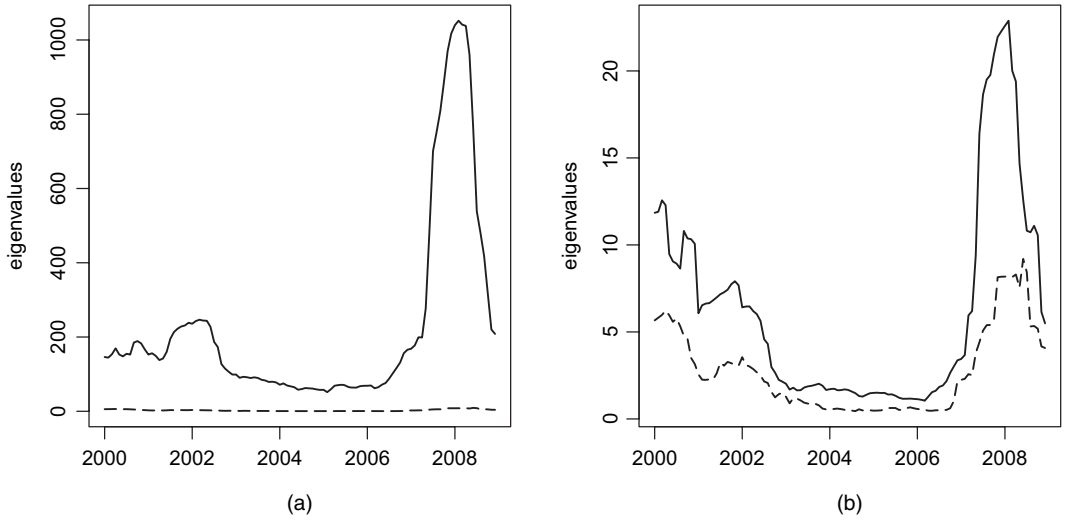

Fig. 18. Estimated eigenvalues of the matrices $\mathbf{\Sigma}_S$ and $\mathbf{\Sigma}_u$ for the various yearly data samples used in Section 7.2 (the time points on the x-axis indicate the starting date of each sample and $K = 3$ as in the paper; the plots show that the first (i.e. the largest) eigenvalue of $\mathbf{\Sigma}_S$ is much more spiked than the third, the latter roughly having the same magnitude as the largest eigenvalue of $\mathbf{\Sigma}_u$): (a) largest eigenvalue of $\mathbf{\Sigma}_S$ (———) and $\mathbf{\Sigma}_u$ (– – –) in each sample; (b) comparison of the third largest eigenvalue of $\mathbf{\Sigma}_S$ (———) with the largest eigenvalue of $\mathbf{\Sigma}_u$ (– – –)

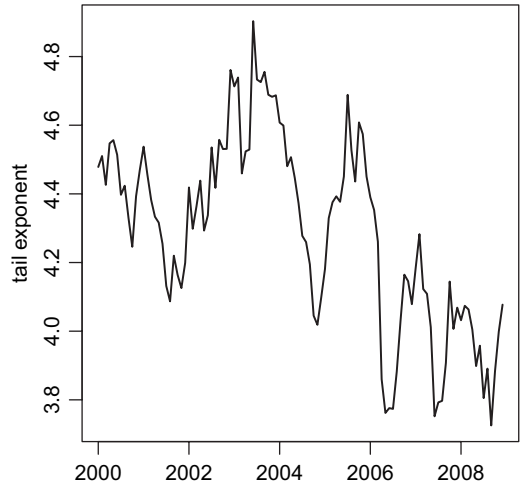

Fig. 19. Estimated tail exponents of the model residuals for the various yearly data samples in Section 7.2: for each sample we calculate the residuals and apply a log-rank regression to them to obtain estimates of the Pareto tail exponents; ———, median of the estimated exponents obtained for each sample (as can be seen, the exponents take values roughly between 3.5 and 5, indicating that in many cases only the first few moments will exist)

Another remark concerns the technical assumption 2, part (c), which imposes exponentially decaying tails on the model residuals. This is a very strong condition which in particular implies that all moments exist. In applications to daily equity returns this is likely to be violated. Fig. 19 shows a log-rank plot of the data from Section 7 which suggests that the residuals are far from having exponentially decaying tails. To have a better idea of how the estimation procedure works when applied to financial data, it would thus be important to understand which parts of the procedures are robust to weaker moment conditions and which are not.

Regarding the portfolio choice application, the authors use shrinkage methods to impose sparsity on the idiosyncratic part of the covariance matrix of returns. An alternative or perhaps complementary approach here (see Yen (2011)) is to impose sparsity on the portfolio weights through an L_1-penalty. Each non-zero investment entails a transaction cost and so it makes financial sense to minimize the number of such transactions; this is especially relevant for very large portfolios turned over daily. One further concern with the portfolio methodology is that no smoothness assumptions on the thresholded idiosyncratic covariances are exploited. In particular, the location of the 0s in the thresholded matrices (and thus their eigenstructure) may change abruptly over time, even though the rolling window data overlap considerably from period to period.

Han Liu (*Princeton University*) **and Lie Wang** (*Massachusetts Institute of Technology, Cambridge*)

We congratulate Professor Fan, Professor Liao and Miss Mincheva for their thought-provoking paper. We believe that the proposed methodology will have profound impact and stimulate many further researches.

Estimating a large covariance matrix under a small sample size is a fundamental problem. However, it suffers from the challenge that the eigenvalues of the sample covariance matrix do not converge to the population truth when the population eigenvalues are bounded or grow at a slow rate. In this paper, Professor Fan, Dr Liao and Miss Mincheva avoid this problem by exploiting an approximate factor model with a spiked eigenvalue condition: they assume that the population covariance matrix decomposes into a low rank component and a residual component. The eigenvalues of the low rank component are spiked and diverge at a fast rate, whereas the eigenvalues of the residual component are bounded. The POET-estimator directly runs the singular value decomposition on the sample covariance matrix. It estimates the low rank component by the top principal components and applies thresholding methods to estimate the residual component according to different sparsity and smoothness conditions. The covariance matrix is then estimated by combining these two components.

Their paper stimulated us to consider the following two extensions.

Semiparametric extension

The POET-method requires exponential-type tails of the data to establish large deviation results. It is interesting to extend this method to handle data from the semiparametric non-paranomral family (Liu *et al.*, 2012).

A random vector $\mathbf{X}$ belongs to a non-paranomral family if there is a set of univariate monotone functions $\{f_j\}_{j=1}^d$ such that $f(\mathbf{X}) := (f_1(X_1), \ldots, f_d(X_d))^{\mathrm{T}}$ is Gaussian, i.e. $f(\mathbf{X}) \sim N(\mathbf{0}, \mathbf{\Sigma}^*)$. For identifiability $\mathbf{\Sigma}^*$ is constrained to be a correlation matrix. Under the non-paranomral model, Liu *et al.* (2012) suggested replacing the sample correlation matrix by the Kendall's τ rank correlation matrix $\hat{\mathbf{\Sigma}}$ with

$$\hat{\mathbf{\Sigma}}^r_{jk} = \sin\left(\frac{\pi}{2}\hat{\tau}_{jk}\right), \tag{7}$$

where $\hat{\tau}_{jk}$ is the empirical Kendall τ-statistic between X_j and X_k. By assuming that $\mathbf{\Sigma}^*$ admits the 'low rank plus sparse' structure, we could apply the POET-method on $\hat{\mathbf{\Sigma}}^\tau$. We have obtained encouraging numerical results; further theoretical investigation is on the way.

Tuning-insensitive extension

Another extension is to apply more sophisticated methods to estimate the residual component matrix. For Gaussian models, Liu and Wang (2012) proposed a sparse inverse covariance estimation method named TIGER. The TIGER-estimator is tuning sensitive and achieves the optimal rates of convergence for both covariance and inverse covariance estimation under different norms. It would be interesting to see whether these good theoretical properties still hold for the corresponding POET-estimator.

Jorge Mateu (*University Jaume I, Castellón*)

Fan and his colleagues are to be congratulated on a valuable and thought-provoking contribution on the estimation of high dimensional covariances with a conditional sparsity structure. As they note, this problem can be encountered in a wide variety of practical examples and scientific fields. In particular they mention the problem of high dimensional classification.

I would like to comment on this problem in the context of spatial point processes. Byers and Raftery (1998) considered the problem of detecting features in spatial point processes in the presence of substantial clutter with the aim of outlining seismic faults. They used kth-nearest-neighbour distances to produce high breakdown point robust estimators of a covariance matrix in a high dimensional problem. If in this context we assume that we have some common but unknown factors, we can then use the idea of the principal orthogonal complement thresholding method to explore such an approximate factor structure with sparsity. We have sound strategies to calculate in this spatial context the sample covariance matrix and the factor-based covariance matrix so that we can use the idea of conditional sparsity.

In the context of spatial point processes, Collins and Cressie (2001) developed exploratory data analytic tools, in terms of local indicators of spatial association functions based on the product density, to examine individual points in the point pattern in terms of how they relate to their neighbouring points. For each point of the point pattern we have a local indicators of spatial association function. To perform statistical inference, needed for example in testing for local clustering, Collins and Cressie (2001) developed closed expressions of the autocovariance and cross-covariance between any two such functions. These covariance structures are complicated to work with as they live in (very) high dimensional spaces. Again, it is not difficult to assume common factors among these functions and thus it could be appropriate to consider conditional sparsity to estimate the covariance matrix consistently. This will provide a new insight in such a problem.

Consider, finally, an approach based on latent process modelling and principal component analysis to obtain a computationally feasible exploratory tool for discovering patterns of association between components of a highly multivariate point process. The latent Gaussian fields are obtained as linear combinations of some independent Gaussian processes. Again it is easy to think about the POET-method to estimate the complicated covariance matrix.

Guangming Pan (*Nanyang Technological University, Singapore*) **and Heng Peng** (*Hong Kong Baptist University*)

We congratulate Professor Fan, Dr Liao and Ms Mincheva for such a timely paper. We enjoyed reading it since there are some good and novel ideas here, particularly the idea to estimate a covariance matrix by reducing $\mathbf{\Sigma}$ to a low rank matrix plus a sparse matrix and the concept of conditional sparsity.

The idea of this paper can be applied also to an ultrahigh dimensional linear model. Consider a high dimensional linear regression model

$$y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, 2, \ldots, n. \tag{8}$$

When p, the dimension of $\boldsymbol{\beta}$, is large, $\boldsymbol{\beta}$ is always assumed to be sparse. In some applications, y_i would depend on the first finite principal components of $\mathbf{x}_i$. The regressor $\mathbf{x}_i$ can be then supposed to follow a factor model like

$$\mathbf{x}_i = \mathbf{B}^{\mathrm{T}}\mathbf{f}_i + \mathbf{u}_i, \qquad i = 1, 2, \ldots, n, \tag{9}$$

where $\mathbf{B}$ is the $r \times p$ factor loading matrix, $\mathbf{f}_i$ is the $r \times 1$ factor process and $\mathbf{u}_i$ are the $p \times 1$ idiosyncratic error components. Combining model (8) with model (9) we then have

$$y_i = \mathbf{f}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}} + \mathbf{u}_i^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, 2, \ldots, n, \tag{10}$$

where $\tilde{\boldsymbol{\beta}} = \mathbf{B}\boldsymbol{\beta}$. When considering the new model (10), $\tilde{\boldsymbol{\beta}}$ is not necessarily sparse. Though the dimension of the model is still ultrahigh, as in Wang (2012) and Ke *et al.* (2012), it can be efficiently reduced by the sure independence screening procedure (Fan and Lv, 2008) if we impose some simple structure on the covariance matrix of $\mathbf{u}_i$.

Fan, Liao and Mincheva assume that the number of principal components, K, is known. In many applications K is unknown and must be estimated. There is some literature focusing on this problem, e.g. Bai and Ng (2002) and Onatski (2009, 2010). Their approaches require similar spike conditions such that the first K largest eigenvalues go to ∞ and the remaining eigenvalues are bounded. But what would happen if $\mathbf{\Sigma}$ is structured in a different way, say, $\mathbf{\Sigma}$ is a Toeplitz matrix where the eigenvalues are not spiked? Can the number of factors still be estimated consistently? Below we consider the problem of estimating K (the first K largest eigenvalues do not tend to ∞).

In some sense, estimating the number of factors is equivalent to finding the number of eigenvalues of the population covariance matrix $\mathbf{\Sigma}$ which are greater than a constant number C, i.e. $K = \#\{i : \lambda_i \geqslant C\}$, where $\#\{i : i \in A\}$ denotes the number of λ_i which satisfies the property A and $\lambda_i, \#i = 1, 2, \ldots, p$, are eigenvalues of $\mathbf{\Sigma}$.

Note that when $p/n \to 0$, under some regularity conditions, eigenvalues of the sample covariance matrix $\hat{\mathbf{\Sigma}}$ are consistent estimates of the respective population eigenvalues of $\mathbf{\Sigma}$ (theorem 4 of Chen and Pan (2012)). Hence K can be determined by the sample eigenvalues as $\hat{K} = \#\{i : \lambda_i \geqslant C\}$. When $p/n \to c \in (0, \infty)$, Baik and Silverstein (2006) and Bai and Yao (2008) stated that the eigenvalues of spiked $\mathbf{\Sigma}$ that are greater than $1 + \sqrt{c}$ can be recovered from those of $\hat{\mathbf{\Sigma}}$. Each population eigenvalue λ_i outside $[1 - \sqrt{c}, 1 + \sqrt{c}]$ pulls one sample eigenvalue away from the support $[(1 - \sqrt{c})^2, (1 + \sqrt{c})^2]$ of the Marchenko–Pastur distribution and positions it at $\lambda_i + c\lambda_i/(\lambda_i - 1)$. Therefore, when $C > 1 + \sqrt{c}$, we can estimate K as

$$\hat{K} = \# \left\{ i : \frac{\hat{\lambda}_i + 1 - c + \sqrt{\{(c - 1 - \hat{\lambda}_i)^2 - 4\hat{\lambda}_i\}}}{2} \geqslant C \right\}.$$

When $p/n \to \infty$, we can firstly split the random vector y_i into several subgroups by some criteria. For every subgroup, its dimension would be smaller than or proportional to n. Hence the number of factors for every subgroup can be determined by the method suggested above. Since every subgroup uses the same factors, those estimated numbers of factors for every different subgroup can be averaged, or maximized to obtain the final estimate of K, the number of factors for all of the random vector y_i. Though such an idea including computational burden would need further investigation, we believe that, when p is significantly larger than n, the number of factors in the model should be able to be estimated accurately. We would be very interested in hearing the authors' views on this point.

Mohsen Pourahmadi (*Texas A&M University, College Station*)

How do we go beyond the prevalent *sparsity* assumption in the recent literature and estimate a large, *non-sparse* covariance matrix? The question arises naturally in the factor models of the form $\Sigma = \Lambda\Lambda' + \Psi$, which is a low rank plus a sparse (non-diagonal) matrix, known as *approximate factor models* (Chamberlain and Rothschild, 1983). Given the sample data $Y_1, \ldots, Y_n$ with dimension $p \geqslant n$, the attraction of the POET covariance estimator proposed by Fan, Liao and Mincheva is in its simplicity, transparency, generality and rigour. These attributes are highly desirable and we would like to see more of them in the rapidly growing algorithm-driven area of high dimensional data analysis (Pourahmadi, 2013).

Construction of a POET-estimator is simple and proceeds as follows.

Step 1: start with the spectral decomposition of the sample covariance matrix of the data,

$$S = \sum_{i=1}^{q} \hat{\lambda}_i \hat{e}_i \hat{e}_i' + \hat{R},$$

where q is the number of selected principal components and $\hat{R} = (\hat{r}_{ij})$ is the residuals.

Step 2: apply the adaptive thresholding (Cai and Liu, 2011) to $\hat{R}$

$$\hat{R}^{\delta} = (\hat{r}_{ij}^{\delta}), \qquad \hat{r}_{ij}^{\delta} = \hat{r}_{ij} I(|\hat{r}_{ij}| \geqslant \delta_{ij}),$$

where

$$\delta_{ij} = \delta(\hat{r}_{ii}\hat{r}_{jj})^{1/2}, \qquad \delta \geqslant 0.$$

Step 3: a (q, δ) POET-estimator of Σ is

$$\hat{\Sigma}^{(q,\delta)} = \sum_{i=1}^{q} \hat{\lambda} \hat{e}_i \hat{e}_i' + \hat{R}^{\delta}. \tag{11}$$

In spite of its simplicity, the POET-estimator is quite general in that it subsumes some important old and new covariance estimators for various choices of (q, δ) in equation (11).

(a) When $\delta = 0$ and $q = p$, it reduces to the sample covariance matrix.
(b) When $\delta = 1$, the estimator reduces to that based on the standard factor model.
(c) When $q = 0$ it reduces to the thresholded estimator of Bickel and Levina (2008) or the adaptive thresholded estimator of Cai and Liu (2011) depending on the choice of δ_{ij}.

In addition as a bonus, using equation (11) and the Sherman–Morrison–Woodbury formula we obtain estimators of the the precision matrix.

The asymptotic properties of the POET covariance estimators are established when the data are temporally correlated and under the strict stationarity assumption. In this set-up, it is desirable to go beyond estimating $\Sigma = \Gamma_0$ and to have consistent estimators of the autocovariance matrices $\{\Gamma_h\}$ or the spectral density matrix of the underlying process. I wonder whether the authors have thought of this problem and

can shed any light on how their conditions relate to those in Forni *et al.* (2004) in the context of generalized dynamic factor models.

Cheng Yong Tang (*University of Colorado, Denver*) **and Yingying Fan** (*University of Southern California, Los Angeles*)

We most heartily congratulate Fan and his colleagues for their thought-provoking and impactful work on estimating the large covariance matrix, which is pivotal in many contemporary scientific and practical studies. Facilitated by a factor model, a parsimonious structure is proposed for the large covariance matrix by combining a low rank matrix and a sparse covariance matrix. In the authors' framework, a factor model is used to characterize the systematic common components underlying the target large-scale dynamics in various problems, and a sparse covariance matrix is imposed to incorporate the remaining idiosyncratic contributions to the variations and covariations. Our comments are mainly on the treatment for the idiosyncratic component, i.e. the remaining dynamics after identifying and removing the systematic part.

An important assumption of the approach proposed is that a sparse covariance matrix $\mathbf{\Sigma}_u$ is imposed for modelling the idiosyncratic component. One may naturally wonder that, in situations when a sparse $\mathbf{\Sigma}_u$ is inadequate, what alternative approach can be used for modelling the idiosyncratic component. Further, can a similar idea of parsimonious modelling by structural decomposition be extended for solving other problems such as large precision matrix estimation? In the framework of graphical models, Tang and Fan (2013) investigate the problem of large precision matrix estimation by parsimoniously modelling the idiosyncratic component by using a sparse precision matrix $\mathbf{\Omega}_u = \mathbf{\Sigma}_u^{-1}$. They observe that the large-scale precision matrix $\mathbf{\Omega}_u = \mathbf{\Sigma}^{-1}$ depends on the idiosyncratic component only through the precision matrix $\mathbf{\Omega}_u$. Thus a similar idea of structural decomposition can be equally applied for estimating the large precision matrix, with the systematic component being captured by a factor model. Facilitated by the interpretation that 0s in a precision matrix imply conditional independence between the corresponding components, a sparse $\mathbf{\Omega}_u$ can have useful practical implications. For example, in the famous Fama–French factor model (Fama and French, 1993) in finance, a non-diagonal sparse precision matrix for the idiosyncratic component characterizes the interpretable market effects among returns of stocks at different levels, such as the industrial segmentwise connections, and the intrinsic within-industry associations, say, among financial firms. Existence of such effects after removing the dynamics corresponding to the systematic component may result in a non-sparse $\mathbf{\Sigma}_u$, yet sparse modelling can still be valid by exploring the sparse precision matrix $\mathbf{\Omega}_u$.

Joong-Ho Won (*Korea University, Seoul*) **and Woncheol Jang and Johan Lim** (*Seoul National University*)

We congratulate Fan and his colleagues for a stimulating paper in which they have made a substantial contribution to challenging problems in large covariance estimation.

As practitioners, we are most interested in finite sample positive definiteness of the estimator proposed by the authors. They suggest using a scaling constant C in the threshold for the idiosyncratic covariance matrix $\mathbf{\Sigma}_{u,K}^{\mathcal{T}}$ and adjusting C to render its minimum eigenvalue positive. This idea leads to the univariate root finding procedure of expression (4.1). Although this procedure looks apparently simple, it requires computing the minimum eigenvalue of a $p \times p$ matrix, which is computationally expensive by itself for even a modest value of p, for every value of C tried. Furthermore, altering C means that the thresholding must be recomputed in every iteration, changing the sparsity pattern of the initial $\mathbf{\Sigma}_{u,K}^{\mathcal{T}}$. Thus we are concerned that the resulting cost of solving expression (4.1) may not be so cheap, especially when the target function in it is not smooth (Fig. 1).

Here we consider an alternative procedure that ensures positive definiteness while preserving the initial sparsity pattern. First, project $\mathbf{\Sigma}_{u,K}^{\mathcal{T}}$ onto a space of positive definite matrices. This can be done by solving

$$\text{minimize} \|X - \mathbf{\Sigma}_{u,K}^{\mathcal{T}}\|^2, \qquad \text{subject to } \lambda_{\min}(X) \geqslant \mu, \tag{12}$$

for a matrix variable X and some $\mu > 0$. The solution to problem (12) is given by $X^* = \sum_{i=1}^{p} \max\{\lambda_i, \mu\} q_i q_i'$ for the spectral decomposition of $\mathbf{\Sigma}_{u,K}^{\mathcal{T}} = \sum_{i=1}^{p} \lambda_i q_i q_i'$ (Boyd and Vandenberghe, 2004). Second, replace the entries of X^* that correspond to the zero-thresholded entries of $\mathbf{\Sigma}_{u,K}^{\mathcal{T}}$ with 0. Repeat these two steps until convergence. This alternating projections procedure is guaranteed to converge, as both steps are convex (Boyd and Dattorro, 2003). The first step (12) requires a spectral decomposition of $\mathbf{\Sigma}_{u,K}^{\mathcal{T}}$ as in the root finding procedure, but the second step is free of comparisons with varying thresholds.

Table 12. Comparison of the root finding and the alternating projections procedures†

Procedure	*Computing time (s)*	*Minimum eigenvalue*	*Number of iterations to converge*
POET	2.34 (0.148)	< 0	—
Root finding	62.5 (9.93)	0.149 (0.054)	20.7 (4.14)
Alternating projections	2.43 (0.118)	0.0997 (0.000)	7.91 (0.71)

†Numbers in cells are averages over 100 data sets along with their empirical standard deviations in parentheses. The code is written in Octave 3.2.3 (Eaton, 2002) on a laptop computer (MacBook Air, 1.8 GHz i5 processor with 4 Gbytes memory). Covariance hard thresholding was used in the ordinary POET with $C = 0.1$ and $K = 3$. In the root finding, Octave function `fzero()` was used to find $C_{\min}$, the root of equation (4.1), starting from $C = 0.1$; final thresholding was conducted for $C_{\min} + 0.1$. In the alternating projections, the lower bound μ for the minimum eigenvalue was set to 0.1. Both procedures terminated if $C_{\min}$ or $\lambda_{\min}$ does not change up to the third digit after the decimal point.

We numerically compared two procedures in a simple setting. The comparison was done for 100 data sets with $n = 50$ samples of a $p = 100$-dimensional standard normal vector. The results are summarized in Table 12. The root finding took roughly 1 min to converge, whereas the alternating projections converged in 2.5 s, with little additional time to the ordinary POET-estimator, i.e. POET without adjustment for positive definiteness, in less than half the iterations.

The POET-method may theoretically be optimization free, but the *post hoc* adjustment to make the ordinary POET-estimator positive definite involves some numerical optimization anyway. A little more attention to this step may greatly improve the practicality of the method proposed.

Lingzhou Xue (*Princeton University*) **and Hui Zou** (*University of Minnesota, Minneapolis*)

We first congratulate Fan, Liao and Mincheva for their innovative and timely contribution to high dimensional covariance matrix estimation. POET is a statistically and computationally appealing method for estimating a large covariance matrix with a conditional sparsity structure. We discuss two alternative methods for estimating the error covariance matrix in POET.

POET2 via positive definite adaptive thresholding estimation

POET uses adaptive thresholding estimation (Cai and Liu, 2011) on the principal orthogonal complement $\hat{\mathbf{\Sigma}}_{u,\hat{K}} = (\hat{\sigma}_{ij}^{u,\hat{K}})_{p\times p}$ to estimate the sparse error covariance matrix, namely

$$\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}} = (\hat{\sigma}_{ij}^{u,\hat{K}} I_{\{i=j\}} + s_{ij}(\hat{\sigma}_{ij}^{u,\hat{K}}) I_{\{i\neq j\}})_{p\times p}$$

where $\tau_{ij} = C w_{\mathcal{T}} \sqrt{\hat{\theta}_{ij}} > 0$ is the entry-dependent threshold. In Section 4.1 Fan, Liao and Mincheva discussed the importance of choosing a proper threshold to guarantee the finite sample positive definiteness of $\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}}$. POET chooses the threshold C in the range $(C_{\min} + \varepsilon, M)$ where $C_{\min}$ is defined in expression (4.1). Xue *et al.* (2012) proposed a direct convex programme to deliver a positive definite thresholding covariance matrix estimator. We adopt the idea thereof to construct another positive definite adaptive thresholding estimator for POET. Specifically, we consider the following constrained l_1-minimization problem:

$$\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}_2} = \arg\min_{\mathbf{\Sigma}\succeq\varepsilon\mathbf{I}} \frac{1}{2}\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_{u,\hat{K}}\|_{\mathrm{F}}^2 + \sum_{(i,j):i\neq j} \tau_{ij}|\sigma_{ij}|,$$

where $\varepsilon > 0$ is some arbitrarily small constant. The alternating direct method of multipliers algorithm in Xue *et al.* (2012) can be easily modified to solve $\hat{\Sigma}_{u,\hat{K}}^{\mathcal{T}_2}$. We introduce a new variable $\mathbf{\Theta}$ and an equality constraint $\mathbf{\Sigma} = \mathbf{\Theta}$, namely

$$(\hat{\mathbf{\Theta}}^{+}, \hat{\mathbf{\Sigma}}^{+}) = \arg\min_{\mathbf{\Theta},\mathbf{\Sigma}} \Big\{ \frac{1}{2}\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}\|_{\mathrm{F}}^2 + \sum_{(i,j):i\neq j} \tau_{ij}|\sigma_{ij}| : \mathbf{\Sigma} = \mathbf{\Theta}, \mathbf{\Theta} \succeq \varepsilon\mathbf{I} \Big\}.$$

We minimize its augmented Lagrangian function for some given parameter $\rho > 0$ (for simplicity we can fix $\rho = 1$), i.e.

$$L(\boldsymbol{\Theta}, \boldsymbol{\Sigma}; \boldsymbol{\Lambda}) = \frac{1}{2}\|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}_{u,\hat{K}}\|_{\mathrm{F}}^2 + \sum_{(i,j):i\neq j} \tau_{ij}|\sigma_{ij}| - \langle \boldsymbol{\Lambda}, \boldsymbol{\Theta} - \boldsymbol{\Sigma}\rangle + \frac{1}{2\rho}\|\boldsymbol{\Theta} - \boldsymbol{\Sigma}\|_{\mathrm{F}}^2$$

We iteratively solve $L(\boldsymbol{\Theta}, \boldsymbol{\Sigma}; \boldsymbol{\Lambda})$ for $(\boldsymbol{\Theta}^{i+1}, \boldsymbol{\Sigma}^{i+1})$ by alternating minimization, and then we update the Lagrange multiplier $\boldsymbol{\Lambda}^{i+1}$. The complete alternating direct method of multipliers algorithm proceeds as follows.

For $i = 1, 2, \ldots,$: $\boldsymbol{\Theta}$ step,

$$\boldsymbol{\Theta}^{i+1} = \arg\min_{\boldsymbol{\Theta}\geq\varepsilon\mathbf{I}} L(\boldsymbol{\Theta}, \boldsymbol{\Sigma}^i; \boldsymbol{\Lambda}^i) = (\boldsymbol{\Sigma}^i + \rho\boldsymbol{\Lambda}^i)_+;$$

$\boldsymbol{\Sigma}$ step

$$\boldsymbol{\Sigma}^{i+1} = \arg\min_{\boldsymbol{\Sigma}} L(\boldsymbol{\Theta}^{i+1}, \boldsymbol{\Sigma}; \boldsymbol{\Lambda}^i) = \frac{1}{1+\rho}(\mathrm{ST}\{\rho(\hat{\sigma}_{jk}^n - \boldsymbol{\Lambda}_{jk}^i) + \boldsymbol{\Theta}_{jk}^{i+1}, \tau_{jk}\rho\})_{p\times p};$$

$\boldsymbol{\Lambda}$ step

$$\boldsymbol{\Lambda}^{i+1} = \boldsymbol{\Lambda}^i - \frac{1}{\rho}(\boldsymbol{\Theta}^{i+1} - \boldsymbol{\Sigma}^{i+1}).$$

The two operators $(\cdot)_+$ and $\mathrm{ST}(\cdot)$ are defined in Xue *et al.* (2012).

We call $\hat{\boldsymbol{\Sigma}}_{\hat{K}}^{\mathcal{T}_2} = \boldsymbol{\Sigma}_{i=1}^{\hat{K}} \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i' + \hat{\boldsymbol{\Sigma}}_{u,\hat{K}}^{\mathcal{T}_2}$ the POET2 estimator of $\boldsymbol{\Sigma}$. We compared POET2 and POET by using simulation models 1–3 with $T = 200$ and $p = 200$ from Section 6.5.2. As can be seen from Table 13, the two versions of POET have very similar performance.

POET3 via principal orthogonal complement banding

If $\boldsymbol{\Sigma}_u$ is in fact bandable, another version of POET can use banding instead of thresholding to regularize the principal orthogonal complement. The bandable structure is widely used to model dependence between ordered variables. Given a banding parameter k, principal orthogonal complement banding yields $\hat{\boldsymbol{\Sigma}}_{u,\hat{K}}^{\mathcal{B}} = (\hat{\sigma}_{ij}^{u,\hat{K}} I(|i-j| \leqslant k))_{p\times p}$. To guarantee the positive definiteness, we consider the eigendecomposition of $\hat{\boldsymbol{\Sigma}}_{u,\hat{K}}^{\mathcal{B}} \Sigma_i \hat{\lambda}_i \mathbf{v}_i \mathbf{v}_i'$, and then define $\hat{\boldsymbol{\Sigma}}_{u,\hat{K}}^{\mathcal{B}+} = \Sigma_i \max(\hat{\lambda}, 0)\mathbf{v}_i\mathbf{v}_i'$. The POET3-estimator of $\boldsymbol{\Sigma}$ is defined as $\hat{\boldsymbol{\Sigma}}_{\hat{K}}^{\mathcal{T}_3} = \Sigma_{i=1}^{\hat{K}} \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i' + \hat{\boldsymbol{\Sigma}}_{u,\hat{K}}^{\mathcal{B}+}$. We compared POET3 and POET by using simulation models 1 and 2. As shown in Table 14, POET3 performs better than POET by taking advantage of the bandable structure. However, POET3

Table 13. Comparison of POET2 and POET in terms of average spectral norm loss over 100 replications ($T = 200$, $p = 200$)

	Results for model 1		*Results for model 2*		*Results for model 3*	
	POET	*POET2*	*POET*	*POET2*	*POET*	*POET2*
$\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|$	26.20	26.18	2.04	2.04	7.73	7.74
$\|\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1}\|$	1.31	1.30	2.07	2.06	8.48	8.50

Table 14. Comparison of POET3 and POET in terms of average spectral norm loss over 100 replications ($T = 200$, $p = 200$)

	Results for model 1		*Results for model 2*	
	POET	*POET3*	*POET*	*POET3*
$\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|$	26.20	25.76	2.04	1.68
$\|\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1}\|$	1.31	1.26	2.07	1.73

is potentially better only when the bandable structure is reliable and the ordering information is accurate. Otherwise, POET (or POET2) should be preferred.

The **authors** replied later, in writing, as follows.

We are very grateful to all contributors for their stimulating comments and questions on high dimensional covariance matrix estimation in the presence of common factors. They have touched many important issues, from theoretical understanding to methodological improvements and applications. Their contribution is important for the better understanding of the proposed POET-estimator. We shall not be able to resolve all points in a brief rejoinder. Indeed, the discussion can be seen as a collective research agenda for the future, and some of the agendas have already been undertaken by the discussants.

Spiked eigenvalues

Several discussants (Critchley, Jung and Fine, Lam and Hu, Linton and Vogt, Onatski, and Yu and Samworth) gave their detailed comments and questions regarding the spikiness of the eigenvalues. They express some concern that the separation between large and remaining eigenvalues is too distinct. Their concerns are very relevant. If there are no large gaps between the large eigenvalues and the small ones, the systematic component of the covariance cannot even be differentiated from the idiosyncratic part in our factor model: $\mathbf{\Sigma} = \mathbf{B}\mathbf{B}' + \mathbf{\Sigma}_u$. We impose the pervasiveness of the factors through the assumption that the eigenvalues of the $K \times K$ matrix

$$\mathbf{A}_p \equiv \frac{1}{p}\mathbf{B}'\mathbf{B} = \frac{1}{p}\sum_{i=1}^{p} \mathbf{b}_i \mathbf{b}_i'$$

are bounded away from both 0 and ∞ as p grows. The interpretation of this is very natural: the factors are common to the majority of variables. Under this condition and the sparsity assumption on $\mathbf{\Sigma}_u$, the first K eigenvalues are of order p whereas the remaining eigenvalues are bounded.

This pervasiveness is not the minimum condition to make the problem identifiable. As correctly pointed out by Jung and Fine, the spikiness of the eigenvalues of the low rank matrix $\mathbf{B}\mathbf{B}'$ and sparseness of $\mathbf{\Sigma}_u$ together play an important role in distinguishing the systematic and idiosyncratic components. As long as $\|\mathbf{\Sigma}_u\|$ is much smaller than $\|\mathbf{B}\mathbf{B}'\|$, these two components can be distinguished. Of course, the rates of convergence depend on the size of the gaps and other parameters. For example, Yu and Samworth suggested a weaker version of the pervasive condition, which replaces p^{-1} in the definition of $\mathbf{A}_p$ with $p^{-\alpha}$ for some $\alpha \in (0, 1)$. With this weaker condition, all results should still go through, and carefully inspecting our technical proofs should yield the rates of convergence. In contrast, there is also recent literature that requires $\alpha = 0$ or replaces $p^{-\alpha}$ with $\log(p)^{-1}$, which corresponds to approximately 'sparse loading matrices' (Pati *et al.*, 2012; Carvalho *et al.*, 2009). See also the discussion by Pan and Peng for a novel approach. Intuitively, this allows for non-pervasive (weak) factors that have no effect on a non-negligible portion of the individuals. However, this will bring more difficulty to estimating the number of spiked eigenvalues, and identifying the low rank part from the idiosyncratic part, because the signal is too weak.

We agree wholeheartedly with H. Huang, Y. Liu, Marron, D. Shen and H. Shen that now is a good time to study asymptotic contexts, where the first K eigenvalues of $\mathbf{\Sigma}$ grow quickly. Indeed, sparsity appears rarely in applications, yet conditional sparsity is likely to be more relevant for many applications. Studying spiked eigenvalues amounts to exploring the main structure of the covariance matrix.

We agree on the existence of weaker factors in applications (Lam and Hu, Linton and Vogt, and Onatski). These factors are usually difficult to differentiate from the idiosyncratic components and do not play a noticeable role without a large amount of data. We would like to add that our assumption on the spikiness of eigenvalues is imposed on the population covariance, not on the sample covariance matrix. Model diagnostics based on sample eigenvalues should be interpreted with care owing to large estimation errors in high dimensional matrices.

Choice of the number of factors K

The gaps between the spiked eigenvalues and the remaining eigenvalues have impact on the choice of the number of factors K, Fryzlewicz and N. Huang, Lam and Hu, and other discussants carried out many interesting simulations about the issue of choosing K, the number of these spiked eigenvalues. In many simulations by the contributors, the responses are not driven by a few common factors. In contrast, POET builds on the principal components analysis based on the sample covariance matrix, whose first K eigenvalues are growing at rate $O(p)$. The existence of these spiked eigenvalues is implied by the pervasive condition for the common factors. This gap can be made smaller if Yu and Samworth's assumption is

imposed instead. As pointed out by the discussants (H. Huang, Y. Liu, Marron, D. Shen and H. Shen) and Shen *et al.* (2012), the existence of spiked eigenvalues is necessary to achieve the principal components analysis consistency in the high dimension, low sample size context.

One can apply standard testing procedures to test the existence of spiked eigenvalues (e.g. Onatski (2010)), and consistently estimate the number of these eigenvalues by

$$\hat{K} = \arg \min_{0 \leqslant K \leqslant M} \frac{1}{p} \mathrm{tr}(\hat{\mathbf{R}}_K) + \mathrm{IC}(p, T) K$$

where $\hat{\mathbf{R}}_K = \Sigma_{i=K+1}^{p} \hat{\lambda}_i \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i'$ is the orthogonal complement of the sample covariance matrix; M is a given upper bound and $\mathrm{IC}(p, T)$ is one of the information criteria in Bai and Ng (2002). However, if there is no gap among the eigenvalues, then there are either no pervasive common factors (all eigenvalues are small) or too many common factors (most eigenvalues are very large). In the first case, the consistent method will estimate K as 0. In the latter case, factor analysis is inappropriate because it does not effectively reduce the dimension. Pan and Peng suggested a new way of choosing K by comparing the sample eigenvalues with a given threshold. Their method should work well even when the factors are weak. Tests based on the gaps among eigenvalues were also proposed by Onatski (2010).

We assume that K is fixed in the current version of the paper. Our working paper version allows K also to grow with p but is assumed to be known. We agree that allowing a growing and unknown K can be a good extension, as commented by Ferreira, and the first step should be consistently estimating it. This can be done by carefully reviewing the proofs in Bai and Ng (2002). Successfully solving this problem will also contribute to the literature on approximate factor models.

Generalized dynamic factor model and spectral matrix

Our paper was written for modelling large covariance matrices in genomics and finance. The former typically assumes that data are collected independently across the population, whereas the latter assumes that markets are efficient so past data play limited roles in asset returns. Hence, a conditional sparsity structure (conditioning here means specifically taking the linear dependence out) is imposed without considering the time-lagged variables.

As correctly pointed out by Hallin and Pourahmadi, allowing lagged factors is important for other applied time series problems. In addition, Farnè and Pourahmadi raise the question of estimating the spectral matrix and frequency domain analysis. Indeed, the generalized dynamic factor model, except requiring second-order stationarity, holds without any further assumptions. The method of Forni *et al.* (2000) should naturally extend POET to the frequency domain analysis, which will also enable us to estimate the autocovariance matrices and the spectral matrix. This will further broaden the scope of POET in both theory and application.

By expanding the vector of state variables, POET can be used to construct the factors that depend on the lagged variables. For a given lag q, let

$$\mathbf{x}_t = (\mathbf{y}_{t-q}', \ldots, \mathbf{y}_t')', \qquad t = q+1, \ldots, T.$$

The sample covariance matrix of $\mathbf{x}_t$ involves the cross-covariance matrices. An application of POET to the data vector $\{\mathbf{x}_t\}_{t=q+1}^{T}$ would yield the factors that are constructed on the basis of the present and past data.

Holzmann and Leister provided an interesting derivation for estimating the lagged covariance under the max-norm. In terms of the convergence rate under the max-norm, knowing the factor structure does not yield a significant improvement over the sample covarianee. This has been demonstrated and explained earlier in Fan *et al.* (2008).

Alternative and related methods

We thank several discussants for suggesting useful alternative steps to improve POET. Won, Jang and Lim propose iterative procedures to produce a finite sample positive definite error covariance matrix. A similar and conceptually simpler method is given by the contribution of Xue and Zou. Zhang and Peng also propose an iterative version of our method. The cost of the potential improvements is the loss of the simplicity of the POET method. Onatski's linear shrinkage to the principal orthogonal components provides a useful alternative approach to regularizing the error covariance matrix.

Fryzlewicz and N. Huang suggest an interesting alternative covariance estimation based on the aggregation of the sample covariance matrix (unbiased) and regularized covariance matrix (biased but with low variance). This aggregation allows a trade-off between the bias and variance in the estimation. Although NOVELIST works well in their simulations, theoretical understanding of the procedures is needed. For

example, in the approximate factor model, $\mathbf{\Sigma}$ is a dense matrix. Hence, it is difficult to explain why a thresholding is applied to the sample covariance matrix, which should also be dense.

Critchley, Dang and Yu, and Gijbels, Herrmann and Verhasselt recommend robust principal components analysis to protect against outliers and possible extensions to deal with non-stationary time series. The transformed Kendall's τ rank correlation matrix suggested by H. Liu and Wang provides an answer to the robust issue on the tails of errors, raised by Linton and Vogt. Bouveyron proposes alternative methods based on l_1-penalization. He suggests use of the covariance matrix approximation approach in Bouveyron *et al.* (2007). The effectiveness of the proposed approach for high dimensional covariance matrices hinges on good approximations and remains to be seen. We emphasize that POET works particularly well in the presence of common factors, as it takes out a few principal components in the first step of the singular value decomposition. Moreover, based on singular value decomposition, POET is optimization free except for choosing the number of factors (which involves a one-dimensional optimization). It is also adaptive to locally stationary processes through time localization and time domain smoothing.

As pointed out by Gijbels, Herrmann and Verhasselt, and Ferreira, the issue of choosing the tuning parameters for thresholding is important in practice and also exists in all regularization procedures. Besides the method suggested by Gijbels and her colleagues, additional research is still needed.

Extensions

Xue and Zou suggest a banding orthogonal complements estimator, to deal with banded idiosyncratic components. This case is asymptotically nested in POET because thresholding can also produce a banded matrix. But, if the structure is indeed 'conditionally banded' (given the common factors), their suggested method should improve the finite sample performance. Technically, it is not difficult to achieve similar rates of convergence in this case.

It is also interesting to work with the sparse inverse idiosyncratic covariance, as suggested by Tang and Fan, and H. Liu and Wang. Tang and Fan also mention a couple of interesting applications that fit into this case. Under the high dimensionality, estimating a sparse precision matrix usually involves optimizations that may introduce some computational burdens. H. Liu and Wang suggest a viable column-by-column penalized square-root lasso method to explore sparsity in inverse idiosyncratic covariance matrices, which is insensitive to the tuning parameter. Lam and Hu suggest an idea to deal with weak factors, which complements our method.

Applications

One of the immediate applications of POET is portfolio allocation, as commented by Linton and Vogt, and Gijbels and her colleagues because the problem crucially depends on estimating a high dimensional covariance matrix, and financial returns are often driven by a few common factors. Once the volatility matrix has been well estimated, we can proceed to portfolio selection via the Markowitz framework. We agree with Linton and Vogt that sparse portfolio allocation is another interesting idea to enhance the stability and the performance of portfolios. It has been thoroughly studied by Jagannathan and Ma (2003) and Fan *et al.* (2012).

We appreciate the detailed comments provided by Bailey, Pesaran and Yamagata and Coad and Maruri-Aguilar on testing the multifactor capital asset pricing model. When the POET estimator is used to replace $\mathbf{\Sigma}_u^{-1}$, if we simply bound the estimation error by

$$\frac{\tau_T' M_F \tau_T}{\sqrt{(2p)}} |\hat{\alpha}'(\mathbf{\Sigma}_u^{-1} - \hat{\mathbf{\Sigma}}_u^{-1})\hat{\alpha}| \leqslant \frac{\tau_T' M_F \tau_T}{\sqrt{(2p)}} \|\hat{\alpha}\|^2 \|\mathbf{\Sigma}_u^{-1}\hat{\mathbf{\Sigma}}_u^{-1}\|^2,$$

then indeed the upper bound is not $o_p(1)$ unless $p \log(p) \ll T$ even when the factors are observable. We would like to note that the above upper bound is too crude. To show that the estimation error is asymptotically negligible, we should not separate $\hat{\alpha}$ and $\mathbf{\Sigma}_u^{-1} - \hat{\mathbf{\Sigma}}_u^{-1}$ because $|\hat{\alpha}'(\mathbf{\Sigma}_u^{-1} - \hat{\mathbf{\Sigma}}_u^{-1})\hat{\alpha}|$ is a weighted estimator error. More careful investigation of this term can yield an improved rate of convergence. However, we wholeheartedly agree with the notion in Pesaran and Yamagata (2012) that ignoring the correlation structure in constructing testing statistics yields more stable test statistics, whose sizes of tests can be more accurately determined.

Pan and Peng, and J. Huang and Zhou connect POET with an application to the high dimensional linear regression. Indeed, when the regressors depend on a few common factors, POET can be applied to estimate their joint covariance, which will help variable selections and prediction. J. Huang and Zhou suggest the use of POET to improve sure independence screening in Fan and Lv (2008) and this can be a fruitful direction to pursue. Sparse principal components can also be used as predictors. Further research along this line is required. We thank H. Huang, Y. Liu, Marron, D. Shen and H. Shen for reminding us

of the literature on SigClust. POET works well in dealing with spiked eigenvalues, so we can foresee the success of combining POET and SigClust or other methods to estimate the principal eigenvalues in high dimension, low sample size contexts.

Ahmad, Hashmi and Halawani suggest genomics applications as an interesting test bed of POET. We agree with them. In fact, Fan and Han (2013) address large-scale hypothesis testing problems in considerable detail. It includes applications to the type of genomic data as Ahmad and his colleagues suggested.

POET is very fast to compute. We are also excited to learn about the potential applications of POET to the spatial point processes suggested by Mateu, source localization problems discussed by Jian Zhang and computer experiments suggested by Coad and Maruri-Aguilar, among others. They open broad areas where POET can be successfully applied and achieve important scientific discoveries.

Comments

Various contributors raise excellent comments. For brevity, some of them have been partially answered above, and many can be seen as a good research agenda.

We appreciate that Kent and Critchley remind us of the issues on invariance or equivariance under affine transformation. We agree that, if the measurement units change, the principal components will then change and hence POET does not have equivariance. In many applications like finance and genomics, the measurement units are comparable and affine transformations will create some interpretation problems. If the units used are a concern, one can apply POET to a correlation matrix. The sparsity of $\mathbf{\Sigma}_u$ remains intact. The equivariance issue in high dimensional problems is very challenging. It is at a very different scale of details from high dimensional data analysis. If affine transforms are considered as in Tyler *et al.* (2009), sparsity should be imposed to enhance the interpretability.

Viroli makes an interesting comment on the dual problem of estimating the factors and loadings, corresponding to the principal components analysis on either $\mathbf{YY}'$ or $\mathbf{Y}'\mathbf{Y}$, but give the same estimate for the common components. Theorem 1 in the paper is similar to what was obtained by Stock and Watson (2002b). But the result presented here allows any $K \leqslant p$, not just the true $K = \dim(\mathbf{f}_t)$.

Montanari is concerned about the precision of estimated factors in our empirical illustration. Here $p = 50 \approx T^{0.7}$. As every time series loads on the same common factors, 50 series contain enough information to estimate the factors and therefore the idiosyncratic components. So the heat maps presented in the paper indeed demonstrate the sparsity of $\mathbf{\Sigma}_u$. We feel that the spurious correlations that are created by not accurately estimated factors should be small. Besides PCA, one can perform a quasi-maximum-likelihood method, which is typically used for classical factor analysis (Lawley and Maxwell, 1971) and is also consistent under high dimensionality (Bai and Li, 2012).

Gilks gives an interesting example where each off-diagonal entry of $\mathbf{\Sigma}_u$ is a non-zero constant with probability π. This is a Bayesian perspective in which the population covariance is generated according to some probability distribution. From this perspective, in many applied problems, the probability of being 0 for each off-diagonal entry should not be a universal constant but varies over the entries. For instance, correlations between the companies in the same industry may have smaller probabilities of being 0 than those across industries. Hence we can write $\pi = \pi_{p,ij}$ for each position (i, j) and require that $\pi_{p,ij} \to 0$ fast for most of the (i, j)s.

Frommlet makes a comment on our simulated results where $\mathbf{\Sigma}$ is generated from a cross-sectional auto-regressive AR(1) process instead of a factor model. We do not claim that POET can solve all large covariance estimation problems, but indicate its power by first controlling a few relatively large eigenvalues. Since POET is reasonably robust to the overestimation of the number of factors, it works well also with sparse matrices. This explains why it works well with AR(1) covariance structure, which is effectively sparse.

Conclusion

In summary, the contributors have provided a wide range of discussions about many aspects of estimating a high dimensional covariance matrix. Many applications suggested have motivated exciting opportunities for interdisciplinary collaborations. We feel very pleased to exchange our ideas and very much look forward to new tools in this important research area. We conclude by reiterating our thanks to all the contributors, and to the Royal Statistical Society and the journal for hosting this forum.

References in the discussion

Abadir, K. M., Distaso, W. and Zikes, F. (2010) Model-free estimation of large variance matrices. *Working Paper 10–17*. Rimini Centre For Economic Analysis, Rimini.

Agarwal, A., Negahban, S. and Wainwright, M. J. (2012) Noisy matrix decomposition via convex relaxation: optimal rates in high dimensions. *Ann. Statist.*, **40**, 1171–1197.
Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D. and Levine, A. J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natn. Acad. Sci. USA*, **96**, 6745–6750.
Amini, A. A. and Wainwright, M. J. (2009) High-dimensional analysis of semidefinite relaxations for sparse principal components. *Ann. Statist.*, **37**, 2877–2921.
Antoniadis, A. (2007) Wavelet methods in statistics: some recent developments and their application. *Statist. Surv.*, **1**, 16–55.
Bai, J. and Li, K. (2012) Statistical analysis of factor models of high dimension. *Ann. Statist.*, **40**, 436–465.
Bai, J. and Ng, S. (2002) Determining the number of factors in approximate factor models. *Econometrica*, **70**, 191–221.
Bai, Z. D. and Yao, J. F. (2008) Central limit theorems for eigenvalues in a spiked population model. *Ann. Inst. H. Poincaré*, **44**, 447–474.
Baik, J. and Silverstein, J. W. (2006) Eigenvalues of large sample covariance matrices of spiked population models. *J. Multiv. Anal.*, **97**, 1382–1408.
Bickel, P. and Levina, E. (2008) Covariance regularization by thresholding. *Ann. Statist.*, **36**, 2577–2604.
Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009) Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, **37**, 1705–1732.
Boehm, H. and von Sachs, R. (2008) Structural shrinkage of nonparametric spectral estimators for multivariate time series. *Electron. J. Statist.*, **2**, 696–721.
Boehm, H. and von Sachs, R. (2009) Shrinkage estimation in the frequency domain of multivariate time series. *J. Multiv. Anal.*, **100**, 913–935.
Bouveyron, C., Girard, S. and Schmid, C. (2007) High-dimensional data clustering. *Computnl Statist. Data Anal.*, **52**, 502–519.
Boyd, S. and Dattorro, J. (2003) Alternating projections. (Available from `http://www.stanford.edu/class/ee392o/alt_proj.pdf`.)
Boyd, S. and Vandenberghe, L. (2004) *Convex Optimization*. Cambridge: Cambridge University Press.
Brillinger, D. R. (1981) *Time Series: Data Analysis and Theory*. San Francisco: Holden-Day.
Byers, S. and Raftery, A. E. (1998) Nearest-neighbor clutter removal for estimating features in spatial point processes. *J. Am. Statist. Ass.*, **93**, 577–584.
Cai, T. and Liu, W. (2011) Adaptive thresholding for sparse covariance matrix estimation. *J. Am. Statist. Ass.*, **106**, 672–684.
Carvalho, C., Chang, J., Lucas, J., Nevins, J., Wang, Q. and West, M. (2008) High-dimensional sparse factor modeling: applications in gene expression genomics. *J. Am. Statist. Ass.*, **103**, 1438–1456.
Chamberlain, G. and Rothschild, M. (1983) Arbitrage, factor structure and mean-variance analysis in large asset markets. *Econometrica*, **51**, 1305–1324.
Chen, B. B. and Pan, G. M. (2012) Convergence of the largest eigenvalue of normalized sample covariance matrices when p and n both tend to infinity with their ratio converging to zero. *Bernoulli*, **18**, 1405–1420.
Chudik, A., Pesaran, M. H. and Tosetti, E. (2011) Weak and strong cross-section dependence and estimation of large panels. *Econmetr. J.*, **14**, 45–90.
Collins, L. B. and Cressie, N. (2001) Analysis of spatial point patterns using bundles product LISA function. *J. Agric. Biol. Environ. Statist.*, **6**, 118–135.
Eaton, J. W. (2002) *GNU Octave Manual*. Bristol: Network Theory Limited.
Engelen, S., Hubert, M. and Vanden Branden, K. (2005) A comparison of three procedures for Robust PCA in high dimensions. *Austn J. Statist.*, **34**, 117–126.
Fama, E. and French, K. R. (1993) Common risk factors in the returns on stocks and bonds. *J. Finan. Econ.*, **33**, 3–56.
Fan, J., Fan, Y. and Lv, J. (2008) High dimensional covariance matrix estimation using a factor model. *J. Econmetr.*, **147**, 186–197.
Fan, J. and Han, X. (2013) Estimation of false discovery proportion with unknown dependence. *Manuscript*.
Fan, J., Liao, Y. and Mincheva, M. (2011) High dimensional covariance matrix estimation in approximate factor models. *Ann. Statist.*, **39**, 3320–3356.
Fan, J. and Lv, J. (2008) Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. R. Statist. Soc.* B, **70**, 849–911.
Fan, J., Zhang, J. and Yu, K. (2012) Vast portfolio selection with gross-exposure constraints. *J. Am. Statist. Ass.*, **107**, 592–606.
Farnè, M. and Montanari, A. (2013) Different estimators of the spectral matrix: an empirical comparison; testing a new shrinkage estimator. *Communs Statist. Theor. Meth.*, to be published.
Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2000) The generalized dynamic factor model: identification and estimation. *Rev. Econ. Statist.*, **82**, 540–554.
Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2004) The generalized dynamic factor model consistency and rates. *J. Econmetr.*, **119**, 231–255.

Forni M. and Lippi, M. (2001) The generalized dynamic factor model: representation theory. *Econmetr. Theor.*, **17**, 1113–1341.
Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gassenbeek, M., Mesirov, J. P., Coller, H., Loh, M., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.
Hall, P., Marron, J. S. and Neeman, A. (2005) Geometric representation of high dimension, low sample size data. *J. R. Statist. Soc.* B, **67**, 427–444.
Huang, H., Liu, Y., Yuan, M. and Marron, J. S. (2013) Statistical significance of clustering using soft thresholding. *Preprint arXiv:1305.5879*.
Jagannathan, R. and Ma, T. (2003) Risk reduction in large portfolios: why imposing the wrong constraints helps. *J. Finan.*, **58**, 1651–1683.
Johnstone, I. M. and Lu, A. Y. (2009) On consistency and sparsity for principal components analysis in high dimensions. *J. Am. Statist. Ass.*, **104**, 682–693.
Jolliffe, I. T. (2002) *Principal Component Analysis*, 2nd edn. New York: Springer.
Jung, S. and Marron, J. S. (2009) PCA consistency in high dimension, low sample size context. *Ann. Statist.*, **37**, no. 6B, 4104–4130.
Jung, S., Sen A. and Marron, J. (2012) Boundary behavior in High Dimension, Low Sample Size asymptotics of PCA. *J. Multiv. Anal.*, **109**, 196–203.
Kaufman, C. G., Bingham, D., Habib, S., Heitmann, K. and Frieman, J. A. (2011) Efficient emulators of computer experiments using compactly supported correlation functions with an application to cosmology. *Ann. Appl. Statist.*, **5**, 2470–2492.
Kaufman, C. G., Schervish, M. J. and Nychka, D. W. (2008) Covariance tapering for likelihood-based estimation in large spatial data sets. *J. Am. Statist. Ass.*, **103**, 1545–1555.
Ke, T., Jin, J. and Fan, J. (2012) Covariance assisted screening and estimation. *Preprint arXiv:1205.4645v2*.
Kourtis, A., Dotsis, G. and Markellos, R. (2012) Parameter uncertainty in portfolio selection: shrinking the inverse covariance matrix. *J. Bankng Finan.*, **36**, 2522–2531.
Lam, C. and Yao, Q. (2012) Factor modeling for high-dimensional time series: inference for the number of factors. *Ann. Statist.*, **40**, 694–726.
Lam, C., Yao, Q. and Bathia, K. (2011) Estimation of latent factors for high-dimensional time series. *Biometrika*, **98**, 901–918.
Lansangan, J. R. G. and Barrios, E. B. (2009) Principal components analysis of nonstationary time series data. *Statist. Comput.*, **19**, 173–187.
Lawley, D. and Maxwell, A. (1971) *Factor Analysis as a Statistical Method*, 2nd edn. London: Butterworth.
Ledoit, O. and Wolf, M. (2003) Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *J. Empir. Finan.*, **10**, 603–621.
Ledoit, O. and Wolf, M. (2004) A well-conditioned estimator for large-dimensional covariance matrices. *J. Multiv. Anal.*, **88**, 365–411.
Liu, H., Han, F., Yuan, M., Lafferty, J. and Wasserman, L. (2012) High dimensional semiparametric gaussian copula graphical models. *Ann. Statist.*, **40**, 2293–2326.
Liu, Y., Hayes, D. N., Nobel, A. and Marron, J. S. (2008) Statistical significance of clustering for high-dimension, low-sample size data. *J. Am. Statist. Ass.*, **103**, 1281–1293.
Liu, H. and Wang, L. (2012) Tiger: a tuning-insensitive approach for optimally estimating gaussian graphical models. Princeton University, Princeton. (Available from `http://arxiv.org/abs/1209.2437`.)
Markowitz, H. (1952) Portfolio selection. *J. Finan.*, **7**, 77–91.
Onatski, A. (2009) Testing hypotheses about the number of factors in large factor models. *Econometrica*, **77**, 1447–1479.
Onatski, A. (2010) Determining the number of factors from empirical distribution of eigenvalues. *Rev. Econ. Statist.*, **92**, 1004–1016.
Pati, D., Bhattacharaya, A., Pillai, N. and Dunson, D. (2012) Posterior contraction in sparse Bayesian factor models for massive covariance matrices. *Manuscript*. Duke University, Durham.
Paul, D. (2007) Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sin.*, **17**, 1617–1642.
Pesaran, M. H. and Yamagata, T. (2012) Testing CAPM with a large number of assets. *American Finance Association San Diego Meetings Paper*. (Available from `http://ssrn.com/abstracts`.)
Pourahmadi, M. (2013) *High-dimensional Covariance Estimation*. New York: Wiley.
Schaefer, J. and Strimmer, K. (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statist. Applic. Genet. Molec. Biol.*, **4**, article 32.
Shen, D., Shen, H. and Marron, J. (2012) A general framework for consistency of principal component analysis. *Preprint arXiv:1211.2671*.
Shen, D., Shen, H. and Marron, J. S. (2013) Consistency of sparse PCA in high dimension, low sample size contexts. *J. Multiv. Anal.*, **115**, 317–333.
Stock, J. H. and Watson, M. W. (2002a) Macroeconomic forecasting using diffusion indexes. *J. Bus. Econ. Statist.*, **20**, 147–162.

Stock, J. H. and Watson, M. W. (2002b) Forecasting using principal components from a large number of predictors. *J. Am. Statist. Ass.*, **97**, 1167–1179.
Stock, J. H. and Watson, M. W. (2005) Implications of dynamic factor models for VAR analysis. *Working Paper 11467*. National Bureau of Economic Research, Cambridge.
Tang, C. Y. and Fan, Y. (2013) Precision matrix estimation by inverse principle orthogonal decomposition. *Manuscript*.
Tyler, D. E., Critchley, F., Dümbgen, L. and Oja, H. (2009) Invariant co-ordinate selection (with discussion). *J. R. Statist. Soc.* B, **71**, 549–592.
Wang, H. (2012) Factor profiled sure independence screening. *Biometrika*, **99**, 15–28.
Won, J.-H., Lim, J., Kim, S.-J. and Rajaratnam, B. (2013) Condition-number-regularized covariance estimation. *J. R. Statist. Soc.* B, **75**, 427–450.
Xue, L., Ma, S. and Zou, H. (2012) Positive definite l_1 penalized estimation of large covariance matrices. *J. Am. Statist. Ass.*, **107**, 1480–1491.
Yen, Y. M. (2011) Sparse weighted norm minimum variance portfolio. *Preprint*.
Zehetmayer, S., Bauer, P. and Posch, M. (2005) Two-stage designs for experiments with a large number of hypotheses. *Bioinformatics*, **21**, 3771–3777.
Zehetmayer, S., Bauer, P. and Posch, M. (2008) Optimized multi-stage designs controlling the false discovery or the family-wise error rate. *Statist. Med.*, **27**, 4145–4160.
Zhang, J. (2012) Source screening for neuroimaging. *Technical Report*. School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury.
Zhang, J., Liu, C. and Green, G. (2012) Source localization with MEG data. *Technical Report*. School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury.