Large Covariance Estimation by Thresholding Principal Orthogonal Complements

Jianqing Fan[*,†], **Yuan Liao**[‡], and **Martina Mincheva**[*]

[*]Department of Operations Research and Financial Engineering, Princeton University [†]Bendheim Center for Finance, Princeton University [‡]Department of Mathematics, University of Maryland

Abstract

This paper deals with the estimation of a high-dimensional covariance with a conditional sparsity structure and fast-diverging eigenvalues. By assuming 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 (POET) method to explore such an approximate factor structure with sparsity. The POET estimator includes the sample covariance matrix, the factor-based covariance matrix (Fan, Fan, and Lv, 2008), the thresholding estimator (Bickel and Levina, 2008) and the adaptive thresholding estimator (Cai and Liu, 2011) 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 impact 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

High-dimensionality; approximate factor model; unknown factors; principal components; sparse matrix; low-rank matrix; thresholding; cross-sectional correlation; diverging eigenvalues

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, Fan and Lv, 2008), high-dimensional classification such as Fisher discriminant (Hastie, Tibshirani and Friedman, 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, Fan, and Wu, 2009; Xiong et al. 2011), and testing the capital asset pricing model (Sentana, 2009), among others. See Section 5 for some of those applications. Yet, the dimensionality is often either comparable to 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 brought by the high dimensionality, in recent years researchers have proposed various regularization techniques to consistently estimate Σ. One of the key assumptions is that the covariance matrix is sparse, namely, many entries are zero or nearly so (Bickel and Levina, 2008, Rothman et al, 2009, Lam and Fan 2009, Cai and Zhou, 2010, Cai and Liu, 2011). In many applications, however, the sparsity assumption directly on Σ is not appropriate. For example, financial returns depend on the equity market risks, housing prices depend on the economic health, gene expressions can be stimulated by cytokines, among others. Due to the presence of common factors, it is unrealistic to assume that many outcomes are uncorrelated. An alternative method is to assume a factor model structure, as in Fan, Fan and Lv (2008). However, they restrict themselves to the strict factor models with known factors.

A natural extension is the conditional sparsity. Given the common factors, the outcomes are weakly correlated. In order 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 1993; Bai and Ng, 2002, etc):

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

Here y_{it} is the observed response for the ith ($i = 1, ..., p$) individual at time $t = 1, ..., T$; $\mathbf{b}_i$ 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, usually called *idiosyncratic component*, uncorrelated with $\mathbf{f}_t$. Both p and T diverge to infinity, while 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, that is, $p \to \infty$. In a data-rich environment, p can diverge at a rate faster than T. The factor model (1.1) can be put in a matrix form as

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

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

$$\sum = \mathbf{B}\mathrm{cov}(\mathbf{f}_t)\mathbf{B}' + \sum\nolimits_u, \quad (1.3)$$

where $\Sigma_u = (\sigma_{u,ij})_{p \times p}$ 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 Σ_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 Σ_u is *approximately sparse* as in Bickel and Levina (2008) and Rothman et al. (2009): for some $q \in [0, 1)$,

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

does not grow too fast as $p \to \infty$. In particular, this includes the exact sparsity assumption ($q = 0$) under which $m_p = \max_i {}_p \Sigma_j {}_p I_{(\sigma_{u,ij} \ 0)}$, the maximum number of nonzero elements in each row.

The conditional sparsity structure of (1.2) was explored by Fan, Liao and Mincheva (2011) 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 have to be inferred. Our approach is simple, optimization-free and it uses the data only through the sample covariance matrix. Run the singular value decomposition on the sample covariance matrix $\hat{\Sigma}_{\text{sam}}$ of $\mathbf{y}_t$, keep the covariance matrix 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 (POET) estimator. When the number of common factors K is unknown, it can be estimated from the data. See Section 2 for additional details. We will 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 Σ and Σ_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 Σ_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 ten years, much theory on the estimation and inference of the static factor model has been developed, for example, Stock and Watson (1998, 2002), Bai and Ng (2002), Bai (2003), Doz, Giannone and Reichlin (2011), among others. Our contribution is on the estimation of covariance matrices and their inverse in large factor models.

The *static* model considered in this paper is to be distinguished from the *dynamic factor model* as in Forni, Hallin, Lippi and Reichlin (2000); the latter allows $\mathbf{y}_t$ to also depend 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 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, Bach and El Ghaoui (2008), Shen and Huang (2008), Witten, Tibshirani, and Hastie (2009) and Ma (2011) 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), 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, Nagahban, Wainwright (2012), 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 Σ are spiked and grow at a rate $O(p)$, whereas the eigenvalues of the matrices studied in the existing literature on covariance estimation are usually assumed to be either bounded or slowly growing. Due to 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 spanned by the eigenvectors of Σ. 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 Σ in absolute term, but it can be estimated accurately in relative term (see Section 3.3 for details). In addition, Σ^{-1} can be estimated accurately.

We would like to further note 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), Lin et al. (2009)[1] considered such a representation on the data matrix. As there is no Σ 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 purpose, we require the idiosyncratic components and common factors to be uncorrelated and the data generating process to be strictly stationary. The covariances 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 Σ_u demonstrates that the 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 principal components analysis 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 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 the appendix. 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 $\|\mathbf{A}\|_F$, $\|\mathbf{A}\|$, $\|\mathbf{A}\|_1$ and $\|\mathbf{A}\|_{\max}$ the Frobenius norm, spectral norm (also called operator norm), L_1-norm, and elementwise norm of a matrix $\mathbf{A}$, defined respectively by $\|\mathbf{A}\|_F = \mathrm{tr}^{1/2}(\mathbf{A}'\mathbf{A})$,

$\|\mathbf{A}\| = \lambda_{\max}^{1/2}(\mathbf{A}'\mathbf{A})$, $\|\mathbf{A}\|_1 = \max_j \Sigma_i |a_{ij}|$ and $\|\mathbf{A}\|_{\max} = \max_{i,j} |a_{ij}|$. Note that when $\mathbf{A}$ is a vector, both $\|\mathbf{A}\|_F$ 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 $b_T = O(a_T)$.

---

[1]We thank a referee for reminding us these related works.

2 Regularized Covariance Matrix via PCA

There are three main objectives of this paper: (i) understand the relationship between principal component analysis (PCA) and the high-dimensional factor analysis; (ii) estimate both covariance matrices Σ and the idiosyncratic Σ_u and their precision matrices in the presence of common factors, and (iii) investigate the impact of estimating the unknown factors on the covariance estimation. The propositions in Section 2.1 below show that the space spanned by the principal components in the population level Σ is close to the space spanned by the columns of the factor loading matrix $\mathbf{B}$.

2.1 High-dimensional PCA and factor model

Consider a factor model

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

where the number of common factors, $K = \dim(\mathbf{f}_t)$, is small compared to 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 Σ are no longer bounded, but growing fast with the dimensionality. We illustrate this in the following example.

Example 2.1: 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 impact on a non-vanishing proportion of outcomes. It is then reasonable to assume $\sum_{i=1}^p b_i^2 > cp$ for some $c > 0$. Therefore, assuming that $\lambda_{\max}(\Sigma_u) = o(p)$, an application of (1.3) yields,

$$\lambda_{\max}(\sum) \ge \mathrm{var}(f_t) \sum_{i=1}^p b_i^2 - \lambda_{\max}(\sum_u) > \frac{c}{2} \mathrm{var}(f_t) p$$

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

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 spanned by the first K principal components of $\mathbf{B}\mathrm{cov}(\mathbf{f}_t)\mathbf{B}'$ is the same as that spanned by the columns of $\mathbf{B}$ when $\mathrm{cov}(\mathbf{f}_t)$ is non-degenerate. Thus, we can assume without loss of generality that the columns of $\mathbf{B}$ are orthogonal and $\mathrm{cov}(\mathbf{f}_t) = \mathbf{I}_K$, the identity matrix. This canonical form corresponds to the identifiability condition in decomposition (1.3). Let $\tilde{\mathbf{b}}_1, \cdots, \tilde{\mathbf{b}}_K$ 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 zero. We will impose the pervasiveness assumption that all eigenvalues of the $K \times K$ matrix $p^{-1}\mathbf{B}'\mathbf{B}$ are bounded away from zero, 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 Σ in a descending order and $\{\boldsymbol{\xi}_j\}_{j=1}^p$ be their corresponding eigenvectors. Then, an application of Weyl's eigenvalue theorem (see the appendix) yields that

Proposition 2.1: Assume that the eigenvalues of $p^{-1}\mathbf{B}'\mathbf{B}$ are bounded away from zero for all large p. For the factor model (1.3) with the canonical condition

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

we have

$$\left|\lambda_j-\left\|\tilde{\mathbf{b}}_j\right\|^2\right| \le \left\|\textstyle\sum_u\right\|, \quad \text{for } j \le K, \quad |\lambda_j| \le \left\|\textstyle\sum_u\right\|, \quad \text{for } j>K.$$

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

Using Proposition 2.1 and the $\sin\theta$ theorem of Davis and Kahn (1970, see the appendix), we have the following:

Proposition 2.2: Under the assumptions of Proposition 2.1, if $\left\{\left\|\tilde{\mathbf{b}}_j\right\|\right\}_{j=1}^K$ are distinct, then

$$\left\|\boldsymbol{\xi}_j-\tilde{\mathbf{b}}_j/\|\tilde{\mathbf{b}}_j\|\right\|=O(p^{-1}\|\textstyle\sum_u\|), \quad \text{for } j \le K.$$

Propositions 2.1 and 2.2 state that PCA and factor analysis are approximately the same if $\|\Sigma_u\| = o(p)$. This is assured through a sparsity condition on $\Sigma_u = (\sigma_{u,ij})_{p\times p}$, which is frequently measured through

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

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

$$\left\|\textstyle\sum_u\right\| \le \left\|\textstyle\sum_u\right\|_1 \le \max_i\sum_{j=1}^p|\sigma_{u,ij}|^q(\sigma_{u,ii}\sigma_{u,jj})^{(1-q)/2}=O(m_p),$$

if the noise variances $\{\sigma_{u,ii}^2\}$ are bounded. Therefore, when $m_p = o(p)$, Proposition 2.1 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 2.2 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 forecast[2]. More recently, Phan (2012) empirically estimated the level of sparsity of the idiosyncratic covariance using the UK market data.

Recent developments on random matrix theory, for example, Johnstone and Lu (2009) and Paul (2007), have shown that when p/T is not negligible, the eigenvalues and eigenvectors of Σ might not be consistently estimated from the sample covariance matrix. A distinguished feature of the covariance considered in this paper is that there are some very spiked eigenvalues. By Propositions 2.1 and 2.2, in the factor model, the pervasiveness condition

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

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 of the space spanned by the columns of the factor loading matrix $\mathbf{B}$. In addition, due to (2.3), the signals of the first K eigenvalues are stronger than those of the spiked covariance model 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. On the other hand, this also shows that, under our setting the PCA is a valid approximation to factor analysis only if $p \to \infty$. The fact that the PCA on the sample covariance is inconsistent when p is bounded was also previously demonstrated in the literature (See e.g., Bai (2003)).

With assumption (2.3), the standard literature on approximate factor models has shown that the PCA on the sample covariance matrix $\widehat{\Sigma}_{\text{sam}}$ can consistently estimate the space spanned by the factor loadings (e.g., Stock and Watson (1998), Bai (2003)). Our contribution in Propositions 2.1 and 2.2 is that we connect the high-dimensional factor model to the principal components, and obtain the consistency of the spectrum in the population level Σ instead of the sample level $\widehat{\Sigma}_{\text{sam}}$. 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 POET

Sparsity assumption directly on Σ is inappropriate in many applications due to the presence of common factors. Instead, we propose a nonparametric estimator of Σ based on the principal component analysis. Let $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$ be the ordered eigenvalues of the sample covariance matrix $\widehat{\Sigma}_{\text{sam}}$ and $\{\hat{\boldsymbol{\xi}}_i\}_{i=1}^p$ be their corresponding eigenvectors. Then the sample covariance has the following spectral decomposition:

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

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

[2]We thank a referee for this interesting reference.

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

$$\hat{\mathbf{R}}_K^{\tau} = (\hat{r}_{ij}^{\tau})_{p \times p}, \quad \hat{r}_{ij}^{\tau} = \begin{cases} \hat{r}_{ii}, & i = j; \\ s_{ij}(\hat{r}_{ij}) I(|\hat{r}_{ij}| \geq \tau_{ij}), & i \neq j. \end{cases} \quad (2.5)$$

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 $s_{ij}(x) = xI(|x| \quad \tau_{ij})$ (Bickel and Levina, 2008) and the constant thresholding parameter $\tau_{ij} = \delta$ are allowed. In practice, it is more desirable to have τ_{ij} be entry-adaptive. An example of the adaptive thresholding is

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

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 Σ is then defined as:

$$\widehat{\sum}_K = \sum_{i=1}^{K} \hat{\lambda}_i \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i' + \hat{\mathbf{R}}_K^{\mathscr{T}}. \quad (2.7)$$

We will call this estimator the Principal Orthogonal complEment thresholding (POET) estimator. It is obtained by thresholding the remaining components of the sample covariance matrix, after taking out the first K principal components. One of the attractiveness of POET is that it is optimization-free, and hence is computationally appealing. [3]

With the choice of τ_{ij} in (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, Fan, and Lv, 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 based on the sample covariance matrix. Determining K in a data-driven way is an important topic, and is well understood in the literature. We will describe the POET with a data-driven K in Section 2.4.

2.3 Least squares point of view

The POET (2.7) has an equivalent representation using a constrained least squares method.

The least squares method seeks for $\hat{\mathbf{\Lambda}}_K = (\hat{\mathbf{b}}_1^K, \ldots, \hat{\mathbf{b}}_p^K)'$ and $\hat{\mathbf{F}}_K' = (\hat{\mathbf{f}}_1^K, \ldots, \hat{\mathbf{f}}_T^K)$ such that

---

[3]We have written an R package for POET, which outputs the estimated Σ, Σ_u, K, the factors and loadings.

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

subject to the normalization

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

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, that is, $Ey_{it} = Ef_{jt} = 0$ for all i p, j K and t T. Putting it in a matrix form, the optimization problem can be written as

$$\arg\min_{\mathbf{B},\mathbf{F}}\left\|\mathbf{Y} - \mathbf{B}\mathbf{F}'\right\|_F^2$$
$$T^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_K, \quad \mathbf{B}'\mathbf{B} \text{ is diagonal.} \quad (2.10)$$

where $\mathbf{Y} = (\mathbf{y}_1, ..., \mathbf{y}_T)$ and $\mathbf{F}' = (\mathbf{f}_1, \cdots, \mathbf{f}_T)$. 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 (2.10), the objective function now becomes $\left\|\mathbf{Y} - T^{-1}\mathbf{Y}\mathbf{F}\mathbf{F}'\right\|_F^2 = \text{tr}[(I_T - T^{-1}\mathbf{F}\mathbf{F}')\mathbf{Y}'\mathbf{Y}]$. The minimizer is now clear: the columns of $\widehat{\mathbf{F}}_K/\sqrt{T}$ are the eigenvectors corresponding to the K largest eigenvalues of the $T \times T$ matrix $\mathbf{Y}'\mathbf{Y}$ and $\widehat{\mathbf{\Lambda}_K} = T^{-1}\mathbf{Y}\widehat{\mathbf{F}_K}$ (see e.g., Stock and Watson (2002)).

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

$$\widehat{\sum}_{u,K}^{\mathscr{T}} = (\widehat{\sigma}_{ij}^{\mathscr{T}})_{p\times p}, \quad \widehat{\sigma}_{ij}^{\mathscr{T}} = \begin{cases} \widehat{\sigma}_{ii}, & i=j \\ s_{ij}(\widehat{\sigma}_{ij}), & i \neq j, \end{cases} \quad (2.11)$$

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

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

It is easy to verify that $s_{ij}(\cdot)$ includes many interesting thresholding functions such as the hard thresholding ($s_{ij}(z) = zI_{(|z| \ \tau_{ij})}$), soft thresholding ($s_{ij}(z) = \text{sign}(z)(|z| - \tau_{ij})_+$), SCAD, and adaptive lasso (See Rothman et al. (2009)).

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

$$\tilde{\sum}_K = \hat{\mathbf{\Lambda}}_K \hat{\mathbf{\Lambda}}'_K + \widehat{\sum}_{u,K}^{\mathscr{T}}, \quad (2.12)$$

and by the Sherman-Morrison-Woodbury formula, noting that $\frac{1}{T}\sum_{t=1}^{T} \hat{\mathbf{f}}_t^K \hat{\mathbf{f}}_t^{K'} = \mathbf{I}_K$,

$$\left(\tilde{\sum}_K\right)^{-1} = \left(\widehat{\sum}_{u,K}^{\mathscr{T}}\right)^{-1} - \left(\widehat{\sum}_{u,K}^{\mathscr{T}}\right)^{-1}\hat{\mathbf{\Lambda}}_K\left[\mathbf{I}_K + \hat{\mathbf{\Lambda}}'_K\left(\widehat{\sum}_{u,K}^{\mathscr{T}}\right)^{-1}\hat{\mathbf{\Lambda}}_K\right]^{-1}\hat{\mathbf{\Lambda}}'_K\left(\widehat{\sum}_{u,K}^{\mathscr{T}}\right)^{-1}, \quad (2.13)$$

In practice, the true number of factors K might be unknown to us. However, for any

determined $K_1 \quad p$, we can always construct either $(\Sigma_{\hat{K}_1}, \hat{\mathbf{R}}_{K_1}^{\mathscr{T}})$ as in (2.7) or $(\Sigma_{\tilde{K}_1}, \widehat{\sum}_{u,K_1}^{\mathscr{T}})$ as in (2.12) to estimate (Σ, Σ_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 2.1: Suppose that the entry-dependent threshold in (2.5) is the same as the thresholding parameter used in (2.11). Then for any $K_1 \quad p$, the estimator (2.7) is equivalent to the substitution estimator (2.12), that is,

$$\widehat{\sum}_{K_1} = \tilde{\sum}_{K_1}, \quad and \quad \widehat{\sum}_{u,K_1}^{\mathscr{T}} = \hat{\mathbf{R}}_{K_1}^{\mathscr{T}}.$$

In this paper, we will use a data-driven $\hat{K}$ to construct the POET (see Section 2.4 below), which has two equivalent representations according to Theorem 2.1.

2.4 POET with Unknown *K*

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

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

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

where *M* is a prescribed upper bound, $\widehat{\mathbf{F}}_{K_1}$ 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 $\mathbf{Y}'\mathbf{Y}$; $g(T, p)$ 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) are

$$\text{IC1}: g(T, p) = \frac{p+T}{pT} \log \left(\frac{pT}{p+T} \right),$$
$$\text{IC2}: g(T, p) = \frac{p+T}{pT} \log \min\{p, T\}.$$

Throughout the paper, we let $\hat{K}$ be the solution to (2.14) 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

$$\widehat{\sum}_{\hat{K}} = \sum_{i=1}^{\hat{K}} \hat{\lambda}_i \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i' + \hat{\mathbf{R}}_{\hat{K}}^{\mathscr{T}}. \quad (2.15)$$

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 the 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 ones 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), Bai (2003), etc.).

Assumption 3.1: All the eigenvalues of the $K \times K$ matrix $p^{-1}\mathbf{B}'\mathbf{B}$ are bounded away from both zero and infinity as $p \to \infty$.

Remark 3.1

1. It implies from Proposition 2.1 in Section 2 that the first K eigenvalues of Σ grow at rate $O(p)$. This unique feature distinguishes our work from most of other low-rank plus sparse covariances considered in the literature, e.g., Luo (2011), Pati et al. (2012), Agarwal et al. (2012), Birnbaum et al. (2012). [4]

2. Assumption 3.1 requires the factors to be pervasive, that is, to impact a non-vanishing proportion of individual time series. See Example 2.1 for its meaning. [5]

3. As to be illustrated in Section 3.3 below, due to the fast diverging eigenvalues, one can hardly achieve a good rate of convergence for estimating Σ 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 to those convergence results in the existing literature.

---

[4]To our best knowledge, the only other papers that estimate large covariances with diverging eigenvalues (growing at the rate of dimensionality $O(p)$) are Fan et al. (2008, 2011) and Bai and Shi (2011). While Fan et al. (2008, 2011) assumed the factors are observable, Bai and Shi (2011) considered the strict factor model in which Σ_u is diagonal.

[5]It is important to distinguish the model we consider in this paper from the "sparse factor model" in the literature, e.g., Carvalho et al. (2009), 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 Σ are bounded.

Assumption 3.2

i. $\{\mathbf{u}_t, \mathbf{f}_t\}_{t \geq 1}$ is strictly stationary. In addition, $Eu_{it} = Eu_{it}f_{jt} = 0$ for all $i \leq p, j \leq K$ and $t \leq T$.

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

iii. There exist $r_1, r_2 > 0$ and $b_1, b_2 > 0$, such that for any $s > 0$, $i \leq p$ and $j \leq K$,

$$P(|u_{it}| > s) \leq \exp(-(s/b_1)^{r_1}), \qquad P(|f_{jt}| > s) \leq \exp(-(s/b_2)^{r_2}).$$

Condition (i) 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 (ii) requires that $\mathbf{\Sigma}_u$ be well-conditioned. The condition $\|\mathbf{\Sigma}_u\|_1 \leq c_2$ instead of a weaker condition $\lambda_{\max}(\mathbf{\Sigma}_u) \leq c_2$ is imposed here in order to consistently estimate K. But it is still standard in the approximate factor model literature as in Bai and Ng (2002), Bai (2003), etc. When K is known, such a condition can be removed. Our working paper[6] shows that the results continue to hold for a growing (known) K under the weaker condition $\lambda_{\max}(\mathbf{\Sigma}_u) \leq c_2$. Condition (iii) requires exponential-type tails, which allows us to apply the large deviation theory to

$$\frac{1}{T}\sum_{t=1}^{T} u_{it}u_{jt} - \sigma_{u,ij} \text{ and } \frac{1}{T}\sum_{t=1}^{T} f_{jt}u_{it}.$$

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

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

Assumption 3.3: 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 $T \in \mathbb{Z}^{+}$,

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

In addition, we impose the following regularity conditions.

Assumption 3.4: There exists $M > 0$ such that for all $i \leq p, t \leq T$ and $s \leq T$,

i. $\|\mathbf{b}_i\|_{\max} < M$,

ii. $E[p^{-1/2}(\mathbf{u}_s'\mathbf{u}_t - E\mathbf{u}_s'\mathbf{u}_t)]^4 < M$,

iii. $E\left\|p^{-1/2}\sum_{i=1}^{p}\mathbf{b}_i u_{it}\right\|^4 < M$.

These conditions are needed to consistently estimate the transformed common factors as well as the factor loadings. Similar conditions were also assumed in Bai (2003), and Bai and

---

[6] See Fan, Liao and Mincheva (2011), working paper, arxiv.org/pdf/1201.0175.pdf

Ng (2006). The number of factors is assumed to be fixed. Our conditions in Assumption 3.4 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 Σ_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, Han and Gu, 2012). See Section 5.

We estimate Σ_u by thresholding the principal orthogonal complements after the first $\hat{K}$ principal components of the sample covariance are taken out: $\widehat{\sum}_{u,\hat{K}}^{\mathscr{T}} = \hat{\mathbf{R}}_{\hat{K}}^{\mathscr{T}}$. By Theorem 2.1, it also has an equivalent expression given by (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

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

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 chosen in the current paper is in fact obtained from the optimal uniform rate of convergence of $\max_{i \le p, j \le p} |\sigma_{ij} - \sigma_{u,ij}|$. 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 3.1: Suppose $\log p = o(T^{\gamma/6})$, $T = o(p^2)$, and Assumptions 3.1–3.4 hold. Then for a sufficiently large constant $C > 0$ in the threshold (3.2), the POET estimator $\widehat{\sum}_{u,\hat{K}}^{\mathscr{T}}$ satisfies

$$\left\| \widehat{\sum}_{u,\hat{K}}^{\mathscr{T}} - \sum_u \right\| = O_p(\omega_T^{1-q} m_p).$$

If further $\omega_T^{1-q} m_p = o(1)$, then the eigenvalues of $\widehat{\sum}_{u,\hat{K}}^{\mathscr{T}}$ are all bounded away from zero with probability approaching one, and

$$\left\| (\widehat{\sum}_{u,\hat{K}}^{\mathscr{T}})^{-1} - \sum_u^{-1} \right\| = O_p(\omega_T^{1-q} m_p).$$

When estimating Σ_u, p is allowed to grow exponentially fast in T, and $\widehat{\Sigma}_{u,\widehat{K}}^{\mathcal{T}}$ can be made

consistent under the spectral norm. In addition, $\widehat{\Sigma}_{u,\widehat{K}}^{\mathcal{T}}$ is asymptotically invertible while the classical sample covariance matrix based on the residuals is not when $p > T$.

Remark 3.2

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

2. Fan, Liao and Mincheva (2011) 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 $\left\| \widehat{\Sigma}_u^{\mathcal{T}} - \Sigma_u \right\| = O_p\left(m_p \sqrt{\frac{\log p}{T}} \right) = \left\| (\widehat{\Sigma}_u^{\mathcal{T}})^{-1} - \Sigma_u^{-1} \right\|$. Hence when the common factors are unobservable, the rate of convergence has an additional term $m_p / \sqrt{p}$, coming from the impact of estimating the unknown factors. This impact vanishes when $p \log p \gg T$, in which case the minimax rate as in Cai and Zhou (2010) is achieved. As p increases, more information about the common factors is collected, which results in more accurate estimation of the common factors $\{\mathbf{f}_t\}_{t=1}^T$.

3. When K is known and grows with p and T, with slightly weaker assumptions, our working paper (Fan et al. 2011) shows that under the exactly sparse case (that is, $q = 0$), the result continues to hold with convergence rate $m_p(K^2 \sqrt{\frac{\log p}{T}} + \frac{K^3}{\sqrt{p}})$.

3.3 Convergence of the POET estimator

Since the first K eigenvalues of Σ grow with p, one can hardly estimate Σ with satisfactory accuracy in the absolute term. This problem arises not from the limitation of any estimation method, but is due to the nature of the high-dimensional factor model. We illustrate this using a simple example.

Example 3.1: Consider an ideal case where we know the spectrum except for the first eigenvector of Σ. Let $\{\lambda_j, \xi_j\}_{j=1}^p$ be the eigenvalues and vectors, and assume that the largest eigenvalue $\lambda_1 \quad cp$ for some $c > 0$. Let $\widehat{\xi}_1$ be the estimated first eigenvector and define the covariance estimator $\widehat{\Sigma} = \lambda_1 \widehat{\xi}_1 \widehat{\xi}_1' + \sum_{j=2}^p \lambda_j \xi_j \xi_j'$. Assume that $\widehat{\xi}_1$ is a good estimator in the sense that $\|\widehat{\xi}_1 - \xi_1\|^2 = O_p(T^{-1})$. However,

$$\left\| \widehat{\Sigma} - \Sigma \right\| = \left\| \lambda_1(\widehat{\xi}_1 \widehat{\xi}_1' - \xi_1 \xi_1') \right\| = \lambda_1 O_p(\|\widehat{\xi} - \xi\|) = O_p(\lambda_1 T^{-1/2}),$$

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

In the presence of very spiked eigenvalues, while the covariance Σ cannot be consistently estimated in absolute term, it can be well estimated in terms of the *relative error* matrix

$$\sum{}^{-1/2}\widehat{\sum}\widehat{\sum}{}^{-1/2}-\mathbf{I}_p$$

which is more relevant for many applications (see Example 5.2). The relative error matrix can be measured by either its spectral norm or the normalized Frobenius norm defined by

$$p^{-1/2}\left\|\sum{}^{-1/2}\widehat{\sum}\sum{}^{-1/2}-\mathbf{I}_p\right\|_F=\left(p^{-1}tr\left[\left(\sum{}^{-1/2}\widehat{\sum}\sum{}^{-1/2}-\mathbf{I}_p\right)^2\right]\right)^{1/2}. \quad (3.3)$$

In the last equality, 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, introduced by James and Stein (1961). Also note that

$$p^{-1/2}\left\|\sum{}^{-1/2}\widehat{\sum}\sum{}^{-1/2}-\mathbf{I}_p\right\|_F=\left\|\widehat{\sum}-\sum\right\|_{\sum}$$

where $\|\mathbf{A}\|_{\Sigma}=p^{-1/2}\|\Sigma^{-1/2}\mathbf{A}\Sigma^{-1/2}\|_F$ 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 $\left\|\widehat{\sum}_{\text{sam}}-\sum\right\|_{\sum}=O_p\left(\sqrt{p/T}\right)$, which does not converge if $p > T$. On the other hand, Theorem 3.2 below shows that $\|\Sigma_{\widehat{K}}-\Sigma\|_{\Sigma}$ can still be convergent as long as $p = o(T^2)$. Technically, the impact of high-dimensionality on the convergence rate of $\Sigma_{\widehat{K}}-\Sigma$ is via the number of rows in $\mathbf{B}$. We show in the appendix that $\mathbf{B}$ appears in $\|\Sigma_{\widehat{K}}-\Sigma\|_{\Sigma}$ through $\mathbf{B}'\Sigma^{-1}\mathbf{B}$ whose eigenvalues are bounded. Therefore it successfully cancels out the curse of high-dimensionality introduced by $\mathbf{B}$.

Compared to estimating Σ, 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 Σ^{-1} has bounded eigenvalues.

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

Theorem 3.2: Under the assumptions of Theorem 3.1, the POET estimator defined in (2.15) satisfies

$$\left\|\widehat{\sum}_{\widehat{K}}-\sum\right\|_{\sum}=O_p\left(\frac{\sqrt{p}\log p}{T}+m_p\omega_T^{1-q}\right), \quad \left\|\widehat{\sum}_{\widehat{K}}-\sum\right\|_{\max}=O_p(\omega_T).$$

In addition, if $m_p\omega_T^{1-q}=o(1)$, then $\Sigma_{\widehat{K}}$ is nonsingular with probability approaching one, with

$$\left\|\widehat{\sum}_{\widehat{K}}^{-1}-\sum{}^{-1}\right\|=O_p(m_p\omega_T^{1-q}).$$

Remark 3.3

1. When estimating Σ^{-1}, p is allowed to grow exponentially fast in T, and the estimator has the same rate of convergence as that of the estimator $\widehat{\Sigma}^{\mathscr{T}}_{u,\widehat{K}}$ in Theorem 3.1. When p becomes much larger than T, the precision matrix can be estimated at the same rate as if the factors were observable.

2. As in Remark 3.2, when $K > 0$ is known and grows with p and T, the working paper Fan et al. (2011) proves the following results (when $q = 0$) [7]:

$$\left\| \widehat{\Sigma}^{\mathscr{T}} - \Sigma \right\|_{\Sigma} = O_p\left(\frac{K\sqrt{p}\log p}{T} + K^2 m_p \sqrt{\frac{\log p}{T}} + \frac{m_p K^3}{\sqrt{p}} \right),$$

$$\left\| \widehat{\Sigma}^{\mathscr{T}} - \Sigma \right\|_{\max} = O_p\left(K^3 \sqrt{\frac{\log p}{T}} + \frac{K^3}{\sqrt{p}} \right),$$

$$\left\| (\widehat{\Sigma}^{\mathscr{T}})^{-1} - \Sigma^{-1} \right\| = O_p\left(K^2 m_p \sqrt{\frac{\log p}{T}} + \frac{K^3 m_p}{\sqrt{p}} \right),$$

    The results state explicitly the dependence of the rate of convergence on the number of factors.

3. The relative error $\|\Sigma^{-1/2} \widehat{\Sigma}_K \Sigma^{-1/2} - \mathbf{I}_p\|$ in operator norm can be shown to have the same order as the maximum relative error of estimated eigenvalues. It does not converge to zero nor diverge. It is much smaller than $\|\widehat{\Sigma}_K - \Sigma\|$, which is of order $p/\sqrt{T}$ (see Example 3.1).

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 $\mathbf{H}'\mathbf{H} = \mathbf{I}_K$, $\mathbf{Bf}_t = \mathbf{BH}'\mathbf{Hf}_t$. Hence $(\mathbf{B}, \mathbf{f}_t)$ cannot be identified from $(\mathbf{BH}', \mathbf{Hf}_t)$. Note that the linear space spanned by the rows of $\mathbf{B}$ is the same as that by those of $\mathbf{BH}'$. In practice, it often does not matter which one is used.

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

Bai (2003) obtained the rate of convergence for both $\widehat{\mathbf{b}_i}$ and $\widehat{\mathbf{f}_t}$ for any fixed (i, t). However, the uniform rate of convergence is more relevant for many applications (see Example 5.1). 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 $\widehat{\mathbf{f}_t}$.

<u>Theorem 3.3:</u> Under the assumptions of Theorem 3.1,

$$\max_{i \leq p} \left|\left|\widehat{\mathbf{b}}_i - \mathbf{H}\mathbf{b}_i\right|\right| = O_p(\omega_T), \quad \max_{t \leq T} \left|\left|\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t\right|\right| = O_p\left(\frac{1}{T^{1/2}} + \frac{T^{1/4}}{\sqrt{p}}\right).$$

As a consequence of Theorem 3.3, we obtain the following: (recall that the constant r_2 is defined in Assumption 3.2.)

Corollary 3.1: Under the assumptions of Theorem 3.1,

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

The rates of convergence obtained above also explain the condition $T = o(p^2)$ in Theorems 3.1 and 3.2. It is needed in order to estimate the common factors $\{\mathbf{f}_t\}_{t=1}^T$ uniformly in $t \leq T$. 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 introduced by the unknown factors 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 other authors, achieving uniform accuracy is essential for large covariance estimations.

4 Choice of Threshold

4.1 Finite-sample positive definiteness

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

$$C_{\min} = \inf\{C > 0 : \lambda_{\min}(\widehat{\sum}_{u,\widehat{K}}^{\mathscr{T}}(M)) > 0, \quad \forall M > C\} \quad (4.1)$$

When C is sufficiently large, the estimator becomes diagonal, while its minimum eigenvalue must retain strictly positive. Thus, $C_{\min}$ is well defined and for all $C > C_{\min}$, $\widehat{\sum}_{u,\widehat{K}}^{\mathscr{T}}(C)$ is positive definite under finite sample. We can obtain $C_{\min}$ by solving $\lambda_{\min}(\widehat{\sum}_{u,\widehat{K}}^{\mathscr{T}}(C)) = 0$, C

0. We can also approximate $C_{\min}$ by plotting $\lambda_{\min}(\widehat{\sum}_{u,\widehat{K}}^{\mathscr{T}}(C))$ as a function of C, as illustrated in Figure 1. In practice, we can choose C in the range $(C_{\min} + \varepsilon, M)$ for a small ε and large enough M. Choosing the threshold in a range to guarantee the finite-sample positive definiteness has also been previously suggested by Fryzlewicz (2010).

4.2 Multifold Cross-Validation

In practice, C can be data-driven, and chosen through multifold cross-validation. After obtaining the estimated residuals $\{\hat{\mathbf{u}}_t\}_{t \leq T}$ by the 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, 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 to choose $T(J_1) = T(1 - (\log T)^{-1})$.

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

$$C^* = \arg \min_{C_{\min} + \varepsilon \leq C \leq M} \frac{1}{H}\sum_{j=1}^{H}\left\|\widehat{\sum}_{u}^{\mathscr{T},j}(C) - \widehat{\sum}_{u}^{j}\right\|_F^2. \quad (4.2)$$

Here $C_{\min}$ is the minimum constant that guarantees the positive definiteness of $\widehat{\sum}_{u,\hat{K}}^{\mathscr{T}}(C)$ for $C > C_{\min}$ as described in the previous subsection, and M is a large constant such that $\widehat{\sum}_{u,\hat{K}}^{\mathscr{T}}(M)$ is diagonal. The resulting C^* is data-driven, so depends on $\mathbf{Y}$ as well as p and T via the data. On the other hand, for each given $N \times T$ data matrix $\mathbf{Y}$, C^* is a universal constant in the threshold $\tau_{ij} = C^*\sqrt{\hat{\theta}_{ij}}\omega_T$ 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 Σ_u rather than Σ 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 $\widehat{\sum}_{u,\hat{K}}^{\mathscr{T}}(C^*)$ under the current model setting, but it ought to be much more technically involved than the regular sparse matrix estimation 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 3.1–3.3 can be applied. Detailed pursuits of these are beyond the scope of the paper.

Example 5.1 (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 hypothesis

$$H_{i0}: \mu_i = 0 \quad \text{vs.} \quad H_{i1}: \mu_i \neq 0$$

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

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

$$\mathbf{Z} = \boldsymbol{\mu} + \mathbf{Bf} + \mathbf{u}, \quad \text{where} \sum_u \text{is sparse.} \quad (5.1)$$

By the principal factor approximation (Theorem 1, Fan, Han, Gu, 2012)

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

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

Now suppose that we have n repeated measurements from the model (5.1). Then, by Corollary 3.1, $\{\eta_i\}$ can be uniformly consistently estimated, and hence $p^{-1}V(x)$ and FDP(x) can be consistently estimated. Efron (2010) obtained these repeated test statistics based on the bootstrap sample from the original raw data. Our theory (Theorem 3.3) gives a formal justification to the framework of Efron (2007, 2010).

<u>Example 5.2 (Risk management):</u> The maximum elementwise estimation error $\|\hat{\boldsymbol{\Sigma}}_K^{\hat{}} - \boldsymbol{\Sigma}\|_{\max}$ appears in risk assessment as in Fan, Zhang and Yu (2012). For a fixed portfolio allocation vector $\mathbf{w}$, the true portfolio variance and the estimated one are given by $\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}$ and $\mathbf{w}'\hat{\boldsymbol{\Sigma}}_K\hat{\mathbf{w}}$ respectively. The estimation error is bounded by

$$\left| \mathbf{w}' \widehat{\sum}_{\hat{K}} \mathbf{w} - \mathbf{w}' \sum \mathbf{w} \right| \leq \left\| \widehat{\sum}_{\hat{K}} - \sum \right\|_{\max} \|\mathbf{w}\|_1^2,$$

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 $\|\mathbf{w}\|_1 \leq c$ for some $c > 0$. In particular, $c = 1$ corresponds to a portfolio with no-short positions (all weights are nonnegative). Theorem 3.2 quantifies the maximum approximation error.

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

$$\left| \mathbf{w}' \widehat{\sum}_{\hat{K}} \mathbf{w} / \mathbf{w}' \sum \mathbf{w} - 1 \right| \leq \left\| \sum^{-1/2} \widehat{\sum}_{\hat{K}} \sum^{-1/2} - \mathbf{I}_p \right\|$$

for any allocation vector $\mathbf{w}$. Theorem 3.2 quantifies this relative error.

Example 5.3 (Panel regression with a factor structure in the errors): Consider the following panel regression model

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

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 β. The above panel regression model has been considered by many researchers, such as Ahn, Lee and Schmidt (2001), Pesaran (2006), and has broad applications in social sciences.

Although OLS (ordinary least squares) produces a consistent estimator of β, a more efficient estimation can be obtained by GLS (generalized least squares). The GLS method depends, however, on an estimator of $\sum_{\varepsilon}^{-1}$, the inverse of the covariance matrix of $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, ..., \varepsilon_{pt})'$. By assuming the covariance matrix of $(u_{1t}, ..., u_{pt})$ to be sparse, we can successfully solve this problem by applying Theorem 3.2. Although ε_{it} is unobservable, it can be replaced by the regression residuals $\widehat{\varepsilon}_{it}$, obtained via first regressing Y_{it} on $\mathbf{x}_{it}$. We then apply the POET estimator to $T^{-1}\sum_{t=1}^{T}\widehat{\varepsilon}_t\widehat{\varepsilon}_t'$. By Theorem 3.2, the inverse of the resulting estimator is a consistent estimator of $\sum_{\varepsilon}^{-1}$ under the spectral norm. A slight difference lies in the fact that when we apply POET, $T^{-1}\sum_{t=1}^{T}\varepsilon_t\varepsilon_t'$ is replaced with $T^{-1}\sum_{t=1}^{T}\widehat{\varepsilon}_t\widehat{\varepsilon}_t'$, which introduces an additional term $O_p(\sqrt{\frac{\log p}{T}})$ in the estimation error.

Example 5.4 (Validating an asset pricing theory): A celebrated financial economic theory is the capital asset pricing model (CAPM, Sharpe 1964) that makes William Sharpe win the Nobel prize in Economics in 1990, whose extension is the multi-factor 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 noises. In the multi-period model, the excess return y_{it} of firm i at time t follows model (1.1), in which $\mathbf{f}_t$ is the excess returns of the risk factors at time t. To test the null hypothesis (1.2), one embeds the model into the multivariate linear model

$$\mathbf{y}_t = \alpha + \mathbf{B}\mathbf{f}_t + \mathbf{u}_t, \quad t = 1, \cdots, T \quad (5.3)$$

and wishes to test $H_0 : \boldsymbol{\alpha} = 0$. The F-test statistic involves the estimation of the covariance matrix Σ_u, whose estimates are degenerate without regularization when $p \quad T$. Therefore, in the literature (Sentana, 2009, and references therein), one focuses on the case p is relatively small. The typical choices of parameters are $T = 60$ monthly data and the number of assets $p = 5$, 10 or 25. However, the CAPM should hold for all tradeable assets, not just a small fraction of assets. With our regularization technique, non-degenerate estimate $\widehat{\sum}_{u,\widehat{K}}^{\mathcal{T}}$ 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 $\widehat{\boldsymbol{\alpha}}$ be the least-squares estimator of (5.3). Then, when $\mathbf{u}_t \sim N(0, \Sigma_u)$, $\widehat{\boldsymbol{\alpha}} \sim N(\boldsymbol{\alpha}, \Sigma_u/c_T)$ for a constant c_T which depends on the observed factors. When Σ_u is known, the Wald test statistic is $W = c_T\widehat{\alpha}'\sum_u^{-1}\widehat{\alpha}$. When it is unknown and p is large, it is

natural to use the F-type of test statistic $\widehat{W} = c_T \hat{\alpha}'(\widehat{\sum}_{u,\widehat{K}}^{\mathscr{T}})^{-1} \hat{\alpha}$. The difference between these two statistics is bounded by

$$\left| \widehat{W} - W \right| \leq c_T \left| \left| (\widehat{\sum}_{u,\widehat{K}}^{\mathscr{T}})^{-1} - \sum_u^{-1} \right| \right| \; ||\hat{\alpha}||^2.$$

Since under the null hypothesis $\hat{a} \sim N(0, \Sigma_u/c_T)$, we have $c_T \left| \left| \sum_u^{-1/2} \hat{\alpha} \right| \right|^2 = O(p)$. Thus, it follows from boundness of $||\Sigma_u||$ that $|\widehat{W} - W| = O(p) \left| \left| (\widehat{\sum}_{u,\widehat{K}}^{\mathscr{T}})^{-1} - \sum_u^{-1} \right| \right|$. Theorem 3.1 provides the rate of convergence for the above difference. Detailed development is out of the scope of the current paper, and we will leave it as a separate research project.

6 Monte Carlo Experiments

In this section, we will examine the performance of the POET method in a finite sample. We will also demonstrate the effect of this estimator on the asset allocation and risk assessment. Similarly to Fan, et al. (2008, 2011), we simulated from a standard Fama-French three-factor model, assuming a sparse error covariance matrix and three factors. Throughout this section, the time span 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 following 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 $\mathbf{b} \sim N_3(\boldsymbol{\mu}_B, \boldsymbol{\Sigma}_B)$, the idiosyncratic errors from $\mathbf{u}_t \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}_u)$, and the factor returns $\mathbf{f}_t$ follow a 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 website of Kenneth French, and the data on 3-month Treasury bill rates from the CRSP database. These industrial portfolios are formed as the intersection of 10 portfolios based on size (market equity) and 10 portfolios based on 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.

1. 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 described in Section 3.1.

2. 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 $\mathbf{b}_i = (b_{i1}, b_{i2}, b_{i3})^T$ for $i = 1, ..., p$ are drawn from $N_3(\boldsymbol{\mu}_B, \boldsymbol{\Sigma}_B)$.

3. We run the stationary vector autoregressive model $\mathbf{f}_t = \boldsymbol{\mu} + \boldsymbol{\Phi}\mathbf{f}_{t-1} + \boldsymbol{\varepsilon}_t$, a VAR(1) model, to the data $\mathbf{F}$ to obtain the multivariate least squares estimator for $\boldsymbol{\mu}$ and $\boldsymbol{\Phi}$,

and estimate Σ_ε. Note that all eigenvalues of Φ in Table 2 fall within the unit circle, so our model is stationary. The covariance matrix $\text{cov}(\mathbf{f}_t)$ can be obtained by solving the linear equation $\text{cov}(\mathbf{f}_t) = \Phi \text{cov}(\mathbf{f}_t) \Phi' + \Sigma_\varepsilon$. The estimated parameters are depicted in Table 2 and are used to generate $\mathbf{f}_t$.

4. For each value of p, we generate a sparse covariance matrix Σ_u of the form:

$$\sum\nolimits_u = \mathbf{D} \sum\nolimits_0 \mathbf{D}.$$

Here, Σ_0 is the error correlation matrix, and $\mathbf{D}$ is the diagonal matrix of the standard deviations of the errors. We set $\mathbf{D} = \text{diag}(\sigma_1, \dots, \sigma_p)$, 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 Fan et al. (2011) has been used in this calibration step. The off-diagonal entries of Σ_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 among the estimated residuals, conditional on their absolute values being no larger than 0.95. We then employ hard thresholding to make Σ_0 sparse, where the threshold is found as the smallest constant that provides the positive definiteness of Σ_0. More precisely, start with threshold value 1, which gives $\Sigma_0 = \mathbf{I}_p$ and then decrease the threshold values in a grid until positive definiteness is violated.

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.

1. Generate independently $\{\mathbf{b}_i\}_{i=1}^p \sim N_3(\mu_B, \sum\nolimits_B)$, and set $\mathbf{B} = (\mathbf{b}_1, \dots, \mathbf{b}_p)'$.

2. Generate independently $\{\mathbf{u}_t\}_{t=1}^T \sim N_p(\mathbf{0}, \sum\nolimits_{\mathbf{u}})$.

3. Generate $\{\mathbf{f}_t\}_{t=1}^T$ as a vector autoregressive sequence of the form $\mathbf{f}_t = \mu + \Phi \mathbf{f}_{t-1} + \varepsilon_t$.

4. Calculate $\{\mathbf{y}_t\}_{t=1}^T$ from $\mathbf{y}_t = \mathbf{B}\mathbf{f}_t + \mathbf{u}_t$.

5. Set hard-thresholding with threshold $0.5 \sqrt{\hat{\theta}_{ij}} \left(\sqrt{\frac{\log p}{T}} + \frac{1}{\sqrt{p}} \right)$. Estimate K using Bai and Ng (2002)'s IC1. Calculate covariance estimators using the POET method. Calculate the sample covariance matrix $\widehat{\Sigma}_{sam}$.

In the graphs below, we plot the averages and standard deviations of the distance from $\widehat{\Sigma}_{\hat{K}}$ and $\widehat{\Sigma}_{sam}$ to the true covariance matrix Σ, under norms $\|.\|_\Sigma$, $\|.\|$ and $\|.\|_{\max}$. We also plot the means and standard deviations of the distances from $(\widehat{\Sigma}_{\hat{K}})^{-1}$ and $\widehat{\sum\nolimits_{sam}}^{-1}$ to Σ^{-1} under the spectral norm. The dimensionality p ranges from 20 to 600 in increments of 20. Due to invertibility, the spectral norm for $\widehat{\sum\nolimits_{sam}}^{-1}$ 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. Notice that we also plot the distance from $\widehat{\Sigma}_{obs}$ to Σ for comparison, where $\widehat{\Sigma}_{obs}$ is the estimated covariance matrix proposed by Fan et al. (2011), assuming the factors are observable.

6.3 Results

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

From the simulation results, reported in Figures 2–5, we observe that POET under the unobservable factor model performs just as well as the estimator in Fan et al. (2011) if the factors are known, when p is large enough. The cost of not knowing the factors is approximately of order $O_p(1/\sqrt{p})$. It can be seen in Figures 2 and 3 that this cost vanishes for p 200. To give a better insight of the impact of estimating the unknown factors for small p, a separate set of simulations is conducted for p 100. As we can see from Figures 2 (bottom panel) and 3 (middle and bottom panels), the impact decreases quickly. In addition, when estimating Σ^{-1}, it is hard to distinguish the estimators with known and unknown factors, whose performances are quite stable compared to the sample covariance matrix. Also, the maximum absolute elementwise error (Figure 4) of our estimator performs very similarly to that of the sample covariance matrix, which coincides with our asymptotic result. Figure 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*

The POET estimator depends on the estimated number of factors. Our theory uses a consistent esimator K. To assess the robustness of our procedure to $\hat{K}$ in finite sample, we calculate $\widehat{\sum}^{\mathcal{T}}_{u,K}$ for $K = 1, 2, ..., 10$. Again, the threshold is fixed to be $0.5\sqrt{\hat{\theta}_{ij}}\left(\sqrt{\frac{\log p}{T}} + \frac{1}{\sqrt{p}}\right)$.

6.4.1 Design 1—The simulation setup is the same as before where the true $K_0 = 3$. We calculate $\left| \|\widehat{\sum}^{\mathcal{T}}_{u,K} - \sum_u\|, \|(\widehat{\sum}^{\mathcal{T}}_{u,K})^{-1} - \sum_u^{-1}\|, \|\widehat{\sum}_K^{-1} - \sum^{-1}\| \right|$ and $\|\hat{\Sigma}_K - \Sigma\|_\Sigma$ for $K = 1, 2, ..., 10$. Figure 6 plots these norms as p increases but with a fixed $T = 300$. The results demonstrate a trend that is quite robust when K 3; especially, the estimation accuracy of the spectral norms for large p are close to each other. When $K = 1$ or 2, the estimators perform badly due to modeling bias. Therefore, POET is robust to over-estimated K, but not to under-estimation.

6.4.2 Design 2—We also simulated from a new data generating process for the robustness assessment. Consider a banded idiosyncratic matrix

$$\sigma_{u,ij} = \begin{cases} 0.5^{|i-j|}, & |i-j| \leq 9 \\ 0, & |i-j| > 9 \end{cases}, \quad (\mathbf{u}_1, \ldots, \mathbf{u}_T) \sim^{i.i.d.} N_p\left(0, \sum_u\right).$$

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

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

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 of all, the best estimation accuracy is achieved when $K = K_0$. Second, the estimation is robust for $K \geq K_0$. 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 under-estimated, corresponding to $K = 1, 2$, all the estimators perform badly, which demonstrates the danger of missing any common factors. Therefore, over-estimating the number of factors, while still maintaining a satisfactory estimation accuracy of the covariance matrices, is much better than under-estimating. The resulting bias caused by under-estimation is more severe than the additional variance introduced by over-estimation. Finally, estimating Σ, the covariance of $\mathbf{y}_t$, does not achieve a good accuracy even when $K = K_0$ in the absolute term $\|\hat{\Sigma} - \Sigma\|$, but the relative error $\|\Sigma^{-1/2}\hat{\Sigma}_K\Sigma^{-1/2} - \mathbf{I}_p\|$ 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, LOREC proposed by Luo (2012), the strict factor model (SFM) by Fan, Fan and Lv (2008), the Dual Method (Dual) by Lin et al. (2009), and finally, the singular value thresholding (SVT) by Cai, Candès and Shen (2008). In particular, SFM is a special case of POET which employs a large threshold that forces Σ_u to be diagonal even when the true Σ_u might not be. Note that Dual, SVT and many others dealing with low-rank plus sparse, such as Candès et al. (2011) and Wright et al. (2009), assume a known Σ and focus on recovering the decomposition. Hence they do not estimate Σ 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 $\widehat{\sum}_u^{-1}$ and $\widehat{\Sigma^{-1}}$.

Data are generated from the same setup as Design 2 in Section 6.4. Table 4 reports the averaged estimation error of the four comparing methods, calculated based on 50 replications for each simulation. Dual and SVT assume the data matrix has a low-rank plus sparse representation, which is not the case for the sample covariance matrix (though the population Σ has such a representation). The tuning parameters for POET, LOREC, Dual and SVT are chosen to achieve the best performance for each method.[8]

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 Σ itself is sparse and has bounded eigenvalues, corresponding to the case $K = 0$. Three models are considered and both POET and THR use the soft thresholding. We fix $T = 200$. Reported results are the average of 100 replications.

<u>Model 1: one-factor:</u> 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 Σ has one diverging eigenvalue.

---

[8]We used the R package for LOREC developed by Luo (2012) and the Matlab codes for Dual and SVT provided on Yi Ma's website "Low-rank matrix recovery and completion via convex optimization" at University of Illinois. The tuning parameters for each method have been chosen to minimize the sum of relative errors $\left|\left|\sum^{-1/2}\widehat{\sum}\sum^{-1/2} - \mathbf{I}_p\right|\right| + \left|\left|\sum_u^{-1/2}\widehat{\sum}_u\sum_u^{-1/2} - \mathbf{I}_p\right|\right|$. We have also written an R package for POET.

Model 2: sparse covariance: Set $K = 0$, hence $\Sigma = \Sigma_u$ itself is a banded matrix with bounded eigenvalues.

Model 3: cross-sectional AR(1): Set $K = 0$, but $\Sigma = \Sigma_u = (0.85^{|i-j|})_{p \times p}$. Now Σ is no longer sparse (or banded), but is not too dense either since Σ_{ij} decreases to zero 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), while 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. The results are summarized in Table 5.

6.6 Simulated portfolio allocation

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

Let $\hat{\Sigma}$ 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 $R(\mathbf{w}) = \mathbf{w}'\Sigma\mathbf{w}$ and $\hat{R}(\mathbf{w}) = \mathbf{w}'\hat{\Sigma}\mathbf{w}$, respectively. Now, define

$$\hat{\mathbf{w}} = \mathrm{argmin}_{\mathbf{w}'1=1} \mathbf{w}' \widehat{\sum} \mathbf{w},$$

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

For each fixed p, the population Σ 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 Σ and obtain $\hat{\mathbf{w}}$: strict factor model $\hat{\Sigma}_{\mathrm{diag}}$ (estimate Σ_u using a diagonal matrix), our POET estimator $\hat{\Sigma}_{\mathrm{POET}}$, both are with unknown factors, and sample covariance $\hat{\Sigma}_{\mathrm{Sam}}$. 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 to the oracle risk $R^* = \min_{\mathbf{w}'1=1} \mathbf{w}'\Sigma\mathbf{w}$, which is the theoretical risk of the portfolio we would have created if we knew the true covariance matrix Σ. We thus compare the regret $R(\hat{\mathbf{w}}) - R^*$, which is always nonnegative, for three estimators of $\hat{\Sigma}$. They are summarized by using the box plots over the 200 simulations. The results are reported in Figure 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 Figure 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 based on 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, where as the sample covariance does not process such a property.

7 Real Data Example

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

7.1 Sarsity of Idiosyncratic Errors

The data were obtained from the CRSP (The Center for Research in Security Prices) database, and consists 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 5 different industry sectors, (more specifically, Consumer Goods-Textile & Apparel Clothing, Financial-Credit Services, Healthcare-Hospitals, Services-Restaurants, 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 notice 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, while the rest are bounded by 0.0020. Hence $K = 0, 1, 2, 3$ are the possible values of the number of factors. Figure 9 shows the heatmap of the thresholded error correlation matrix (for simplicity, we applied hard thresholding). The threshold has been chosen using the cross validation as described in Section 4. We compare the level of sparsity (percentage of non-zero off-diagonal elements) for the 5 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 5 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 5 blocks, 100% are positive, as opposed to a distribution of 60.3% positive and 39.7% negative amongst 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 are taken out, and the thresholding has preserved them. The results for $K = 1, 2$ and 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 longer period (ten years): 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 using the penalty function proposed by Bai and Ng (2002), as defined in (2.14). For calibration, we use the last 100 consecutive business days of the above data, and both IC1 and IC2 give $K = 3$. On the 1st of each month, we estimate $\widehat{\Sigma}_{\text{diag}}$ (SFM) and $\widehat{\Sigma}_K$ (POET with soft thresholding) using the historical data of excess daily returns for the proceeding 12 months ($T = 252$). The value of the threshold is determined 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 $\hat{\mathbf{w}}_2$ (based on $\widehat{\Sigma} = \widehat{\Sigma}_{\text{diag}}$ and $\widehat{\Sigma}_K$): $\hat{\mathbf{w}} = \arg \min_{\mathbf{w}'\mathbf{1}=1} \mathbf{w}'\widehat{\Sigma}\mathbf{w}$. At the end of the month (21 trading days), their actual risks are compared, calculated by

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

We can see from Figure 10 that the minimum-risk portfolio created by the POET estimator performs significantly better, achieving lower variance 76% of the time. Amongst those months, the risk is decreased by 48.63%. On the other hand, during the months that POET produces a higher-risk portfolio, the risk is increased by only 17.66%.

Next, we demonstrate the impact 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, close to our suggested constant 0.5 used for simulation. We also present the results based on various choices of constant $C = 0.5, 0.75, 1$ and 1.25, with soft threshold $C\sqrt{\widehat{\theta}_{ij}}\omega_T$. 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 unconditional sparsity assumption is inappropriate in many applications, we introduce a latent factor model that has a conditional sparsity feature, and propose the POET estimator 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 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, 2011), 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 gets large enough, 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 the 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 2.1), which will result in one or two very spiked eigenvalues of the gene expressions' covariance matrix. The POET can be applied to estimate such a covariance matrix and its network model.

Acknowledgments

References

Agarwal A, Negahban S, Martin J, Wainwright MJ. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. Ann Statist. 2012; 40:1171–1197.

Ahn S, Lee Y, Schmidt P. GMM estimation of linear panel data models with time-varying individual effects. J Econometrics. 2001; 101:219–255.

Alessi L, Barigozzi M, Capassoc M. Improved penalization for determining the number of factors in approximate factor models. Statistics and Probability Letters. 2010; 80:1806–1813.

Amini AA, Wainwright MJ. High-dimensional analysis of semidefinite relaxations for sparse principal components. Ann Statist. 2009; 37:2877–2921.

Antoniadis A, Fan J. Regularized wavelet approximations. J Amer Statist Assoc. 2001; 96:939–967.

Athreya, K.; Lahiri, S. Measure theory and probability theorey. Springer; New York: 2006.

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

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

Bai J, Ng S. Large dimensional factor analysis. Foundations and trends in econometrics. 2008; 3:89–163.

Bai J, Shi S. Estimating high dimensional covariance matrices and its applications. Annals of Economics and Finance. 2011; 12:199–215.

Bickel P, Levina E. Covariance regularization by thresholding. Ann Statist. 2008; 36:2577–2604.

Birnbaum A, Johnstone I, Nadler B, Paul D. Minimax bounds for sparse PCA with noisy high-dimensional data. To appear in. Ann Statist. 2012

Boivin J, Ng S. Are More Data Always Better for Factor Analysis? J Econometrics. 2006; 132:169–194.

Cai J, Candès E, Shen Z. A singular value thresholding algorithm for matrix completion. SIAM J on Optimization. 2008; 20:1956–1982.

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

Cai, T.; Zhou, H. Optimal rates of convergence for sparse covariance matrix estimation. University of Pennsylvania; 2010. Manuscript

Candès E, Li X, Ma Y, Wright J. Robust principal component analysis? J ACM. 2011; 58:3.

Carvalho C, Chang J, Lucas J, Nevins J, Wang Q, West M. High-dimensional sparse factor modeling: applications in gene expression genomics. J Amer Statist Assoc. 2008; 103:1438–1456.

Chamberlain G, Rothschild M. Arbitrage, factor structure and mean-variance analyssi in large asset markets. Econometrica. 1983; 51:1305–1324.

Doz C, Giannone D, Reichlin L. A two-step estimator for large approximate dynamic factor models based on Kalman filtering. J Econometrics. 2011; 164:188–205.

d'Aspremont A, Bach F, Ghaoui ELL. Optimal solutions for sparse principal component analysis. J Mach Learn Res. 2008; 9:1269–1294.

Davis C, Kahan W. The rotation of eigenvectors by a perturbation III. SIAM J Numer Anal. 1970; 7:1–46.

Efron B. Correlation and large-scale simultaneous significance testing. J Amer Statist Assoc. 2007; 102:93–103.

Efron B. Correlated z-values and the accuracy of large-scale statistical estimates. J Amer Statist Assoc. 2010; 105:1042–1055.

Fama E, French K. The cross-section of expected stock returns. Journal of Finance. 1992; 47:427–465.

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

Fan J, Han X, Gu W. Control of the false discovery rate under arbitrary covariance dependence (with discussion). J Amer Statist Assoc. 2012; 107:1019–1048.

Fan J, Liao Y, Mincheva M. High dimensional covariance matrix estimation in approximate factor models. Ann Statist. 2011; 39:3320–3356.

Fan, J.; Liao, Y.; Mincheva, M. Large covariance estimation by thresholding principal orthogonal complements. Working paper of this article. 2011. arxiv.org/pdf/1201.0175.pdf

Fan J, Zhang J, Yu K. Vast portfolio selection with gross-exposure constraints. J Amer Statist Assoc. 2012; 107:592–606.

Forni M, Hallin M, Lippi M, Reichlin L. The generalized dynamic factor model: identification and estimation. Review of Economics and Statistics. 2000; 82:540–554.

Forni M, Hallin M, Lippi M, Reichlin L. The generalized dynamic factor model consistency and rates. J Econometrics. 2004; 119:231–255.

Forni M, Lippi M. The generalized dynamic factor model: representation theory. Econometric Theory. 2001; 17:1113–1141.

Fryzlewicz, P. Haar-Fisz wavelet method for interpretable estimation of large, sparse, time-varying volatility matrices. London School of Economics and Political Science; 2010. Manuscript

Hallin M, Liška R. Determining the number of factors in the general dynamic factor model. J Amer Statist Assoc. 2007; 102:603–617.

Hallin M, Liška R. Dynamic factors in the presence of blocks. J Econometrics. 2011; 163:29–41.

Hastie, TJ.; Tibshirani, R.; Friedman, J. The elements of Statistical Learning: Data Mining, Inference, and Prediction. 2. Springer; New York: 2009.

James, W.; Stein, C. Proc Fourth Berkeley Symp Math Statist Probab. Vol. 1. Univ. California Press; Berkeley: 1961. Estimation with quadratic loss; p. 361-379.

Johnstone IM. On the distribution of the largest eigenvalue in principal components analysis. Ann Statist. 2001; 29:295–327.

Johnstone IM, Lu AY. On consistency and sparsity for principal components analysis in high dimensions. J Amer Statist Assoc. 2009; 104:682–693.

Jung S, Marron JS. PCA consistency in high dimension, low sample size context. Ann Statist. 2009; 37:4104–4130.

Kapetanios G. A testing procedure for determining the number of factors in approximate factor models with large datasets. J Bus Econom Statist. 2010; 28:397–409.

Lam C, Fan J. Sparsistency and rates of convergence in large covariance matrix estimation. Ann Statist. 2009; 37:4254–4278.

Lawley, D.; Maxwell, A. Factor analysis as a statistical method. 2. London: Butterworths; 1971.

Leek J, Storey J. A general framework for multiple testing dependence. PNAS. 2008; 105:18718–18723. [PubMed: 19033188]

Lin, Z.; Ganesh, A.; Wright, J.; Wu, L.; Chen, M.; Ma, Y. Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix. Microsoft Research Asia; 2009. Manuscript

Luo, X. High dimensional low rank and sparse covariance matrix estimation via convex minimization. 2011. Manuscript

Ma, Z. Sparse principal components analysis and iterative thresholding. 2011. Manuscript

Meinshausen N, Bühlmann P. High dimensional graphs and variable selection with the Lasso. Ann Statist. 2006; 34:1436–1462.

Onatski A. Determining the number of factors from empirical distribution of eigenvalues. The Review of Economics and Statistics. 2010; 92:1004–1016.

Pati, D.; Bhattacharya, A.; Pillai, N.; Dunson, D. Posterior contraction in sparse Bayesian factor models for massive covariance matrices. Duke University; 2012. Manuscript

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

Pesaran MH. Estimation and inference in large heterogeneous panels with a multifactor error structure. Econometrica. 2006; 74:967–1012.

Phan, Q. On the sparsity assumption of the idiosyncratic errors covariance matrix-Support from the FTSE 100 stock returns. University of Warwick; 2012. Manuscript

Ross SA. The arbitrage theory of capital asset pricing. Journal of Economic Theory. 1976; 13:341–360.

Rothman A, Levina E, Zhu J. Generalized thresholding of large covariance matrices. J Amer Statist Assoc. 2009; 104:177–186.

Sentana E. The econometrics of mean-variance efficiency tests: a survey. Econ-metrics Jour. 2009; 12:65–101.

Shen H, Huang J. Sparse principal component analysis via regularized low rank matrix approximation. J Multivariate Analysis. 2008; 99:1015–1034.

Sharpe W. Capital asset prices: A theory of market equilibrium under conditons of risks. Journal of Finance. 1964; 19:425–442.

Stock, J.; Watson, M. Diffusion Indexes, NBER Working Paper 6702. 1998.

Stock J, Watson M. Forecasting using principal components from a large number of predictors. J Amer Statist Assoc. 2002; 97:1167–1179.

Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics. 2009; 10:515–534. [PubMed: 19377034]

Wright, J.; Peng, Y.; MY; Ganesh, A.; Rao, S. Robust principal component analysis: exact recovery of corrupted low-rank matrices by convex optimization. Microsoft Research Asia; 2009. Manuscript

Xiong H, Goulding EH, Carlson EJ, Tecott LH, McCulloch CE, Sen S. A flexible estimating equations approach for mapping function-valued traits. Genetics. 2011; 189:305–316. [PubMed: 21705752]

Yap JS, Fan J, Wu R. Nonparametric modeling of longitudinal covariance structure in functional mapping of quantitative trait loci. Biometrics. 2009; 65:1068–1077. [PubMed: 19302406]

Zhang Y, El Ghaui L. Large-scale sparse principal component analysis with application to text data. NIPS. 2011

APPENDIX

A Estimating a sparse covariance with contaminated data

We estimate Σ_u by applying the adaptive thresholding given by (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 is subject to measurement of errors. Instead, we may observe $\{\widehat{\mathbf{u}}_t\}_{t=1}^T$. For instance, in the approximate factor models, $\widehat{\mathbf{u}}_t=\mathbf{y}_t-\widehat{\mathbf{b}}_i'\widehat{\mathbf{f}}_t$.

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

$$\widehat{\sigma}_{ij}=\frac{1}{T}\sum_{t=1}^T\widehat{u}_{it}\widehat{u}_{jt}, \quad \text{and} \quad \widehat{\theta}_{ij}=\frac{1}{T}\sum_{t=1}^T(\widehat{u}_{it}\widehat{u}_{jt}-\widehat{\sigma}_{ij})^2. \quad \text{(A.1)}$$

$$\widehat{\sum}_u^{\mathscr{T}}=(s_{ij}(\widehat{\sigma}_{ij}))_{p\times p},$$

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

When $\{\widehat{\mathbf{u}}_t\}_{t=1}^T$ is close enough to $\{\mathbf{u}_t\}_{t=1}^T$, we can show that $\widehat{\sum}_u^{\mathscr{T}}$ 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 defined in Assumptions 3.2 and 3.3, let $\alpha=3r_1^{-1}+r_3^{-1}+1$.

Theorem A.1—Suppose $(\log p)^{6a} = o(T)$, and Assumptions 3.2 and 3.3 hold. In addition, suppose there is a sequence $a_T = o(1)$ so that $\max_{i \le p} \frac{1}{T} \sum_{t=1}^{T} |u_{it} - \hat{u}_{it}|^2 = O_p(a_T^2)$, and $\max_{i \le p, t \le T} |u_{it} - \hat{u}_{it}| = o_p(1)$; Then there is a constant $C > 0$ in the adaptive thresholding estimator (A.1) with

$$\omega_T = \sqrt{\frac{\log p}{T}} + a_T$$

such that

$$\left\| \widehat{\sum}_u^{\mathscr{T}} - \sum_u \right\| = O_p(\omega_T^{1-q} m_p).$$

If further $\omega_T m_p = o(1)$, then $\widehat{\sum}_u^{\mathscr{T}}$ is invertible with probability approaching one, and

$$\left\| (\widehat{\sum}_u^{\mathscr{T}})^{-1} - \sum_u^{-1} \right\| = O_p(\omega_T^{1-q} m_p).$$

Proof: By Assumptions 3.2 and 3.3, the conditions of Lemmas A.3 and A.4 of Fan, Liao and Mincheva (2011, *Ann. Statist*, **39**, 3320–3356) are satisfied. Hence for any $\varepsilon > 0$, there are positive constants M, θ_1 and θ_2 such that each of the events

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

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

$$\left|\left|\widehat{\textstyle\sum}_u^{\mathscr{T}} - \textstyle\sum_u\right|\right| \le \max_{i \le p} \sum_{j=1}^p |s_{ij}(\widehat{\sigma}_{ij}) - \sigma_{u,ij}|$$

$$= \max_{i \le p} \sum_{j=1}^p \left| s_{ij}(\widehat{\sigma}_{ij}) I_{(|\widehat{\sigma}_{ij}| > C\omega_T \sqrt{\widehat{\theta}_{ij}})} - \sigma_{u,ij} I_{(|\widehat{\sigma}_{ij}| > C\omega_T \sqrt{\widehat{\theta}_{ij}})} - \sigma_{u,ij} I_{(|\widehat{\sigma}_{ij}| \le C\omega_T \sqrt{\widehat{\theta}_{ij}})} \right|$$

$$\le \max_{i \le p} \sum_{j=1}^p |s_{ij}(\widehat{\sigma}_{ij}) - \widehat{\sigma}_{ij}| I_{(|\widehat{\sigma}_{ij}| > C\omega_T \sqrt{\widehat{\theta}_{ij}})} + \sum_{j=1}^p |\widehat{\sigma}_{ij} - \sigma_{u,ij}| I_{(|\widehat{\sigma}_{ij}| > C\omega_T \sqrt{\widehat{\theta}_{ij}})} + \sum_{j=1}^p |\sigma_{u,ij}| I_{(|\widehat{\sigma}_{ij}| \le C\omega_T \sqrt{\widehat{\theta}_{ij}})}$$

$$\le \max_{i \le p} \sum_{j=1}^p C\omega_T \sqrt{\widehat{\theta}_{ij}} I_{(|\widehat{\sigma}_{ij}| > C\omega_T \theta_2)} + M\omega_T \sum_{j=1}^p I_{(|\widehat{\sigma}_{ij}| > C\omega_T \theta_2)} + \sum_{j=1}^p |\sigma_{u,ij}| I_{(|\widehat{\sigma}_{ij}| \le C\omega_T \theta_1)}$$

$$\le (C\theta_1 + M)\omega_T \max_{i \le p} \sum_{j=1}^p I_{(|\sigma_{u,ij}| > M\omega_T)} + \max_{i \le p} \sum_{j=1}^p |\sigma_{u,ij}| I_{(|\sigma_{u,ij}| \le C\omega_T \theta_1 + M\omega_T)}$$

$$\le (C\theta_1 + M)\omega_T \max_{i \le p} \sum_{j=1}^p \frac{|\sigma_{u,ij}|^q}{M^q \omega_T^q} I_{(|\sigma_{u,ij}| > M\omega_T)} + \max_{i \le p} \sum_{j=1}^p |\sigma_{u,ij}| \frac{(C\theta_1 + M)^{1-q} \omega_T^{1-q}}{|\sigma_{u,ij}|^{1-q}} I_{(|\sigma_{u,ij}| \le (C\theta_1 + M)\omega_T)}$$

$$\le \frac{C\theta_1 + M}{M^q} \omega_T^{1-q} \max_{i \le p} \sum_{j=1}^p |\sigma_{u,ij}|^q + \max_{i \le p} \sum_{j=1}^p |\sigma_{u,ij}|^q (C\theta_1 + M)^{1-q} \omega_T^{1-q}$$

$$= m_p \omega_T^{1-q} (C\theta_1 + M)(M^{-q} + (C\theta_1 + M)^{-q}).$$

Let $M_1 = (C\theta_1 + M)(M^{-q} + (C\theta_1 + M)^{-q})$. Then with probability at least $1 - 2\varepsilon$,

$$\left|\left|\widehat{\textstyle\sum}_u^{\mathscr{T}} - \textstyle\sum_u\right|\right| \le m_p \omega_T^{1-q} M_1$$. Since ε is arbitrary, we have $\left|\left|\widehat{\textstyle\sum}_u^{\mathscr{T}} - \textstyle\sum_u\right|\right| = O_p(\omega_T^{1-q} m_p)$. If

in addition, $\omega_T m_p = o(1)$, then the minimum eigenvalue of $\widehat{\textstyle\sum}_u^{\mathscr{T}}$ is bounded away from zero with probability approaching one since $\lambda_{\min}(\Sigma_u) > c_1$. This then implies

$$\left|\left|(\widehat{\textstyle\sum}_u^{\mathscr{T}})^{-1} - \textstyle\sum_u^{-1}\right|\right| = O_p\left(\omega_T^{1-q} m_p\right).$$

B Proofs for Section 2

We first cite two useful theorems, which are needed to prove propositions 2.1 and 2.2. In Lemma B.1 below, let $\{\lambda_i\}_{i=1}^p$ be the eigenvalues of Σ in descending order and $\{\boldsymbol{\xi}_i\}_{i=1}^p$ be their associated eigenvectors. Correspondingly, let $\{\widehat{\lambda}_i\}_{i=1}^p$ be the eigenvalues of $\widehat{\Sigma}$ in descending order and $\{\widehat{\boldsymbol{\xi}}_i\}_{i=1}^p$ be their associated eigenvectors.

Lemma B.1

1. **(Weyl's Theorem)** $|\widehat{\lambda}_i - \lambda_i| \quad ||\widehat{\Sigma} - \Sigma||$.

2. (sin θ Theorem, Davis and Kahan, 1970)

$$\left|\left|\widehat{\boldsymbol{\xi}}_i - \boldsymbol{\xi}_i\right|\right| \le \frac{\sqrt{2} \left|\left|\widehat{\Sigma} - \Sigma\right|\right|}{\min(|\widehat{\lambda}_{i-1} - \lambda_i|, |\lambda_i - \widehat{\lambda}_{i+1}|)}.$$

Proof of Proposition 2.1

Proof: Since $\{\lambda_j\}_{j=1}^p$ are the eigenvalue of Σ and $\{||\tilde{\mathbf{b}}_j||^2\}_{j=1}^K$ are the first K eigenvalues of $\mathbf{B}\mathbf{B}'$ (the remaining $p - K$ eigenvalues are zero), then by the Weyl's theorem, for each $j \quad K$,

$$\left| \lambda_j - \left\| \tilde{\mathbf{b}}_j \right\|^2 \right| \le \left\| \sum - \mathbf{B}\mathbf{B}' \right\| = \left\| \sum{}_u \right\|.$$

For $j > K$, $|\lambda_j| = |\lambda_j - 0|$ $\|\Sigma_u\|$. On the other hand, 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 zero. Thus when j K, $\|\tilde{\mathbf{b}}_j\|^2/p$ are bounded away from zero for all large p.

Proof of Proposition 2.2

<u>Proof:</u> Applying the sin θ theorem yields

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

For a generic constant $c > 0$, $|\lambda_{j-1} - \|\tilde{\mathbf{b}}_j\|^2|$ $\|\|\tilde{\mathbf{b}}_{j-1}\|^2 - \|\tilde{\mathbf{b}}_j\|^2| - |\lambda_{j-1} - \|\tilde{\mathbf{b}}_{j-1}\|^2|$ cp for all large p, since $\|\|\tilde{\mathbf{b}}_{j-1}\|^2 - \|\tilde{\mathbf{b}}_j\|^2|$ cp but $|\lambda_{j-1} - \|\tilde{\mathbf{b}}_{j-1}\|^2|$ is bounded by Prosposition 2.1. On the other hand, if $j < K$, the same argument implies $\|\|\tilde{\mathbf{b}}_j\|^2 - \lambda_{j+1}|$ cp. If $j = K$, $\|\|\tilde{\mathbf{b}}_j\|^2 - \lambda_{j+1}|$ $= p\|\|\mathbf{b}_K\|^2/p - \lambda_{K+1}/p|$, where $\|\mathbf{b}_K\|^2/p$ is bounded away from zero, but $\lambda_{K+1}/p = O(p^{-1})$. Hence again, $\|\|\tilde{\mathbf{b}}_j\|^2 - \lambda_{j+1}|$ cp.

Proof of Theorem 2.1

<u>Proof:</u> The sample covariance matrix of the residuals using least squares method is given by

$\widehat{\sum}_u \frac{1}{T}(\mathbf{Y} - \hat{\mathbf{\Lambda}}\hat{\mathbf{F}}')(\mathbf{Y}' - \hat{\mathbf{F}}\hat{\mathbf{\Lambda}}') = \frac{1}{T}\mathbf{Y}\mathbf{Y}' - \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}'$. 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 $\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' = \sum_{i=1}^{K} \hat{\lambda}_i\hat{\xi}_i\hat{\xi}_i'$, then from the decompositions of the sample covariance

$$\frac{1}{T}\mathbf{Y}\mathbf{Y}' = \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' + \widehat{\sum}{}_u = \sum_{i=1}^{K} \hat{\lambda}_i\hat{\xi}_i\hat{\xi}_i' + \hat{\mathbf{R}},$$

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

We now show $\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' = \sum_{i=1}^{K} \hat{\lambda}_i\hat{\xi}_i\hat{\xi}_i'$ indeed holds. Consider again the least squares problem (2.8) but with the following alternative normalization constraints: $\frac{1}{p}\sum_{i=1}^{p} \mathbf{b}_i\mathbf{b}_i' = \mathbf{I}_K$, and

$\frac{1}{T}\sum_{t=1}^{T} \mathbf{f}_t\mathbf{f}_t'$ 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 (2.10) is $\tilde{\mathbf{\Lambda}} = (\hat{\xi}_1, \cdots, \hat{\xi}_K)$ and $\tilde{\mathbf{F}} = p^{-1}\mathbf{Y}'\tilde{\mathbf{\Lambda}}$. In addition, $T^{-1}\tilde{\mathbf{F}}'\tilde{\mathbf{F}} = \text{diag}(\hat{\lambda}_1, \cdots, \hat{\lambda}_K)$. From $\hat{\mathbf{\Lambda}}\hat{\mathbf{F}}' = \tilde{\mathbf{\Lambda}}\tilde{\mathbf{F}}'$, it follows that

$\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' = \frac{1}{T}\hat{\mathbf{\Lambda}}\hat{\mathbf{F}}'\hat{\mathbf{F}}\hat{\mathbf{\Lambda}}' = \frac{1}{T}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{F}}'\tilde{\mathbf{F}}\tilde{\mathbf{\Lambda}}' = \sum_{i=1}^{K} \hat{\lambda}_i\hat{\xi}_i\hat{\xi}_i'.$

C Proofs for Section 3

We will proceed by subsequently showing Theorems 3.3, 3.1 and 3.2.

C.1 Preliminary lemmas

The following results are to be used subsequently. The proofs of Lemmas C.1, C.2 and C.3 are found in Fan, Liao and Mincheva (2011).

Lemma C.1—Suppose $\mathbf{A}$, $\mathbf{B}$ are symmetric semi-positive 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

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

Lemma C.2—Suppose that the random variables Z_1, 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 $\forall s > 0$,

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

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

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

Lemma C.3—Under the assumptions of Theorem 3.1,

i.
$$\max_{i,j \leq K} |\tfrac{1}{T}\sum_{t=1}^{T} f_{it} f_{jt} - E f_{it} f_{jt}| = O_p(\sqrt{1/T}).$$

ii.
$$\max_{i,j \leq p} |\tfrac{1}{T}\sum_{t=1}^{T} u_{it} u_{jt} - E u_{it} u_{jt}| = O_p(\sqrt{(\log p)/T})$$

iii.
$$\max_{i \leq K, j \leq p} |\tfrac{1}{T}\sum_{t=1}^{T} f_{it} u_{jt}| = O_p(\sqrt{(\log p)/T})$$

Lemma C.4—Let $\hat{\lambda_K}$ denote the Kth largest eigenvalue of $\widehat{\sum}_{sam} = \tfrac{1}{T}\sum_{t=1}^{T} \mathbf{y}_t \mathbf{y}_t'$, then $\hat{\lambda_K} > C_1 p$ with probability approaching one for some $C_1 > 0$.

<u>Proof:</u> First of all, by Proposition 2.1, under Assumption 3.1, the *Kth* largest eigenvalue λ_K of Σ satisfies: for some $c > 0$,

$$\lambda_K \geq \left\|\tilde{\mathbf{b}}_K\right\|^2 - |\lambda_K - \left\|\tilde{\mathbf{b}}_K\right\|^2| \geq cp\left\|\sum_u\right\| \geq cp/2$$

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

$$\mathbf{D}_1 = (T^{-1}\mathbf{B}\sum_{t=1}^{T}\mathbf{f}_t\mathbf{f}_t' - \mathbf{I}_K)\mathbf{B}', \quad \mathbf{D}_2 = T^{-1}\sum_{t=1}^{T}(\mathbf{u}_t\mathbf{u}_t' - \textstyle\sum_u),$$

$$\mathbf{D}_3 = \mathbf{B}T^{-1}\sum_{t=1}^{T}\mathbf{f}_t\mathbf{u}_t', \quad \mathbf{D}_4 = \mathbf{D}_3'$$

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

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

First of all, by Lemma C.3,

$$\left\|T^{-1}\sum_{t=1}^{T}\mathbf{f}_t\mathbf{f}_t' - \mathbf{I}_K\right\| \leq K\left\|T^{-1}\sum_{t=1}^{T}\mathbf{f}_t\mathbf{f}_t' - \mathbf{I}_K\right\|_{\max} = O_p(\sqrt{1/T}),$$ which is $o_p(p)$ if $K \log p$
$= o(T)$. Consequently, by Assumption 3.1, we have

$$\left\|\mathbf{D}_1\right\| \leq O_p(K\sqrt{(\log K)/T})\left\|\mathbf{B}\mathbf{B}'\right\| = O_p(p\sqrt{1/T}).$$

We now deal with $\mathbf{D}_2$. It follows from Lemma C.3 that

$$\|\mathbf{D}_2\| \leq p\left\|T^{-1}\sum_{t=1}^{T}(\mathbf{u}_t\mathbf{u}_t' - \textstyle\sum_u)\right\|_{\max} = O_p(p\sqrt{(\log p)/T}).$$

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

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

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

Lemma C.5—Under Assumption 3.3, $\max_{t\leq T}\sum_{s=1}^{T}|E\mathbf{u}_s'\mathbf{u}_t|/p = O(1)$.

Proof: Since $\{\mathbf{u}_t\}_{t=1}^{T}$ is weakly stationary, $\max_{t\leq T}\sum_{s=1}^{T}|E\mathbf{u}_s'\mathbf{u}_t|/p \leq 2\sum_{t=1}^{\infty}|E\mathbf{u}_1'\mathbf{u}_t|/p$. In addition, $E|u_{it}|^4 < M$ for some constant M and any i, t since u_{it} has exponential tail. Hence by Davydov's inequality (Corollary 16.2.4 in Athreya and Lahiri 2006), there is a constant $C > 0$, for all $i \leq p$, $t \leq T$, $|Eu_{i1}u_{it}| \leq C\sqrt{\alpha(t)}$, where $\alpha(t)$ is the α-mixing coefficient. By Assumption 3.3, $\sum_{t=1}^{\infty}\sqrt{\alpha(t)} < \infty$. Thus uniformly in T,

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

C.2 Proof of Theorem 3.3

Our derivation below relies on a result obtained by Bai and Ng (2002), which showed that the estimated number of factors is consistent, in the sense that $\widehat{K}$ equals the true K with probability approaching one. Note that under our Assumptions 3.1–3.4, all the assumptions in Bai and Ng (2002) are satisfied. Thus immediately we have the following Lemma.

Lemma C.6 (Theorem 2 in Bai and Ng (2002))—For $\widehat{K}$ defined in (2.14),

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

Proof: See Bai and Ng (2002).

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

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

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

We first prove some preliminary results in the following Lemmas. Denote by $\widehat{\mathbf{f}}_t = (\widehat{f_{1t}}, \ldots, \widehat{f_{Kt}})'$.

Lemma C.7—For all $i \quad \widehat{K}$,

   i.
   $$\frac{1}{T}\sum_{t=1}^{T} \left(\frac{1}{T}\sum_{s=1}^{T} \widehat{f}_{is} E(\mathbf{u}_s'\mathbf{u}_t)/p \right)^2 = O_p(T^{-1}),$$

   ii.
   $$\frac{1}{T}\sum_{t=1}^{T} \left(\frac{1}{T}\sum_{s=1}^{T} \widehat{f}_{is} \zeta_{st} \right)^2 = O_p(p^{-1}),$$

   iii.
   $$\frac{1}{T}\sum_{t=1}^{T} \left(\frac{1}{T}\sum_{s=1}^{T} \widehat{f}_{is} \eta_{st} \right)^2 = O_p(p^{-1}),$$

   iv.
   $$\frac{1}{T}\sum_{t=1}^{T} \left(\frac{1}{T}\sum_{s=1}^{T} \widehat{f}_{is} \xi_{st} \right)^2 = O_p(p^{-1}).$$

Proof

   i.
   We have $\forall i$, $\sum_{s=1}^{T} \widehat{f}_{is}^2 = T$. By the Cauchy-Schwarz inequality,

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

By Lemma C.5, $\max_{t \leq T} \sum_{s=1}^{T} |E\mathbf{u}_s' \mathbf{u}_t / p| = O(1)$, which then yields the result.

ii. By the Cauchy-Schwarz inequality,

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

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

Note that $E(\sum_{s=1}^{T}\sum_{l=1}^{T}(\sum_{t=1}^{T}\zeta_{st}\zeta_{lt})^2) = T^2 E(\sum_{t=1}^{T}\zeta_{st}\zeta_{lt})^2 \leq T^4 \max_{st} E|\zeta_{st}|^4$. By Assumption 3.4, $\max_{st} E\zeta_{st}^4 = O(p^{-2})$, which implies that $\sum_{s,l}(\sum_{t=1}^{T}\zeta_{st}\zeta_{lt})^2 = O_p(T^4/p^2)$, and yields the result.

iii. By definition, $\eta_{st} = \mathbf{f}_s' \sum_{i=1}^{p} \mathbf{b}_i u_{it}/p$. We first bound $\|\sum_{i=1}^{p}\mathbf{b}_i u_{it}\|$. Assumption 3.4 implies $E\frac{1}{T}\sum_{t=1}^{T}\left\|\sum_{i=1}^{p}\mathbf{b}_i u_{it}\right\|^2 = E\left\|\sum_{i=1}^{p}\mathbf{b}_i u_{it}\right\|^2 = O(p)$. Therefore, by the Cauchy-Schwarz inequality,

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

$$\leq \frac{1}{Tp^2}\sum_{t=1}^{T}\left\|\sum_{j=1}^{p}\mathbf{b}_j u_{jt}\right\|^2\left(\frac{1}{T}\sum_{s=1}^{T}\widehat{f}_{is}^2\frac{1}{T}\sum_{s=1}^{T}\|\mathbf{f}_s\|^2\right) = O_p\left(\frac{1}{p}\right).$$

iv. Similar to part (iii), noting that ξ_{st} is a scalar, we have:

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

$$\leq O_p(1)\frac{1}{T}\sum_{s=1}^{T}\left\|\sum_{j=1}^{p}\mathbf{b}_j u_{js}\frac{1}{p}\right\|^2 \cdot \frac{1}{T}\sum_{s=1}^{T}\widehat{f}_{is}^2 \leq O_p\left(\frac{1}{p}\right),$$

where the third line follows from the Cauchy-Schwarz inequality.

Lemma C.8

i. $\max_{t \leq T}\left\|\frac{1}{Tp}\sum_{s=1}^{T}\widehat{\mathbf{f}}_s E(\mathbf{u}_s' \mathbf{u}_t)\right\| = O_p(\sqrt{1/T}),$

ii. $\max_{t \leq T}\left\|\frac{1}{T}\sum_{s=1}^{T}\widehat{\mathbf{f}}_s \zeta_{st}\right\| = O_p(T^{1/4}/\sqrt{p}),$

iii. $\max_{t \leq T}\left\|\frac{1}{T}\sum_{s=1}^{T}\widehat{\mathbf{f}}_s \eta_{st}\right\| = O_p(T^{1/4}/\sqrt{p}),$

iv. $\max_{t \leq T}\left\|\frac{1}{T}\sum_{s=1}^{T}\widehat{\mathbf{f}}_s \xi_{st}\right\| = O_p(T^{1/4}/\sqrt{p}).$

Proof

i.

By the Cauchy-Schwarz inequality and the fact that $\frac{1}{T}\sum_{t=1}^{T}\left\|\widehat{\mathbf{f}}_t\right\|^2=O_p(1)$,

$$\max_{t\leq T}\left\|\frac{1}{Tp}\sum_{s=1}^{T}\widehat{\mathbf{f}}_s E(\mathbf{u}_s'\mathbf{u}_t)\right\| \leq \max_{t\leq T}\left(\frac{1}{T}\sum_{s=1}^{T}\left\|\widehat{\mathbf{f}}_s\right\|^2 \frac{1}{T}\sum_{s=1}^{T}(E\mathbf{u}_s'\mathbf{u}_t/p)^2\right)^{1/2}$$

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

The result then follows from Assumption 3.3.

ii. By the Cauchy-Schwarz inequality,

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

It follows from Assumption 3.4 that $E(\frac{1}{T}\sum_{s=1}^{T}\zeta_{st}^2)^2 \leq \max_{s,t\leq T}E\zeta_{st}^4=O(\frac{1}{p^2})$. It then follows from the Chebyshev's inequality and Bonferroni's method that

$\frac{1}{T}\sum_{s=1}^{T}\zeta_{st}^2=O_p(\sqrt{T}/p)$.

iii.

By Assumption 3.4, $E\left\|\frac{1}{\sqrt{p}}\sum_{i=1}^{p}\mathbf{b}_i u_{it}\right\|^4 \leq K^2 M$. Chebyshev's inequality and Bonferroni's method yield $\max_{t\leq T}\left\|\sum_{i=1}^{p}\mathbf{b}_i u_{it}\right\|=O_p(T^{1/4}\sqrt{p})$ with probability one, which then implies:

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

iv. By the Cauchy-Schwarz inequality and Assumption 3.4, we have demonstrated that

$\left\|\frac{1}{T}\sum_{s=1}^{T}\sum_{i=1}^{p}\mathbf{b}_i u_{is}\frac{1}{p}\widehat{\mathbf{f}}_s\right\|=O_p(p^{-1/2})$. In addition, since $E\|K^{-2}\mathbf{f}_t\|^4 < M$, $\max_{t\ T}\|\mathbf{f}_t\| = O_p(T^{1/4})$. It follows that

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

.

Lemma C.9

i.

$\max_{i\leq K}\frac{1}{T}\sum_{t=1}^{T}(\widehat{\mathbf{f}}_t-\mathbf{Hf}_t)_i^2=O_p(1/T+1/p)$.

ii.

$\frac{1}{T}\sum_{t=1}^{T}\left\|\widehat{\mathbf{f}}_t-\mathbf{Hf}_t\right\|^2+O_p(1/T+1/p)$.

iii.

$\max_{t\leq T}\left\|\widehat{\mathbf{f}}_t-\mathbf{Hf}_t\right\|=O_p(\sqrt{1/T}+T^{1/4}/\sqrt{p})$.

Proof: We prove this lemma conditioning on the event $\widehat{K}=K$. Once this is done, due to $P(\widehat{K}$ $K)=o(1)$, it then implies the unconditional arguments.

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

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

Each of the four terms on the right hand side above are bounded in Lemma C.7, which then yields the desired result.

ii. It follows from part (i) and that

$$\frac{1}{T} \sum_{t=1}^{T} \left\| \hat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t \right\|^2 \leq K \max_{i \leq K} \frac{1}{T} \sum_{t=1}^{T} (\hat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t)_i^2.$$ Part (iii) is implied by (C.2) and Lemma C.8.

Lemma C.10

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

ii. $\mathbf{H}'\mathbf{H} = \mathbf{I}_K + O_p(1/\sqrt{T} + 1/\sqrt{p})$.

Proof: We first condition on $K \hat{=} K$. (i) Lemma C.4 implies $\|\mathbf{V}^{-1}\| = O_p(p^{-1})$. Also $\|\mathbf{F}\| = \lambda_{\max}^{1/2}(\mathbf{F}\mathbf{F}') = \lambda_{\max}^{1/2}(\sum_{t=1}^{T} \mathbf{f}_t \mathbf{f}_t') = O_p(\sqrt{T})$. In addition, $\|\hat{\mathbf{F}}\| = \sqrt{T}$. It then follows from the definition of $\mathbf{H}$ that $\|\mathbf{H}\| = O_p(1)$. Define $\widehat{\mathrm{cov}}(\mathbf{H}\mathbf{f}_t) = \frac{1}{T} \sum_{t=1}^{T} \mathbf{H}\mathbf{f}_t (\mathbf{H}\mathbf{f}_t)'$. Applying the triangular inequality gives:

$$\left\| \mathbf{H}\mathbf{H}' - \mathbf{I}_{\hat{K}} \right\|_F \leq \left\| \mathbf{H}\mathbf{H}' - \widehat{\mathrm{cov}}(\mathbf{H}\mathbf{f}_t) \right\|_F + \left\| \widehat{\mathrm{cov}}(\mathbf{H}\mathbf{f}_t) - \mathbf{I}_{\hat{K}} \right\|_F \quad (C.3)$$

By Lemma C.3, the first term in (C.3) is

$\left\| \mathbf{H}\mathbf{H}' - \widehat{\mathrm{cov}}(\mathbf{H}\mathbf{f}_t) \right\|_F \leq \|\mathbf{H}\|^2 \|\mathbf{I}_K - \widehat{\mathrm{cov}}(\mathbf{f}_t)\|_F = O_p\left(\frac{1}{T}\right)$. The second term of (C.3) can be bounded, by the Cauchy-Schwarz inequality and Lemma C.9, as follows:

$$\left\| \frac{1}{T} \sum_{t=1}^{T} \mathbf{H}\mathbf{f}_t (\mathbf{H}\mathbf{f}_t)' - \frac{1}{T} \sum_{t=1}^{T} \hat{\mathbf{f}}_t \hat{\mathbf{f}}_t' \right\|_F \leq \| \frac{1}{T} \sum_t (\mathbf{H}\mathbf{f}_t - \hat{\mathbf{f}}_t)(\mathbf{H}\mathbf{f}_t)' \|_F + \| \frac{1}{T} \sum_t \hat{\mathbf{f}}_t (\hat{\mathbf{f}}_t' - (\mathbf{H}\mathbf{f}_t)') \|_F$$
$$\leq \left(\frac{1}{T} \sum_t \left\| \mathbf{H}\mathbf{f}_t - \hat{\mathbf{f}}_t \right\|^2 \frac{1}{T} \sum_t \left\| \mathbf{H}\mathbf{f}_t \right\|^2 \right) + \left(\frac{1}{T} \sum_t \left\| \mathbf{H}\mathbf{f}_t - \hat{\mathbf{f}}_t \right\|^2 \frac{1}{T} \sum_t \left\| \hat{\mathbf{f}}_t \right\|^2 \right)^{1/2}$$
$$= O_p\left(\frac{1}{\sqrt{T}} + \frac{1}{\sqrt{p}} \right).$$

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

Proof of Theorem 3.3

<u>Proof:</u> The second part of this theorem was proved in Lemma C.9. We now derive the convergence rate of $\max_{i \ p} \|\widehat{\mathbf{b}}_i - \mathbf{H}\mathbf{b}_i\|$.

Using the facts that $\widehat{\mathbf{b}}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}\widehat{\mathbf{f}}_t$, and that $\frac{1}{T}\sum_{t=1}^{T}\widehat{\mathbf{f}}_t\widehat{\mathbf{f}}_t' = I_k$, we have

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

We bound the three terms on the right hand side respectively. It follows from Lemmas C.3 and C.10 that

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

second term, $Ey_{it}^2 = O(1)$. Therefore, $\max_i T^{-1}\sum_{t=1}^{T}y_{it}^2 = O_p(1)$. The Cauchy-Schwarz inequality and Lemma C.9 imply

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

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

Proof of Corollary 3.1—Under Assumption 3.3, it can be shown by Bonferroni's method that $\max_t \ _T\|\mathbf{f}_t\| = O_p((\log T)^{1/r_2})$. By Theorem 3.3, uniformly in i and t,

$$\left\|\widehat{\mathbf{b}}_i'\widehat{\mathbf{f}}_t - \mathbf{b}_i'\mathbf{f}_t\right\| \le \|\widehat{\mathbf{b}}_i - \mathbf{H}\mathbf{b}_i\|\|\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t\| + \|\mathbf{H}\mathbf{b}_i\|\|\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t\| + \|\widehat{\mathbf{b}}_i - \mathbf{H}\mathbf{b}_i\|\|\mathbf{H}\mathbf{f}_t\| + \|\mathbf{b}_i\|\|\mathbf{f}_t\|\|\mathbf{H}'\mathbf{H} - \mathbf{I}_K\|$$
$$= O_p\left((\log T)^{1/r_2}\sqrt{\frac{\log p}{T}} + \frac{T^{1/4}}{\sqrt{p}}\right).$$

C.3 Proof of Theorem 3.1

Lemma C.11— $\max_{i\le p}\frac{1}{T}\sum_{t=1}^{T}|u_{it} - \widehat{u}_{it}|^2 = O_p(\omega_T^2)$, and $\max_{i,t}|u_{it} - \hat{u}_{it}| = o_p(1)$.

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

$$\max_{i\le p}\frac{1}{T}\sum_{t=1}^{T}(u_{it} - \widehat{u}_{it})^2 \le 4\max_i\left\|\mathbf{b}_i'\mathbf{H}'\right\|^2\frac{1}{T}\sum_{t=1}^{T}\left\|\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t\right\|^2$$
$$+ 4\max_i\left\|\widehat{\mathbf{b}}_i' - \mathbf{b}_i'\mathbf{H}'\right\|^2\frac{1}{T}\sum_{t=1}^{T}\left\|\widehat{\mathbf{f}}_t\right\|^2 + 4\max_i\|\mathbf{b}_i\|^2\frac{1}{T}\sum_{t=1}^{T}\|\mathbf{f}_t\|^2\left\|\mathbf{H}'\mathbf{H} - \mathbf{I}_K\right\|_F^2$$

The first part of the lemma then follows from Theorem 3.3 and Lemma C.9. The second part follows from Corollary 3.1.

Proof of Theorem 3.1—The theorem follows immediately from Theorem A.1 and Lemma C.11.

C.4 Proof of Theorem 3.2

Define

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

Lemma C.12

i.
$$\left\|\mathbf{C}_T\right\|_F^2 = O_p(\omega_T^2 p), \text{ and } \left\|\mathbf{C}_T'\mathbf{C}_T\right\|_{\sum}^2 = O_p(\omega_T^4 p).$$

ii.
$$\left|\widehat{\sum}_{u,\widehat{K}}^{\mathscr{T}} - \sum_u\right|_{\sum}^2 = O_p(\omega_T^{2-2q} m_p^2).$$

iii.
$$\left\|\mathbf{B}\mathbf{H}'\mathbf{C}_T'\right\|_{\sum}^2 = O_p(\omega_T^2).$$

iv.
$$\left\|\mathbf{B}(\mathbf{H}'\mathbf{H} - \mathbf{I}_K)\mathbf{B}'\right\|_{\sum}^2 = O_p(p^{-2} + (pT)^{-1}).$$

Proof

i.
We have $\left\|\mathbf{C}_T\right\|_F^2 \le \max_{i \le p}\left\|\widehat{\mathbf{b}}_i - \mathbf{H}\mathbf{b}_i\right\|^2 p = O_p(\omega_T^2 p)$. Moreover, since all the eigenvalues of $\mathbf{\Sigma}$ are bounded away from zero, for any matrix $\mathbf{A}$,

$$\left\|\mathbf{A}\right\|_{\sum}^2 = O_p(p^{-1})\left\|\mathbf{A}\right\|_F^2. \text{ Hence } \left\|\mathbf{C}_T'\mathbf{C}_T\right\|_{\sum}^2 = O_p(p^{-1}\left\|\mathbf{C}_T\right\|_F^4) = O_p(p\omega_T^4).$$

ii. By Theorem 3.1,

$$\left|\widehat{\sum}_{u,\widehat{K}}^{\mathscr{T}} - \sum_u\right|_{\sum}^2 = O_p(p^{-1}\left|\widehat{\sum}_{u,\widehat{K}}^{\mathscr{T}} - \sum_u\right|_F^2) = O_p(\left|\widehat{\sum}_{u,\widehat{K}}^{\mathscr{T}} - \sum_u\right|^2) = O_p(\omega_T^{2-2q} m_p^2)$$

.

iii. The same argument of the proof of Theorem 2 in Fan, Fan and Lv (2008) implies

that $\|\mathbf{B}'\mathbf{\Sigma}^{-1}\mathbf{B}\| = O(1)$. Thus, $\left\|\mathbf{B}\mathbf{H}'\mathbf{C}_T'\right\|_{\sum}^2 = p^{-1}\text{tr}(\mathbf{H}'\mathbf{C}_T'\sum^{-1}\mathbf{C}_T\mathbf{H}\mathbf{B}'\sum^{-1}\mathbf{B})$ is upper bounded by

$$p^{-1}\left\|\mathbf{H}\right\|^2\left\|\mathbf{B}'\sum^{-1}\mathbf{B}\right\|\left\|\sum^{-1}\right\|\left\|\mathbf{C}_T\right\|_F^2 = O_p(p^{-1}\left\|\mathbf{C}_T\right\|_F^2) = O_p(\omega_T^2).$$

iv. Again, by $\|\mathbf{B}'\mathbf{\Sigma}^{-1}\mathbf{B}\| = O(1)$, and Lemma C.10,

$$\left\|\mathbf{B}(\mathbf{H}'\mathbf{H} - \mathbf{I}_K)\mathbf{B}'\right\|_{\sum}^2 = p^{-1}\text{tr}((\mathbf{H}'\mathbf{H} - \mathbf{I}_K)\mathbf{B}'\sum^{-1}\mathbf{B}(\mathbf{H}'\mathbf{H} - \mathbf{I}_K)\mathbf{B}'\sum^{-1}\mathbf{B})$$
$$\le p^{-1}\left\|\mathbf{H}'\mathbf{H} - \mathbf{I}_K\right\|_F^2\left\|\mathbf{B}'\sum^{-1}\mathbf{B}\right\|^2 = O_p(p^{-2} + (pT)^{-1}). \tag{C.5}$$

Proof of Theorem 3.2 (i)

Proof: By Lemma C.12,

$$\left\|\mathbf{B}(\mathbf{H}'\mathbf{H}-\mathbf{I}_K)\mathbf{B}'\right\|_{\sum}^2+\left\|\mathbf{B}\mathbf{H}'\mathbf{C}_T{}'\right\|_{\sum}^2+\left\|\mathbf{C}_T\mathbf{C}_T{}'\right\|_{\sum}^2=O_p(\omega_T^2+\tfrac{p\log^2 p}{T^2}).$$ Hence for a generic constant $C > 0$,

$$\left\|\widehat{\sum}_K-\sum\right\|_{\sum}^2 \le C\left\|\widehat{\mathbf{\Lambda}}\widehat{\mathbf{\Lambda}}'-\mathbf{B}\mathbf{B}'\right\|_{\sum}^2+C\left\|\widehat{\sum}_{u,\widehat{K}}^{\mathcal{T}}-\sum_u\right\|_{\sum}^2$$

$$\le C[\left\|\mathbf{B}(\mathbf{H}'\mathbf{H}-\mathbf{I}_K)\mathbf{B}'\right\|_{\sum}^2+\left\|\mathbf{B}\mathbf{H}'\mathbf{C}_T{}'\right\|_{\sum}^2+\left\|\mathbf{C}_T\mathbf{C}_T{}'\right\|_{\sum}^2+\left\|\widehat{\sum}_{u,\widehat{K}}^{\mathcal{T}}-\sum_u\right\|_{\sum}^2]$$

$$=O_p(\omega_T^{2-2q}m_p^2+\tfrac{p\log^2 p}{T^2}).$$

Lemma C.13— $\left\|\widehat{\mathbf{\Lambda}}'(\widehat{\sum}_{u,\widehat{K}}^{\mathcal{T}})^{-1}\widehat{\mathbf{\Lambda}}-(\mathbf{B}\mathbf{H}')'\sum_u^{-1}\mathbf{B}\mathbf{H}'\right\|=O_p(p\omega_T^{1-q}m_p)$.

Proof: $\left\|\mathbf{C}_T\right\|_F^2=O_p(\omega_T^2 p)$. Hence

$$\left\|\widehat{\mathbf{\Lambda}}'(\widehat{\sum}_{u,\widehat{K}}^{\mathcal{T}})^{-1}\widehat{\mathbf{\Lambda}}-(\mathbf{B}\mathbf{H}')'\sum_u^{-1}\mathbf{B}\mathbf{H}'\right\| \le \left\|\mathbf{C}_T'(\widehat{\sum}_{u,\widehat{K}}^{\mathcal{T}})^{-1}\mathbf{C}_T\right\|$$

$$+2\left\|\mathbf{C}_T'(\widehat{\sum}_{u,\widehat{K}}^{\mathcal{T}})^{-1}\mathbf{B}\mathbf{H}'\right\|+\left\|\mathbf{B}\mathbf{H}'((\widehat{\sum}_{u,\widehat{K}}^{\mathcal{T}})^{-1}-\sum_u^{-1})\mathbf{B}\mathbf{H}'\right\|=O_p(p\omega_T^{1-q}m_p) \qquad \text{(C.6)}$$

Lemma C.14— If $\omega_T^{1-q}m_p=o(1)$, then with probability approaching one, for some $c > 0$,

  i. $\lambda_{\min}(\mathbf{I}_K+(\mathbf{B}\mathbf{H}')'\sum_u^{-1}\mathbf{B}\mathbf{H}') \ge cp.$

  ii. $\lambda_{\min}(\mathbf{I}_K+\widehat{\mathbf{\Lambda}}'(\widehat{\sum}_{u,\widehat{K}}^{\mathcal{T}})^{-1}\widehat{\mathbf{\Lambda}}) \ge cp.$

  iii. $\lambda_{\min}(\mathbf{I}_K+\mathbf{B}'\sum_u^{-1}\mathbf{B}) \ge cp.$

  iv. $\lambda_{\min}((\mathbf{H}\mathbf{H}')^{-1}+\mathbf{B}'\sum_u^{-1}\mathbf{B}) \ge cp.$

Proof: (i) By Lemma C.10, with probability approaching one, $\lambda_{\min}(\mathbf{H}\mathbf{H}')$ is bounded away from zero. Hence,

$$\lambda_{\min}(\mathbf{I}_K+(\mathbf{B}\mathbf{H}')'\sum_u^{-1}\mathbf{B}\mathbf{H}') \ge \lambda_{\min}((\mathbf{B}\mathbf{H}')'\sum_u^{-1}\mathbf{B}\mathbf{H}'))$$

$$\ge \lambda_{\min}(\sum_u^{-1})\lambda_{\min}(\mathbf{H}\mathbf{B}'\mathbf{B}\mathbf{H}') \ge \lambda_{\min}(\sum_u^{-1})\lambda_{\min}(\mathbf{B}'\mathbf{B})\lambda_{\min}(\mathbf{H}\mathbf{H}') \ge cp.$$

(ii) The result follows from part (i) and Lemma C.13. Part (iii) and (iv) follow from a similar argument of part (i) and Lemma C.10.

Proof of Theorem 3.2

Proof: We derive the rate for $\left|\left|\widehat{\sum}_{\widehat{K}}^{-1} - \sum^{-1}\right|\right|$. Define

$$\tilde{\sum} = \mathbf{BH}'\mathbf{HB}' + \sum_u.$$

Note that $\widehat{\sum}_{\widehat{K}} = \widehat{\mathbf{\Lambda}}\widehat{\mathbf{\Lambda}}' + \widehat{\sum}_{u,\widehat{K}}^{\mathscr{T}}$ and $\Sigma = \mathbf{BB}' + \Sigma_u$. The triangular inequality gives

$$\left|\left|\widehat{\sum}_{\widehat{K}}^{-1} - \sum^{-1}\right|\right| \le \left|\left|\widehat{\sum}_{\widehat{K}}^{-1} - \tilde{\sum}^{-1}\right|\right| + \left|\left|\tilde{\sum}^{-1} - \sum^{-1}\right|\right|.$$

Using the Sherman-Morrison-Woodbury formula, we have $\left|\left|\widehat{\sum}_{\widehat{K}}^{-1} - \tilde{\sum}^{-1}\right|\right| \le \sum_{i=1}^{6} L_i$, where

$$L_1 = \left|\left|(\widehat{\sum}_{u,\widehat{K}}^{\mathscr{T}})^{-1} - \sum_u^{-1}\right|\right|$$

$$L_2 = \left|\left|((\widehat{\sum}_{u,\widehat{K}}^{\mathscr{T}})^{-1} - \sum_u^{-1})\widehat{\mathbf{\Lambda}}[\mathbf{I}_K + \widehat{\mathbf{\Lambda}}'(\widehat{\sum}_{u,\widehat{K}}^{\mathscr{T}})^{-1}\widehat{\mathbf{\Lambda}}]^{-1}\widehat{\mathbf{\Lambda}}'(\widehat{\sum}_{u,\widehat{K}}^{\mathscr{T}})^{-1}\right|\right|$$

$$L_3 = \left|\left|((\widehat{\sum}_{u,\widehat{K}}^{\mathscr{T}})^{-1} - \sum_u^{-1})\widehat{\mathbf{\Lambda}}[\mathbf{I}_K + \widehat{\mathbf{\Lambda}}'(\widehat{\sum}_{u,\widehat{K}}^{\mathscr{T}})^{-1}\widehat{\mathbf{\Lambda}}]^{-1}\widehat{\mathbf{\Lambda}}'\sum_u^{-1}\right|\right|$$

$$L_4 = \left|\left|\sum_u^{-1}(\widehat{\mathbf{\Lambda}} - \mathbf{BH}')[\mathbf{I}_K + \widehat{\mathbf{\Lambda}}'(\widehat{\sum}_{u,\widehat{K}}^{\mathscr{T}})^{-1}\widehat{\mathbf{\Lambda}}]^{-1}\widehat{\mathbf{\Lambda}}'\sum_u^{-1}\right|\right| \quad\text{(C.7)}$$

$$L_5 = \left|\left|\sum_u^{-1}(\widehat{\mathbf{\Lambda}} - \mathbf{BH}')[\mathbf{I}_K + \widehat{\mathbf{\Lambda}}'(\widehat{\sum}_{u,\widehat{K}}^{\mathscr{T}})^{-1}\widehat{\mathbf{\Lambda}}]^{-1}\mathbf{HB}'\sum_u^{-1}\right|\right|$$

$$L_6 = \left|\left|\sum_u^{-1}\mathbf{BH}'([\mathbf{I}_K + \widehat{\mathbf{\Lambda}}'(\widehat{\sum}_{u,\widehat{K}}^{\mathscr{T}})^{-1}\widehat{\mathbf{\Lambda}}]^{-1} - [\mathbf{I}_K + \mathbf{HB}'\sum_u^{-1}\mathbf{BH}']^{-1})\mathbf{HB}'\sum_u^{-1}\right|\right|.$$

We bound each of the six terms respectively. First of all, L_1 is bounded by Theorem 3.1. Let $\mathbf{G} = [\mathbf{I}_K + \widehat{\mathbf{\Lambda}}'(\widehat{\sum}_{u,\widehat{K}}^{\mathscr{T}})^{-1}\widehat{\mathbf{\Lambda}}]^{-1}$, then

$$L_2 \le \left|\left|(\widehat{\sum}_{u,\widehat{K}}^{\mathscr{T}})^{-1} - \sum_u^{-1}\right|\right| \cdot \left|\left|\widehat{\mathbf{\Lambda}}\mathbf{G}\widehat{\mathbf{\Lambda}}'\right|\right| \cdot \left|\left|(\widehat{\sum}_{u,\widehat{K}}^{\mathscr{T}})^{-1}\right|\right|.$$

Note that Theorem 3.1 implies $\left|\left|(\widehat{\sum}_{u,\widehat{K}}^{\mathscr{T}})^{-1}\right|\right| = O_p(1)$. Lemma C.14 then implies $\|\mathbf{G}\| = O_p(p^{-1})$. This shows that $L_2 = O_p(L_1)$. Similarly $L_3 = O_p(L_1)$. In addition, since $||\mathbf{C}_T||_F^2 = \left|\left|\widehat{\mathbf{\Lambda}} - \mathbf{BH}'\right|\right|_F^2 = O_p(\omega_T^2 p)$, $L_4 \le \left|\left|\sum_u^{-1}(\widehat{\mathbf{\Lambda}} - \mathbf{BH}')\right|\right| \, ||\mathbf{G}|| \, \left|\left|\widehat{\mathbf{\Lambda}}\sum_u^{-1}\right|\right| = O_p(\omega_T)$.

Similarly $L_5 = O_p(L_4)$. Finally, let $\mathbf{G}_1 = [\mathbf{I}_K + (\mathbf{BH}')' \sum_u^{-1} \mathbf{BH}']^{-1}$. By Lemma C.14, $\|\mathbf{G}_1\| = O_p(p^{-1})$. Then by Lemma C.13,

$$\left\|\mathbf{G} - \mathbf{G}_1\right\| = \left\|\mathbf{G}(\mathbf{G}^{-1} - \mathbf{G}_1^{-1})\mathbf{G}_1\right\| \leq O_p(p^{-2}) \left\|(\mathbf{BH}')' \textstyle\sum_u^{-1} \mathbf{BH}' - \hat{\mathbf{\Lambda}}' (\widehat{\textstyle\sum}_{u,\hat{K}}^{\mathcal{T}})^{-1} \hat{\mathbf{\Lambda}}\right\|$$
$$= O_p(p^{-1}\omega_T^{1-q}m_p).$$

Consequently, $L_6 \leq \left\|\sum_u^{-1}\mathbf{BH}'\right\|^2 \|\mathbf{G} - \mathbf{G}_1\| = O_p(\omega_T^{1-q}m_p)$. Adding up L_1–L_6 gives

$$\left\|\widehat{\textstyle\sum}_{\hat{K}}^{-1} - \tilde{\textstyle\sum}^{-1}\right\| = O_p(\omega_T^{1-q}m_p).$$

One the other hand, using Sherman-Morrison-Woodbury formula again implies

$$\left\|\tilde{\Sigma}^{-1} - \Sigma^{-1}\right\| \leq \left\|\textstyle\sum_u^{-1}\mathbf{B}([(\mathbf{H}'\mathbf{H})^{-1} + \mathbf{B}'\textstyle\sum_u^{-1}\mathbf{B}]^{-1} - [\mathbf{I}_K + \mathbf{B}'\textstyle\sum_u^{-1}\mathbf{B}]^{-1})\mathbf{B}'\textstyle\sum_u^{-1}\right\|$$
$$\leq O(p) \left\|[(\mathbf{H}'\mathbf{H})^{-1} + \mathbf{B}'\textstyle\sum_u^{-1}\mathbf{B}]^{-1} - [\mathbf{I}_K + \mathbf{B}'\textstyle\sum_u^{-1}\mathbf{B}]^{-1}\right\|$$
$$= O_p(p^{-1}) \left\|(\mathbf{H}'\mathbf{H})^{-1} - \mathbf{I}_K\right\| = o_p(\omega_T^{1-q}m_p).$$

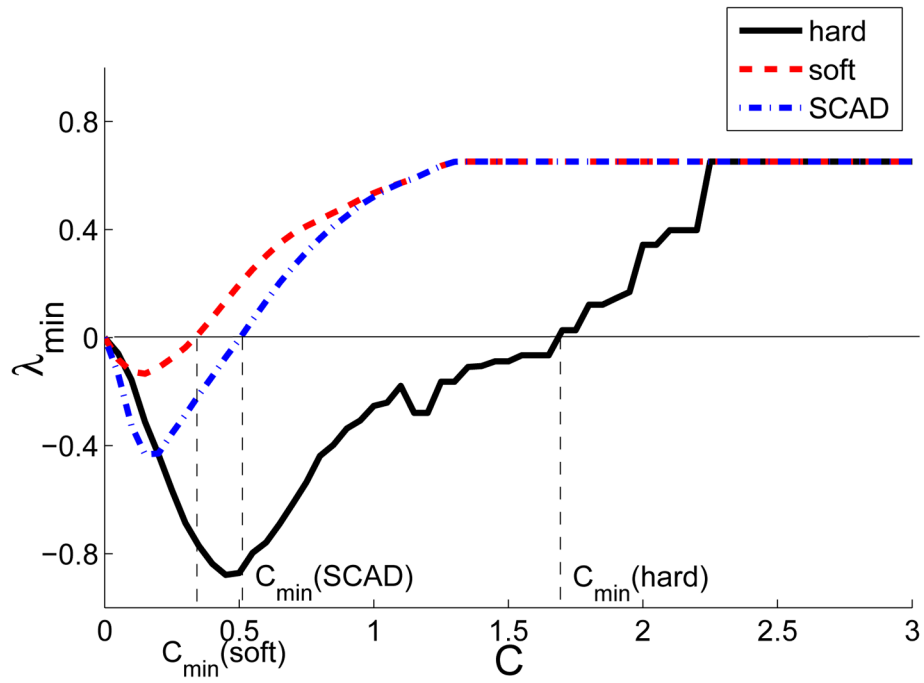**Proof of Theorem 3.2—$\|\Sigma^{\hat{\mathcal{T}}} - \Sigma\|_{\max}$**

<u>**Proof:**</u> We first bound $\|\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' - \mathbf{BB}'\|_{\max}$. Repeatedly using the triangular inequality yields

$$\left\|\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' - \mathbf{BB}'\right\|_{\max} = \max_{i,j \leq p} |\hat{\mathbf{b}}_i'\hat{\mathbf{b}}_j - \mathbf{b}_i'\mathbf{b}_j|$$
$$\leq \max_{ij} [|(\hat{\mathbf{b}}_i - \mathbf{Hb}_i)'\hat{\mathbf{b}}_j| + |\mathbf{b}_i'\mathbf{H}'(\hat{\mathbf{b}}_j - \mathbf{Hb}_j)| + |\mathbf{b}_i'(\mathbf{H}'\mathbf{H} - \mathbf{I}_K)\mathbf{b}_j|]$$
$$\leq (\max_i \|\hat{\mathbf{b}}_i - \mathbf{Hb}_i\|)^2 + 2\max_{ij} \|\hat{\mathbf{b}}_i - \mathbf{Hb}_i\| \|\mathbf{Hb}_j\| + \max_i \|\mathbf{b}_i\|^2 \|\mathbf{H}'\mathbf{H} - \mathbf{I}_K\|$$
$$= O_p(\omega_T).$$

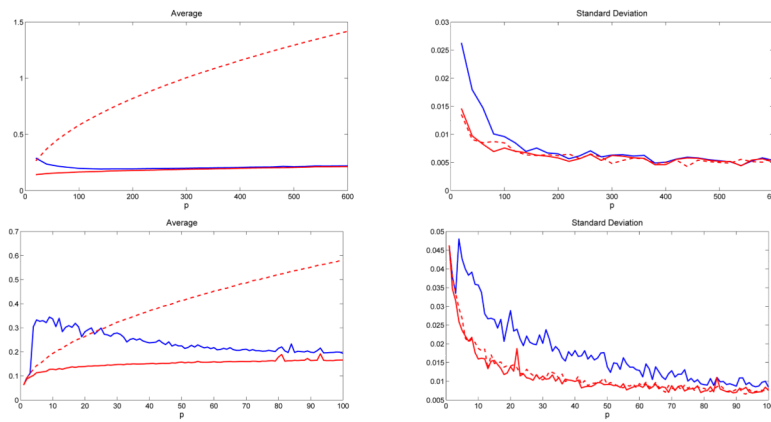On the other hand, let $\sigma_{u,ij}$ be the (i, j) entry of Σ_u. Then $\max_{ij} |\hat{\sigma_{ij}} - \sigma_{u,ij}| = O_p(\omega_T)$.

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

Hence $\left\|\widehat{\textstyle\sum}_{u,\hat{K}}^{\mathcal{T}} - \textstyle\sum_u\right\|_{\max} = O_p(\omega_T)$. The result then follows immediately.
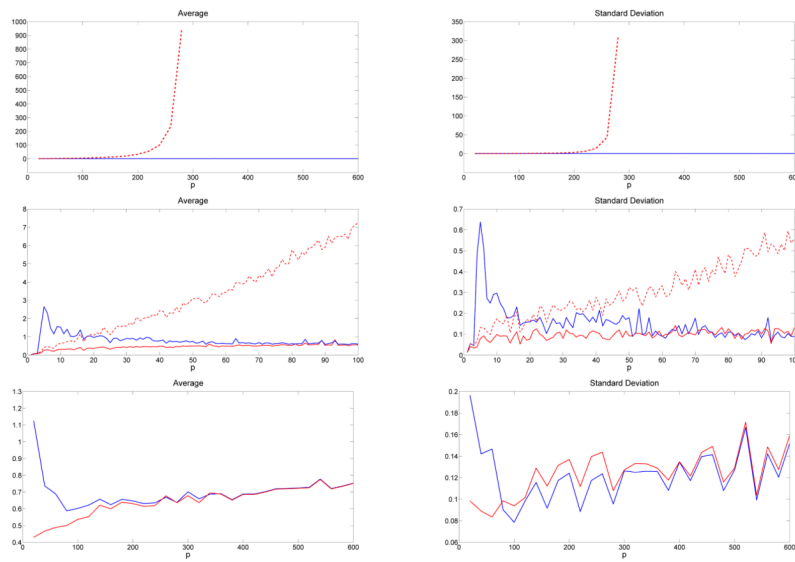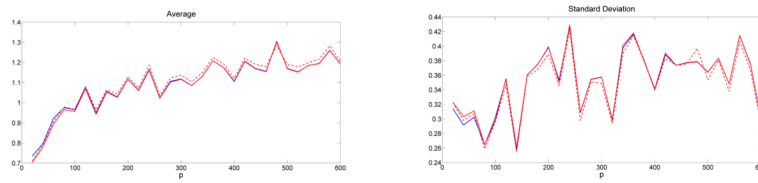
Figure 1.

Minimum eigenvalue of $\widehat{\sum}_{u,\widehat{K}}^{\mathscr{T}}(C)$ as a function of C for three choices of thresholding rules. The plot is based on the simulated data set in Section 6.2.
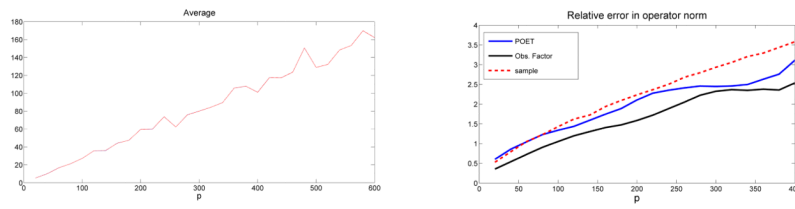
Figure 2.
Averages (left panel) and standard deviations (right panel) of the relative error $p^{-1/2}\|$ $\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - \mathbf{I}_p\|_F$ with known factors ($\hat{\Sigma} = \hat{\Sigma}_{obs}$ solid red curve), POET ($\hat{\Sigma} = \hat{\Sigma}_K$ solid blue curve), and sample covariance ($\hat{\Sigma} = \hat{\Sigma}_{sam}$ dashed curve) over 200 simulations, as a function of the dimensionality p. Top panel: p ranges in 20 to 600 with increment 20; bottom panel: p ranges in 1 to 100 with increment 1.

Figure 3.
Averages (left panel) and standard deviations (right panel) of $\|\widehat{\Sigma^{-1}} - \Sigma^{-1}\|$ with known factors ($\widehat{\Sigma} = \Sigma_{obs}$ solid red curve), POET ($\widehat{\Sigma} = \widehat{\Sigma}_{K}$ solid blue curve), and sample covariance ($\widehat{\Sigma} = \widehat{\Sigma}_{sam}$ dashed curve) over 200 simulations, as a function of the dimensionality p. Top panel: p ranges in 20 to 600 with increment 20; middle panel: p ranges in 1 to 100 with increment 1; Bottom panel: the same as the top panel with dashed curve excluded.
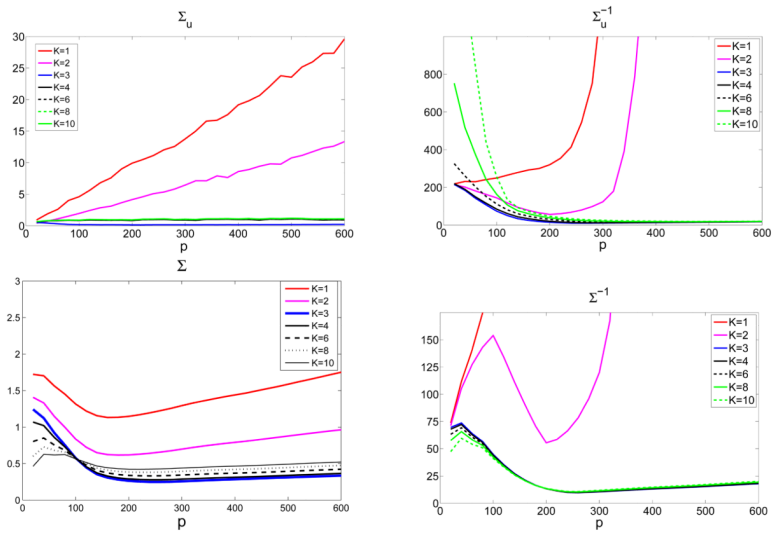
Figure 4.
Averages (left panel) and standard deviations (right panel) of $\|\hat{\Sigma} - \Sigma\|_{\max}$ with known factors ($\hat{\Sigma} = \Sigma_{obs}$ solid red curve), POET ($\hat{\Sigma} = \hat{\Sigma}_K$ solid blue curve), and sample covariance ($\hat{\Sigma} = \hat{\Sigma}_{sam}$ dashed curve) over 200 simulations, as a function of the dimensionality p. They are nearly indifferentiable.

Figure 5.
Averages of $\|\hat{\Sigma} - \Sigma\|$ (left panel) and $\|\Sigma^{-1/2}\Sigma\hat{\Sigma}^{-1/2} - \mathbf{I}_p\|$ with known factors ($\hat{\Sigma} = \hat{\Sigma}_{obs}$ solid red curve), POET ($\hat{\Sigma} = \hat{\Sigma}_K^{\hat{\omega}}$ solid blue curve), and sample covariance ($\hat{\Sigma} = \hat{\Sigma}_{sam}$ dashed curve) over 200 simulations, as a function of the dimensionality p. The three curves are hardly distinguishable on the left panel.
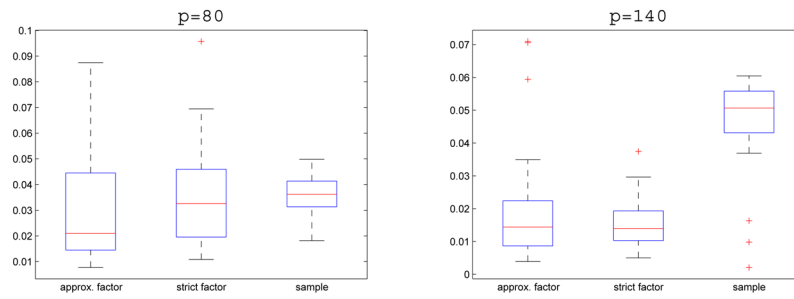
Figure 6.
Robustness of K as p increases for various choices of K (Design 1, $T = 300$). Top left: $\| \widehat{\sum}_{u,K}^{\mathscr{T}} - \sum_u \|$; top right: $\| (\widehat{\sum}_{u,K}^{\mathscr{T}})^{-1} - \sum_u^{-1} \|$; bottom left: $\|\hat{\Sigma}_K - \Sigma\|_{\Sigma}$; bottom right: $\| \widehat{\sum}_K^{-1} - \sum^{-1} \|$.

Figure 7.
Box plots of regrets $R(\hat{\mathbf{w}}) - R^*$ for $p = 80$ and 140. In each panel, the box plots from left to right correspond to $\hat{\mathbf{w}}$ obtained using $\hat{\Sigma}$ based on approximate factor model, strict factor model, and sample covariance, respectively.
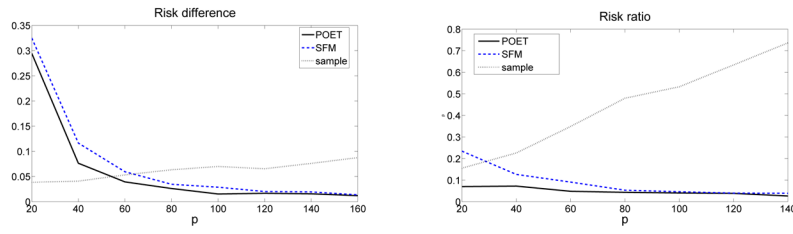
Figure 8.
Estimation errors for risk assessments as a function of the portfolio size p. Left panel plots the average absolute error $|R(\hat{\mathbf{w}}) - \hat{R}(\hat{\mathbf{w}})|$ and right panel depicts the average relative error $|\hat{R}(\hat{\mathbf{w}})/R(\hat{\mathbf{w}}) - 1|$. Here, $\hat{\mathbf{w}}$ and R are obtained based on three estimators of Σ.
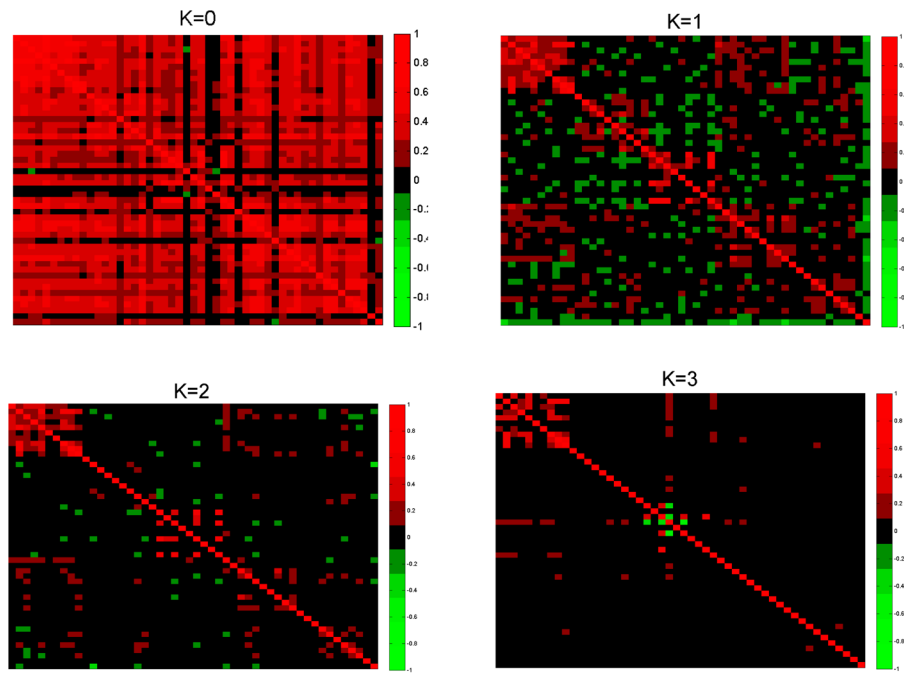
Figure 9.
Heatmap of thresholded error correlation matrix for number of factors $K = 0$, $K = 1$, $K = 2$ and $K = 3$.
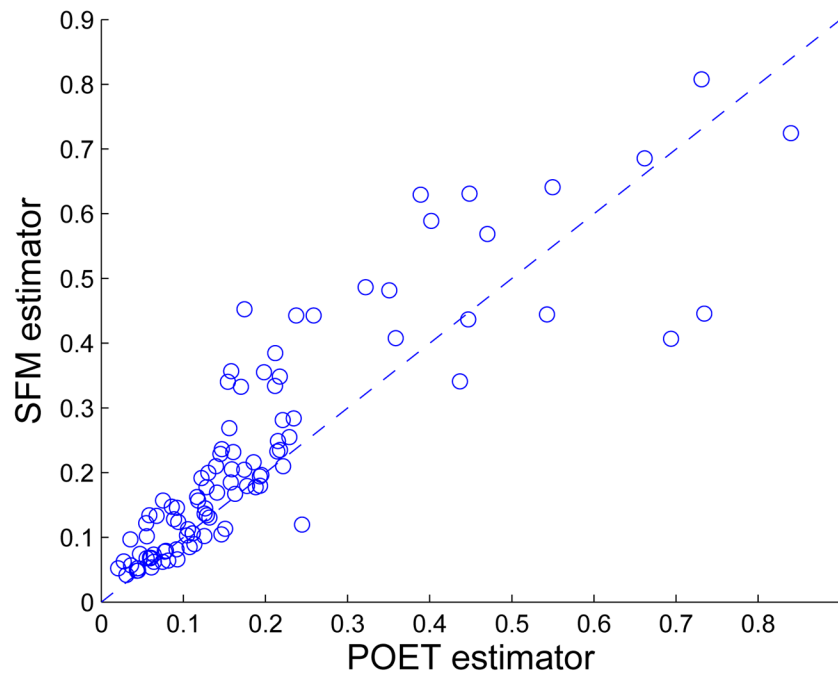
Figure 10.
Risk of portfolios created with POET and SFM (strict factor model)

Table 1

Mean and covariance matrix used to generate **b**

μ_B	Σ_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 $\mathbf{f}_t$ generating process

$\boldsymbol{\mu}$	$\text{cov}(\mathbf{f}_t)$			$\boldsymbol{\Phi}$		
−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

Table 3

Robustness of K. Design 2, estimation errors in spectral norm

				K			
	1	**2**	**3**	**4**	**5**	**6**	**8**
$p = 100$							
$\widehat{\Sigma}_{u,K}$	10.70	5.23	1.63	1.80	1.91	2.04	2.22
$\left(\widehat{\Sigma}_{u,K}\right)^{-1}$	2.71	2.51	1.51	1.50	1.44	1.84	2.82
$\widehat{\Sigma}_K^{-1}$	2.69	2.48	1.47	1.49	1.41	1.56	2.35
$\widehat{\Sigma}_K$	94.66	91.36	29.41	31.45	30.91	33.59	33.48
$\Sigma^{-1/2}\widehat{\Sigma}_K\Sigma^{-1/2}$	17.37	10.04	2.05	2.83	2.94	2.95	2.93
$p = 200$							
$\widehat{\Sigma}_{u,K}$	11.34	11.45	1.64	1.71	1.79	1.87	2.01
$\left(\widehat{\Sigma}_{u,K}\right)^{-1}$	2.69	3.91	1.57	1.56	1.81	2.26	3.42
$\widehat{\Sigma}_K^{-1}$	2.67	3.72	1.57	1.55	1.70	2.13	3.19
$\widehat{\Sigma}_K$	200.82	195.64	57.44	63.09	64.53	60.24	56.20
$\Sigma^{-1/2}\widehat{\Sigma}_K\Sigma^{-1/2}$	20.86	14.22	3.29	4.52	4.72	4.69	4.76
$p = 300$							
$\widehat{\Sigma}_{u,K}$	12.74	15.20	1.66	1.71	1.78	1.84	1.95
$\left(\widehat{\Sigma}_{u,K}\right)^{-1}$	7.58	7.80	1.74	2.18	2.58	3.54	5.45

	K						
	1	2	3	4	5	6	8
$\widetilde{\sum}_K^{-1}$	7.59	7.49	1.70	2.13	2.49	3.37	5.13
$\hat{\Sigma}_K$	302.16	274.12	87.92	92.47	91.90	83.21	92.50
$\Sigma^{-1/2}\hat{\Sigma}_K\Sigma^{-1/2}$	23.43	16.89	4.38	6.04	6.16	6.14	6.20

Table 4

Method Comparison under spectral norm for $T = 100$. RelE represents the relative error $\|\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2} - \mathbf{I}_p\|$

		$\widehat{\Sigma_u}$	$\widehat{\sum}_u^{-1}$	RelE	$\widehat{\Sigma}^{-1}$	$\widehat{\Sigma}$
$p = 100$	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

Table 5

Method Comparison. $T = 200$

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

The reported numbers are the averages based on 100 replications.

Table 6

Comparisons of the risks of portfolios using POET and SFM: The first number is proportion of the time POET outperforms and the second number is percentage of average risk improvements. *C* represents the constant in the threshold.

C	$\hat{K} = 1$	$\hat{K} = 2$	$\hat{K} = 3$
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%