Large Covariance Estimation by Thresholding Principal Orthogonal Complements

Jianqing Fan Yuan Liao Martina Mincheva*

Department of Operations Research and Financial Engineering, Princeton University

Abstract

This paper deals with estimation of high-dimensional covariance with a conditional sparsity structure, which is the composition of a low-rank matrix plus a sparse matrix. By assuming sparse error covariance matrix in a multi-factor model, we allow the presence of the 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. 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, including the spectral norm. 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.

Keywords: High dimensionality, approximate factor model, unknown factors, principal components, sparse matrix, low-rank matrix, thresholding, cross-sectional correlation.

*Address: Department of ORFE, Sherrerd Hall, Princeton University, Princeton, NJ 08544, USA, e-mail: *jqfan@princeton.edu*, *yuanliao@princeton.edu*, *mincheva@princeton.edu*. The research was partially supported by NIH R01GM100474-01, NIH R01-GM072611, and DMS-0704337.

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 covariance matrix or its inverse (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 4 for some of those applications. Yet, the dimensionality is often either comparable to the sample size or even larger. In such a case, the sample covariance is known to have a poor performance due to its diverse spectra, and some regularization is needed.

Realizing the importance of estimating large covariance matrix and the challenges brought by the high dimensionality, researchers have proposed various regularization techniques to consistently estimate $\mathbf{\Sigma}$ in recent years. One of the key assumptions is that the covariance matrix is sparse, namely, many entries are zero (Bickel and Levia, 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 $\mathbf{\Sigma}$ is not appropriate. For example, the 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 structure of factor model, 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 sparse. This approximate factor model is frequently used in economics and financial studies (Chamberlain and Rothschild, 1983; Fama and French 1993; Bai and Ng, 2002). A factor model typically takes the following form:

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

where y_{it} is the observed datum for the ith ($i = 1, ..., p$) variable 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 idiosyncratic error component, uncorrelated with $\mathbf{f}_t$. In a data-rich environment, p can diverge at a rate

faster than T. The factor model (1.1) can be put in the matrix form as

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

where $\mathbf{y}_t = (y_{1t}, ..., y_{pt})'$, $\mathbf{B} = (\mathbf{b}_1, ..., \mathbf{b}_p)'$ and $\mathbf{u}_t = (u_{1t}, ..., u_{pt})'$. Under model (1.1), the covariance matrix $\mathbf{\Sigma}$ is given by

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

where $\mathbf{\Sigma}_u$ is the covariance matrix of $\mathbf{u}_t$. We assume that $\mathbf{\Sigma}_u$ is sparse (instead of diagnonal) and refer to the model as the approximate factor model.

The conditional sparsity of form (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 the 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 very simple. Run the singular value decomposition on the sample covariance matrix $\widehat{\mathbf{\Sigma}}$, 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. See Section 2 for additional details. We will investigate various properties of POET under the assumptions that the data are serial dependence, which include independent observation as a specific example.

High-dimensional approximate factor model (1.2) is innately related to the principal component analysis, as clearly elucidated in Section 2. This is a distinguished feature not shared by the finite-dimensionality. Bai (2003) established the large sample properties of the principal components as the estimators of the factor loading as well as the estimated common factors when they are unobservable. Forni, Hallin, Lippi and Reichlin (2000) established consistency for the estimated common components $\mathbf{b}_i'\mathbf{f}_t$. In addition, Doz, Giannone and Reichlin (2006) proposed quasi-maximum likelihood estimations and investigated the effect of misspecification on the estimation of common factors, and Wang (2010) studied the large sample theory of high dimensional factor models with a multi-level factor structure. Stock and Watson (2002) considered time-varying factor loadings. See Bai and Shi (2011) for a recent review.

There has been an extensive literature in recent years that deals with sparse principal components, which have been widely used to enhance the convergence of the principal components in high-dimensional space. d'Aspremont, Bach and El Ghaoui (2008) proposed a semidefinite relaxation to this problem and derive a greedy algorithm that computes a full set of good solutions for all target numbers of non-vanishing coefficients. Shen and Huang

(2008) proposed the sPCA-rSVD algorithm and Witten, Tibshirani, and Hastie (2009) used the sPC algorithm for computing regularized principal component. The idea is further extended by Ma (2011) who iteratively applied thresholding and the QR decomposition to find sparse principal components and derived the rates of convergence of sparse principal components. Johnstone and Lu (2009) screened the variables first by a marginal variance and then applied the PCA to the screened variables to obtain a sparse principal component. They proved the consistency of such a method. Amini and Wainwright (2009) analyzed semidefinite relaxations for sparse principal components. Zhang and El Ghaoui (2011) proposed a fast block coordinate ascent algorithm for computing sparse PCA.

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. Specific applications of regularized covariance matrix are given in Section 4. Numerical results are reported in Section 5. Finally, Section 6 presents a real data example. 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 $\mathbf{A}$ respectively, defined respectively as $\|\mathbf{A}\|_F = \mathrm{tr}^{1/2}(\mathbf{A}'\mathbf{A})$, $\|\mathbf{A}\| = \lambda_{\max}^{1/2}(\mathbf{A}'\mathbf{A})$, $\|\mathbf{A}\|_1 = \max_j \sum_i |a_{ij}|$ and $\|\mathbf{A}\|_{\max} = \max_{i,j} |a_{ij}|$. Note that when $\mathbf{A}$ is a vector, $\|\mathbf{A}\|$ is equal to the Euclidean norm.

2 Regularized Covariance Matrix via PCA

2.1 POET

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

$$\widehat{\mathbf{\Sigma}} \;=\; \sum_{i=1}^{K} \widehat{\lambda}_i \widehat{\boldsymbol{\xi}}_i \widehat{\boldsymbol{\xi}}_i' + \widehat{\mathbf{R}}, \tag{2.1}$$

where $\widehat{\mathbf{R}} = \sum_{i=K+1}^{p} \widehat{\lambda}_i \widehat{\boldsymbol{\xi}}_i \widehat{\boldsymbol{\xi}}_i' = (\hat{r}_{ij})_{p\times p}$ is the principal orthogonal complement, and K is the number of principal components.

Now we apply thresholding on $\widehat{\mathbf{R}}$ (Bickel and Levina, 2008). Define

$$\widehat{\mathbf{R}}^{\mathcal{T}} = (\hat{r}_{ij}^{\mathcal{T}})_{p\times p}, \quad \hat{r}_{ij}^{\mathcal{T}} = \hat{r}_{ij} I(|\hat{r}_{ij}| \geq \tau_{ij}), \tag{2.2}$$

where $\tau_{ij} > 0$ is an entry-dependent adaptive threshold (Cai and Liu, 2011). In particular, the constant thresholding $\tau_{ij} = \delta$ is allowed. An example of the adaptive thresholding is

$$\tau_{ij} = \tau(\hat{r}_{ii}\hat{r}_{jj})^{1/2}, \quad \text{for a given } \tau > 0 \tag{2.3}$$

where $\hat{r}_{ii}$ is the i^{th} diagonal element of $\widehat{\mathbf{R}}$. This corresponds to applying the thresholding with parameter τ to the correlation matrix of $\widehat{\mathbf{R}}$. Moreover, instead of the hard-thresholding, one can also use more general thresholding functions of Antoniadis and Fan (2001), as in Rothman et al. (2009) and Cai and Liu (2011).

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

$$\widehat{\mathbf{\Sigma}}^{\mathcal{T}} = \sum_{i=1}^{K} \widehat{\lambda}_i \widehat{\boldsymbol{\xi}}_i \widehat{\boldsymbol{\xi}}_i' + \widehat{\mathbf{R}}^{\mathcal{T}}. \tag{2.4}$$

We will call the estimator as 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. In practice, the number of principal components 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. See, for example, Hallin and Liška (2007), Kapetanios (2010), Onatski (2010), etc, in latent factor models.

With the choice of τ_{ij} in (2.3), our estimator encompasses many popular estimators as its specific cases. When $\tau = 0$, the estimator will be the sample covariance matrix and when $\tau = 1$, the estimator becomes that based on the strict factor model. When $K = 0$, our estimator is the same as the threhsolding estimator of Bickel and Levina (2008) or the adaptive thresholding estimator of Cai and Liu (2011) with a slight modification of the thresholding parameter that takes the standard error of the sample covariance.

2.2 High dimensional PCA and factor model

We now elucidate why PCA can be used for the factor analysis. The main reason is that when dimensionality p is large, the covariance matrix $\mathbf{\Sigma}$ has very spiked eigenvalues. For

$\mathbf{\Sigma}_u = (\sigma_{ij})_{p\times p}$ as in (1.3), define

$$m_p = \max_{i\le p} \sum_{j\le p} I(\sigma_{ij} \neq 0). \tag{2.5}$$

which is assumed to grow slowly in p. Hence $\mathbf{\Sigma}_u$ is a sparse covariance matrix. The rationale behind this assumption is that, after the common factors are taken out, many pairs of the cross-sectional units become uncorrelated. As we now demonstrate, we need the following lemma.

Lemma 2.1. *Let $\{\lambda_i\}_{i=1}^p$ be eigenvalues of $\mathbf{\Sigma}$ in descending order and $\{\boldsymbol{\xi}_i\}_{i=1}^p$ be their associated eigenvectors. Correspondingly, let $\{\widehat{\lambda}_i\}_{i=1}^p$ be eigenvalues of $\widehat{\mathbf{\Sigma}}$ in descending order and $\{\widehat{\boldsymbol{\xi}}_i\}_{i=1}^p$ be their associated eigenvectors.*

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

2. **($\sin\theta$ Theorem, Davis and Kahn, 1970)**

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

When the factor loadings $\{\mathbf{b}_i\}_{i=1}^p$ are a random sample from a certain population, we then have

$$p^{-1}\sum_{i=1}^p \mathbf{b}_i\mathbf{b}_i' = p^{-1}\mathbf{B}'\mathbf{B} \to E\mathbf{b}_i\mathbf{b}_i', \text{ as } p\to\infty, \tag{2.6}$$

under some mild conditions. Note that the linear space spanned by the first K principal components of $\mathbf{B}\text{cov}(\mathbf{f}_t)\mathbf{B}'$ is the same as that spanned by the columns of $\mathbf{B}$ when $\text{cov}(\mathbf{f}_t)$ is non-degenerate. Thus, we can assume without loss of generality that the columns of $\mathbf{B}$ are orthogonal and $\text{cov}(\mathbf{f}_t) = \mathbf{I}_K$, the identity matrix. The assumptions correspond to the identifiability condition in decomposition (1.3). Let $\widetilde{\mathbf{b}}_1, \cdots, \widetilde{\mathbf{b}}_K$ be the columns of $\mathbf{B}$, ordered such that $\{\|\widetilde{\mathbf{b}}_j\|\}_{j=1}^K$ is in a non-increasing order. Then, $\{\widetilde{\mathbf{b}}_j/\|\widetilde{\mathbf{b}}_j\|\}_{j=1}^K$ are principal components of the matrix $\mathbf{BB}'$ with eigenvalues $\{\|\widetilde{\mathbf{b}}_j\|\}_{j=1}^K$ and the rest zero.

Let $\{\bar{\lambda}_j\}_{j=1}^K$ be the eigenvalues (in non-increasing order) of $p^{-1}\mathbf{B}'\mathbf{B}$, which are bounded from below and above by (2.6), assuming $K \ll p$. Since the non-vanishing eigenvalues of the matrix $\mathbf{BB}'$ are the same as those of $\mathbf{B}'\mathbf{B}$, the non-vanishing eigenvalues of the matrix $\mathbf{BB}'$ are $\{p\bar{\lambda}_j\}_{j=1}^K$, and $p\bar{\lambda}_j = \|\widetilde{\mathbf{b}}_j\|$. Correspondingly, let $\{\lambda_i\}_{i=1}^p$ be the values of $\mathbf{\Sigma}$ in a descending order and $\{\boldsymbol{\xi}_j\}_{j=1}^p$ be their corresponding eigenvectors. Then, an application of Weyl's theorem yields that

Proposition 2.1. *For the factor model (1.3) with the normalization condition*

$$\mathrm{cov}(\mathbf{f}_t) = \mathbf{I}_K \text{ and } \mathbf{B}'\mathbf{B} \text{ is diagonal}, \tag{2.7}$$

we have

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

If in addition (2.6) holds with $\lambda_{\min}(E\mathbf{b}_i\mathbf{b}_i')$ bounded away from zero, then $\|\widetilde{\mathbf{b}}_j\|/p$ is bounded away from zero for all large p.

The above proposition reveals that the first K eigenvalues of $\mathbf{\Sigma}$ is of order p, whereas the rest is only of order $\|\mathbf{\Sigma}_u\|$. Under the sparsity assumption (2.5), with the notation $\mathbf{\Sigma}_u = (\sigma_{u,ij})$, we have

$$\|\mathbf{\Sigma}_u\| \le \|\mathbf{\Sigma}_u\|_1 \le m_p \max_i \sigma_{u,ii}.$$

Thus, when $m_p = o(p)$ and $\max_i \sigma_{u,ii}$ is bounded, we have distinguished eigenvalues between the principal components and the rest of the components. More generally,

$$\|\mathbf{\Sigma}_u\| \le \|\mathbf{\Sigma}_u\|_1 \le \max_i \sum_{j=1}^{p} |\sigma_{u,ij}|^q (\sigma_{u,ii}\sigma_{u,jj})^{(1-q)/2}, \quad \text{for } q \le 1. \tag{2.8}$$

If we assume that the right-hand side of (2.8) is bounded by $c_p \ll p$, then the conclusion continues to hold. This generalizes the notion of sparsity defined by (2.5), which corresponds to $q = 0$ and was used in Bickel and Levina (2008) and Cai and Liu (2011).

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

Proposition 2.2. *Under the normalization (2.7), if $\{\|\widetilde{\mathbf{b}}_j\|\}_{j=1}^K$ are distinct , then*

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

Proposition 2.2 reveals that when p is large, the principal components $\{\boldsymbol{\xi}_j\}_{j=1}^K$ are close to the normalized vector $\{\widetilde{\mathbf{b}}_j\}_{j=1}^K$. The space spanned by the first K principal components are close to the space spanned by the columns of the factor loading matrix $\mathbf{B}$. This provides the mathematics for using the principal components as proxy of the space spanned by the columns of the loading matrix. Our conditions for the consistency of principal components are much weaker than those in Jung and Marron (2009).

A penalized least-squares approach for the low-rank plus sparse matrix is to minimize,

with respect to $\mathbf{B}$ and $\mathbf{S}$, the objective function

$$\|\widehat{\mathbf{\Sigma}} - \mathbf{B}\mathbf{B}' - \mathbf{S}\|_F^2 + \lambda \text{rank}(\mathbf{B}) + \sum_{i \neq j} p_\lambda(|s_{ij}|). \tag{2.9}$$

For example, Luo (2011) takes $p_\lambda(|\theta|) = \lambda|\theta|$ (but including the penalty in the diagnonal term) and relax the penalty on rank($\mathbf{B}$) by its nuclear norm (the L_1-norm of singular values). The advantage of our approach is that no optimization is needed. Even when the rank of $\mathbf{B}$ is known, and $p_\lambda(|\theta|) = \lambda^2 - (\lambda - |\theta|)_+^2$ (Antoniadis and Fan, 2001) is the hard thresholding penalty, the minimization of (2.9) still consists of two-step iterations: Given $\mathbf{B}$, $\mathbf{S}$ is the thresholded matrix $\widehat{\mathbf{\Sigma}} - \mathbf{B}\mathbf{B}'$, and given $\mathbf{S}$, $\mathbf{B}$ is the un-normalized principal components of $\widehat{\mathbf{\Sigma}} - \mathbf{S}$. With the POET, the above iterations are avoided and Propositions 2.1 and 2.2 provide some mathematical justifications.

2.3 Least squares point of view

The POET (2.4) has an equivalent representation using a constrained least squares method, due to the low-rank decomposition (1.3). In the latent factor model (1.1), the least squares method seeks for $\widehat{\mathbf{\Lambda}} = (\widehat{\mathbf{b}}_1, ..., \widehat{\mathbf{b}}_p)'$ and $\widehat{\mathbf{f}}_t$ such that

$$(\widehat{\mathbf{\Lambda}}, \widehat{\mathbf{f}}_t) = \arg\min_{\mathbf{b}_i, \mathbf{f}_t} \sum_{i=1}^{p} \sum_{t=1}^{T} (y_{it} - \mathbf{b}_i' \mathbf{f}_t)^2, \tag{2.10}$$

subject to the normalization

$$\frac{1}{T} \sum_{t=1}^{T} \widehat{\mathbf{f}}_t \widehat{\mathbf{f}}_t' = \mathbf{I}_K, \text{ and } \frac{1}{p} \sum_{i=1}^{p} \widehat{\mathbf{b}}_i \widehat{\mathbf{b}}_i' \text{ is diagonal.} \tag{2.11}$$

The constraints (2.11) correspond to the normalization (2.7). Here, through inclusion of the intercept terms a_i in (2.10) if necessary, we assume that the mean of each variable $\{y_{it}\}_{t=1}^T$ has been removed so have been the factors $\{\mathbf{f}_t\}_{t=1}^T$. Putting in the matrix form, the optimization problem can be written as

$$\begin{aligned} &\arg\min_{\mathbf{B}, \mathbf{F}} \|\mathbf{Y} - \mathbf{B}\mathbf{F}'\|_F^2 \\ &\mathbf{F}'\mathbf{F} = \mathbf{I}_K, \quad \mathbf{B}'\mathbf{B} \text{ is diagonal.} \end{aligned} \tag{2.12}$$

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 is $\mathbf{\Lambda} = T^{-1}\mathbf{Y}\mathbf{F}$, using the constraint (2.11) on the factors. Substituting this into (2.12), the

object function now becomes

$$\|\mathbf{Y} - T^{-1}\mathbf{Y}\mathbf{F}\mathbf{F}'\|_F^2 = \text{tr}[(\mathbf{I}_T - T^{-1}\mathbf{F}\mathbf{F}')\mathbf{Y}'\mathbf{Y}].$$

The minimizer is now clear: the columns of $\widehat{\mathbf{F}}$ are the eigenvectors corresponding to the K largest eigenvalues of the $T \times T$ matrix $T^{-1}\mathbf{Y}'\mathbf{Y}$ and $\widehat{\mathbf{\Lambda}} = T^{-1}\mathbf{Y}\widehat{\mathbf{F}}$.

We will show that under some mild regularity conditions, as p and $T \to \infty$, $\widehat{\mathbf{b}}_i'\widehat{\mathbf{f}}_t$ consistently estimates $\mathbf{b}_i'\mathbf{f}_t$ for each i and t. Since $\mathbf{\Sigma}_u$ is assumed to be sparse, we can construct an estimator of $\mathbf{\Sigma}_u$ using the adaptive thresholding method by Cai and Liu (2011) as follows. For some pre-determined decreasing sequence $\omega_T > 0$, let

$$\hat{u}_{it} = y_{it} - \widehat{\mathbf{b}}_i'\widehat{\mathbf{f}}_t, \quad \widehat{\sigma}_{ij} = \frac{1}{T}\sum_{t=1}^{T}\hat{u}_{it}\hat{u}_{jt}, \quad \text{and} \quad \hat{\theta}_{ij} = \frac{1}{T}\sum_{t=1}^{T}(\hat{u}_{it}\hat{u}_{jt} - \widehat{\sigma}_{ij})^2.$$

Define $\widehat{\mathbf{\Sigma}}_u = (\hat{u}_{it})$,

$$\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}} = (\widehat{\sigma}_{ij}^{\mathcal{T}})_{p\times p}, \quad \text{and} \quad \widehat{\sigma}_{ij}^{\mathcal{T}} = \widehat{\sigma}_{ij}I(|\widehat{\sigma}_{ij}| \geq \sqrt{\hat{\theta}_{ij}}\omega_T). \tag{2.13}$$

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

$$\widetilde{\mathbf{\Sigma}}^{\mathcal{T}} = \widehat{\mathbf{\Lambda}}\widehat{\mathbf{\Lambda}}' + \widehat{\mathbf{\Sigma}}_u^{\mathcal{T}}, \tag{2.14}$$

and

$$(\widetilde{\mathbf{\Sigma}}^{\mathcal{T}})^{-1} = (\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1} - (\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1}\widehat{\mathbf{\Lambda}}[\mathbf{I}_K + \widehat{\mathbf{\Lambda}}'(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1}\widehat{\mathbf{\Lambda}}]^{-1}\widehat{\mathbf{\Lambda}}^T(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1}, \tag{2.15}$$

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

The following theorem shows that the two estimators based on either regularized PCA or least squares substitution are equivalent.

Theorem 2.1. *If the entry-dependent threshold in (2.2) is the same as the thresholding parameter used in (2.13). Then, $\widehat{\mathbf{\Sigma}}_u = \widehat{\mathbf{R}}$ and the estimator (2.4) is equivalent to the substitution estimator (2.14), that is,*

$$\widehat{\mathbf{\Sigma}}^{\mathcal{T}} = \widetilde{\mathbf{\Sigma}}^{\mathcal{T}}, \qquad \textit{and} \qquad \widehat{\mathbf{\Sigma}}_u^{\mathcal{T}} = \widehat{\mathbf{R}}^{\mathcal{T}}.$$

3 Asymptotic Properties

3.1 Assumptions

This section gives the assumptions for the convergence of $\widehat{\mathbf{\Sigma}}^{\mathcal{T}} - \mathbf{\Sigma}$ as well as $(\widehat{\mathbf{\Sigma}}^{\mathcal{T}})^{-1} - \mathbf{\Sigma}^{-1}$ under model (1.2), in which only $\{\mathbf{y}_t\}_{t=1}^T$ is observable. Recall that we impose the identifiability condition (2.7). We make the following assumptions.

Assumption 3.1. *(i)* $\{\mathbf{u}_t\}_{t\geq 1}$ *is stationary and ergodic with mean vector zero and covariance matrix* $\mathbf{\Sigma}_u$.
(ii) There exit constants $c_1, c_2 > 0$ *such that* $c_1 < \lambda_{\min}(\mathbf{\Sigma}_u) \leq \lambda_{\max}(\mathbf{\Sigma}_u) < c_2$.
(iii) There exist $r_1 > 0$ *and* $b_1 > 0$, *such that for any* $s > 0$ *and* $i \leq p$,

$$P(|u_{it}| > s) \leq \exp(-(s/b_1)^{r_1}).$$

Condition (ii) requires that $\mathbf{\Sigma}_u$ is well-conditioned as in Chamberlain and Rothschild (1983). As noted in Bickel and Levina (2004), this condition is satisfied if for each t, $\{u_{it}\}_{i=1}^{\infty}$ is a stationary and ergodic process with spectral density that is bounded away from both zero and infinity. Condition (iii) requires the distributions of the idiosyncratic terms to have exponential-type tails, which allows us to apply the large deviation theory to $\frac{1}{T}\sum_{t=1}^T u_{it}u_{jt} - \sigma_{u,ij}$.

Assumption 3.2. *(i)* $\{\mathbf{f}_t\}_{t\geq 1}$ *is stationary and ergodic.*
(ii) $\{\mathbf{u}_t\}_{t\geq 1}$ *and* $\{\mathbf{f}_t\}_{t\geq 1}$ *are independent.*

We introduce the strong mixing conditions to conduct asymptotic analysis of the least square estimates. Let $\mathcal{F}_{-\infty}^0$ and $\mathcal{F}_T^{\infty}$ denote the σ-algebras generated by $\{(\mathbf{f}_t, \mathbf{u}_t) : -\infty \leq t \leq 0\}$ and $\{(\mathbf{f}_t, \mathbf{u}_t) : T \leq t \leq \infty\}$ respectively. In addition, define the mixing coefficient

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

Assumption 3.3. *(i) Exponential tail: There exist* $b_2 > 0$ *and* $r_2 > 0$ *such that for all* i, t, *and* $s > 0$,

$$P(|f_{it}| > s) \leq \exp(-(s/b_2)^{r_2}).$$

(ii) 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,

$$\max_{t\le T}\sum_{s=1}^{T}|E\mathbf{u}_s'\mathbf{u}_t|/p = O(1).$$

In addition, we impose the following regularity conditions.

Assumption 3.4. *There exists $M > 0$ such that for all $i \le p$ and $t \le T$,*
(i) $\|\mathbf{b}_i\|_{\max} < M$, and $E\|\mathbf{f}_t\|^4 < K^2M$,
(ii) $E[p^{-1/2}(\mathbf{u}_s'\mathbf{u}_t - E\mathbf{u}_s'\mathbf{u}_t)]^4 < M$,
(iii) $E\|(pK)^{-1/2}\sum_{i=1}^{p}\mathbf{b}_i u_{it}\|^4 < M$.

This assumption provides regularity conditions in order to consistently estimate the transformed common factors as well as the factor loadings. Conditions (ii) and (iii) are satisfied, for example, if $\{u_{it}\}_{i=1}^{\infty}$ and $\{b_{ik}u_{it}\}_{i=1}^{\infty}$ are stationary and ergodic processes with finite fourth moment for each t, k. For simplicity, we only consider nonrandom factor loadings and a known number of factors. Note that the conditions here are weaker than those in Bai (2003), as there is no need to establish the asymptotic normality in order to estimate the covariance matrix.

The following assumption requires that the factors should be pervasive, that is, impact every individual time series (Harding, 2009). See also (2.6).

Assumption 3.5. $\|p^{-1}\mathbf{B}'\mathbf{B} - \Omega\| = o(1)$ *for some $K \times K$ symmetric positive definite matrix Ω such that $\lambda_{\min}(\Omega)$ is bounded away from both zero and infinity.*

3.2 Convergence of the idiosyncratic covariance

Estimating the covariance matrix $\mathbf{\Sigma}_u$ of the idiosyncratic components $\{\mathbf{u}_t\}$ is important for many statistical inference purposes. For example, it is needed for large sample inference of the unknown factors and loadings in Bai (2003), Wang (2010), for testing the capital asset pricing model in Sentana (2009), and large-scale testing in Fan, Han and Gu (2012). See Section 4 for the last two applications.

We estimate $\mathbf{\Sigma}_u$ by applying the adaptive thresholding given by (2.13). By Theorem 2.1, it is also the same as $\widehat{\mathbf{R}}^{\mathcal{T}}$ given by (2.2). We apply the POET estimator with adaptive threshold:

$$\tau_{ij} = C\sqrt{\hat{\theta}_{ij}}\omega_T, \tag{3.2}$$

where $C > 0$ is a sufficiently large constant, and throughout the paper we denote

$$\omega_T = \frac{K\sqrt{\log p} + K^2}{\sqrt{T}} + \frac{K^3}{\sqrt{p}} + \sqrt{\frac{\log p}{T}} \quad \text{and} \quad \delta_T = m_p\omega_T. \tag{3.3}$$

The following theorem shows that $\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}}$ is asymptotically nonsingular as T and $p \to \infty$. The rate of convergence under the spectral norm is also derived. Let $\gamma^{-1} = 3r_1^{-1} + 1.5r_2^{-1} + r_3^{-1}$.

Theorem 3.1. *Suppose* $\max\{(\log p)^{6/\gamma-1}, K^4(\log(pT))^2\} = o(T)$, *and* $T^{1/4}K^3 = o(\sqrt{p})$. *Under Assumptions 3.1-3.5, the POET estimator defined in (2.13) satisfies*

$$\|\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}} - \mathbf{\Sigma}_u\| = O_p(\delta_T).$$

If further $\delta_T = o(1)$, *then* $\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}}$ *is invertible with probability approaching one, and*

$$\|(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1} - \mathbf{\Sigma}_u^{-1}\| = O_p(\delta_T).$$

When estimating $\mathbf{\Sigma}_u$, p is allowed to grow exponentially fast in T, that is, there exists $a > 0$, as long as $\log p = O(T^a)$, $\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}}$ can be made consistent under the spectral norm. In addition, $\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}}$ is invertible while the classical sample covariance matrix based on the residuals is not when $p > T$.

Remark 3.1. Fan, Liao and Mincheva (2011) showed that when $\{\mathbf{f}_t\}_{t=1}^T$ are observable, the rate of convergence of the adaptive thresholding estimator is given by

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

Hence when the common factors are unobservable, the rate of convergence has an additional term $m_p K^3/\sqrt{p}$, coming from the impact of estimating the unknown factors. This impact vanishes when $p \log p \gg K^4 T$. As p increases, more information about the common factors is collected, which results in more accurate estimation of the common factors $\{\mathbf{f}_t\}_{t=1}^T$. The estimation error becomes negligible when p is significantly larger than T. If in addition K is bounded, the rate of convergence achieves the minimax rate as in Cai and Zhou (2010).

3.3 Convergence of the POET estimator

As was found by Fan et al. (2008), deriving the rate of convergence under the spectral norm for $\widehat{\mathbf{\Sigma}}^{\mathcal{T}} - \mathbf{\Sigma}$ is inappropriate. We illustrate this using a simple example.

Example 3.1. Consider an ideal case where we know $\mathbf{b}_i = (1, 0, ..., 0)'$ for each $i = 1, ..., p$, $\mathbf{\Sigma}_u = \mathbf{I}_p$, and $\{\mathbf{f}_t\}_{t=1}^T$ are observable. Then when estimating $\mathbf{\Sigma}$, we only need to estimate $\text{cov}(\mathbf{f}_t)$ using the sample covariance matrix $\widehat{\text{cov}}(\mathbf{f}_t)$, and obtain an estimator for $\mathbf{\Sigma}$:

$$\widehat{\mathbf{\Sigma}} = \mathbf{B}\widehat{\text{cov}}(\mathbf{f}_t)\mathbf{B}' + \mathbf{I}_p.$$

Simple calculations yield to

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

where $\mathbf{1}_p$ denotes the p-dimensional column vector of ones with $\|\mathbf{1}_p\mathbf{1}_p'\| = p$. Therefore, $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\| = O_p(p/\sqrt{T})$, which is $o_p(1)$ only if $p = o(\sqrt{T})$. □

Alternatively, Fan et al. (2008) suggested to use the weighted quadratic loss to study the rate of convergence in high dimensional factor models:

$$\|\mathbf{A}\|_\Sigma = p^{-1/2}\|\mathbf{\Sigma}^{-1/2}\mathbf{A}\mathbf{\Sigma}^{-1/2}\|_F.$$

which is closely related to the entropy loss, introduced by James and Stein (1961). Here $p^{-1/2}$ is a normalization factor such that $\|\mathbf{\Sigma}\|_\Sigma = 1$. Technically, the impact of high dimensionality on the convergence rate of $\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}$ is via the number of factor loadings in $\mathbf{B}$. It was shown by Fan et al. (2008) that $\mathbf{B}$ appears in $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_\Sigma$ through $\mathbf{B}'\mathbf{\Sigma}^{-1}\mathbf{B}$, and that $\lambda_{\max}(\mathbf{B}'\mathbf{\Sigma}^{-1}\mathbf{B})$ is bounded, which successfully cancels out the curse of high dimensionality introduced by $\mathbf{B}$. The following theorem gives the rate of convergence under various norms.

Theorem 3.2. *Under the assumptions of Theorem 3.1, the POET estimator defined in (2.4) satisfies*

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

$$\|\widehat{\mathbf{\Sigma}}^{\mathcal{T}} - \mathbf{\Sigma}\|_{\max} = O_p(K^3\sqrt{\frac{\log K}{T}} + \omega_T).$$

In addition, if $\delta_T = o(1)$, then $\widehat{\mathbf{\Sigma}}^{\mathcal{T}}$ is nonsingular with probability approaching one, with

$$\|(\widehat{\mathbf{\Sigma}}^{\mathcal{T}})^{-1} - \mathbf{\Sigma}^{-1}\| = O_p\left(\delta_T\right).$$

In practical applications, K is typically small compared to p and T. It can even be thought of as a constant. For example, in the Fama-French model, $K = 3$. Our result also covers the case of $K = 0$, when the target covariance $\mathbf{\Sigma} = \mathbf{\Sigma}_u$ is a sparse matrix. The POET estimator is then the same as either the thresholding estimator of Bickel and Levina (2008) or the adaptive thresholding estimator of Cai and Liu (2011). Theorem 3.2 then implies

$$\|\widehat{\mathbf{\Sigma}}^{\mathcal{T}} - \mathbf{\Sigma}\|_\Sigma = O_p\left(\sqrt{\frac{\log p}{T}}\right) = \|(\widehat{\mathbf{\Sigma}}^{\mathcal{T}})^{-1} - \mathbf{\Sigma}^{-1}\|.$$

Remark 3.2. When estimating $\mathbf{\Sigma}^{-1}$, p is allowed to grow exponentially fast in T, and the estimator has the same rate of convergence as the estimator $\widehat{\mathbf{\Sigma}}_u$ in Theorem 3.1. When p becomes much larger than T, the precision matrix can be estimated as if the factors were observable. This fact was also shown in the asymptotic expansions obtained by Bai (2003) when K is finite and $\mathbf{\Sigma}_u$ is diagonal. Therefore in the approximate factor model, we can do a much better job in estimating $\mathbf{\Sigma}^{-1}$ than estimating $\mathbf{\Sigma}$. The intuition follows from the fact that $\mathbf{\Sigma}^{-1}$ has bounded eigenvalues whereas $\mathbf{\Sigma}^{-1}$ has diverging eigenvalues.

3.4 Convergence of unknown factors and factor loadings

Many applications of the factor models contain the latent factors, and as a result, the common factors need to be estimated by the method of principal components. In this case, the factor loadings in $\mathbf{B}$ and the common factors $\mathbf{f}_t$ are not separably identifiable, as for any nonsingular $K \times K$ matrix $\mathbf{H}$, $\mathbf{B}\mathbf{f}_t = \mathbf{B}\mathbf{H}^{-1}\mathbf{H}\mathbf{f}_t$. Hence $(\mathbf{B}, \mathbf{f}_t)$ cannot be identified from $(\mathbf{B}\mathbf{H}^{-1}, \mathbf{H}\mathbf{f}_t)$. However, this ambiguity is eliminated by the identifiability condition (2.7), subject to a permutation. Note that the linear space spanned by $\mathbf{B}$ is the same as that by $\mathbf{B}\mathbf{H}^{-1}$. In practice, it often does not matter which one to be used.

Let $\mathbf{V}$ denote the $K \times K$ diagonal matrix of the first K largest eigenvalues of the sample covariance matrix in decreasing order. Recall that $\mathbf{F}' = (\mathbf{f}_1, ..., \mathbf{f}_T)$ and let

$$\mathbf{H} = \frac{1}{T}\mathbf{V}^{-1}\widehat{\mathbf{F}}'\mathbf{F}\mathbf{B}'\mathbf{B}.$$

Then for $t = 1, ..., T$,

$$\mathbf{H}\mathbf{f}_t = \frac{1}{T}\mathbf{V}^{-1}\widehat{\mathbf{F}}'(\mathbf{B}\mathbf{f}_1, ..., \mathbf{B}\mathbf{f}_T)'\mathbf{B}\mathbf{f}_t.$$

Note that $\mathbf{H}\mathbf{f}_t$ depends only on the data $\mathbf{V}^{-1}\widehat{\mathbf{F}}'$ and identifiable part of parameters $\{\mathbf{B}\mathbf{f}_t\}_{t=1}^T$. Therefore, there is no identifiability issue in $\mathbf{H}\mathbf{f}_t$ no matter the identifiability condition (2.7) is imposed.

Bai (2003) showed that when K is fixed and $\mathbf{\Sigma}_u$ is diagnonal, $\widehat{\mathbf{b}}_i$ and $\widehat{\mathbf{f}}_t$ are consistent estimators of $\mathbf{H}'^{-1}\mathbf{b}_i$ and $\mathbf{H}\mathbf{f}_t$ respectively. The following theorem allows K to grow slowly and $\mathbf{\Sigma}_u$ to be a sparse non-diagonal matrix.

Theorem 3.3. *Under the assumptions of Theorem 3.1,*

$$\max_{i \le p} \|\widehat{\mathbf{b}}_i - (\mathbf{H}')^{-1}\mathbf{b}_i\| = O_p\left(\frac{\omega_T}{\sqrt{K}}\right),$$

and

$$\max_{t \le T} \|\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t\| = O_p(\delta_T^*),$$

where $\delta_T^* = \sqrt{K/T} + K^{3/2}T^{1/4}/\sqrt{p}$.

As a consequence of Theorem 3.3, we obtain the following

Corollary 3.1. *Under the assumptions of Theorem 3.1,*

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

4 Applications

We give four examples which are immediate applications of the results in Theorems 3.1–3.3.

Example 4.1 (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). 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 $Z_i \sim N(\mu_i, 1)$. These test statistics $\mathbf{Z}$ are correlated and follow $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with unknown covariance matrix $\boldsymbol{\Sigma}$. For a given critical value x, the false discovery proportion is then defined as $\text{FDP}(x) = V(x)/R(x)$ where

$$V(x) = p^{-1}\sum_{\mu_i=0} I(|Z_i| > x) \quad \text{and} \quad R(x) = p^{-1}\sum_{i=1}^{p} I(|Z_i| > x)$$

are the total number of false discoveries and the total number of discoveries, respectively. Our interest is to estimate $\text{FDP}(x)$ 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

$$\mathbf{Z} = \boldsymbol{\mu} + \mathbf{B}\mathbf{f} + \mathbf{u}, \qquad \text{where } \boldsymbol{\Sigma}_u \text{ is sparse.} \tag{4.1}$$

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

$$V(x) = \sum_{i=1}^{p}\{\Phi(a_i(z_{x/2} + \eta_i)) + \Phi(a_i(z_{x/2} - \eta_i))\} + o_P(p), \tag{4.2}$$

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 $a_i = \mathrm{var}(u_i)$.

Now suppose that we have n repeated measurements from the model (4.1). Then, by Corollary 3.1, $\{\eta_i\}$ can be uniformly estimated, and hence $p^{-1}V(x)$ can be consistently estimated and hence FDP(x). Efron (2010) obtained these repeated test statistics based on the bootstrap sample from the original raw data. Our theory gives a formal justification to the seminal work of Efron (2007, 2010).

Example 4.2 (Risk management)**.** The maximum elementwise estimation error $\|\widehat{\mathbf{\Sigma}}^{\mathcal{T}} - \mathbf{\Sigma}\|_{\max}$ appears in risk assessment as in Fan, Zhang and Yu (2008). For a fixed portfolio allocation vector $\mathbf{w}$, the true portfolio variance and the estimated one are given by $\mathbf{w}'\mathbf{\Sigma}\mathbf{w}$ and $\mathbf{w}'\widehat{\mathbf{\Sigma}}^{\mathcal{T}}\mathbf{w}$ respectively. The estimation error is bounded by

$$|\mathbf{w}'\widehat{\mathbf{\Sigma}}^{\mathcal{T}}\mathbf{w} - \mathbf{w}'\mathbf{\Sigma}\mathbf{w}| \leq \|\widehat{\mathbf{\Sigma}}^{\mathcal{T}} - \mathbf{\Sigma}\|_{\max}\|\mathbf{w}\|_1^2,$$

where $\|\mathbf{w}\|_1$, the l_1 norm of $\mathbf{w}$, is the gross exposure of the portfolio. Usually a constraint is placed on the total percentage of the short positions, in which case we have a restriction $\|\mathbf{w}\|_1 \leq c$ 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.

Example 4.3 (Panel regression with a factor structure in the errors)**.** Consider the following panel regression model

$$\begin{aligned} Y_{it} &= \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \quad i \leq p, t \leq T, \\ \varepsilon_{it} &= \mathbf{b}_i'\mathbf{f}_t + u_{it}, \end{aligned}$$

where $\mathbf{x}_{it}$ is a vector of observable regressors with fixed dimension. The regression error ε_{it} has a factor structure, but $\mathbf{b}_i$, $\mathbf{f}_t$ and u_{it} are all unobservable. Assuming $\mathbf{x}_{it}$ to be independent of ε_{it}, we are interested in the common regression coefficients $\boldsymbol{\beta}$. The above panel regression model has been considered by many researchers, such as Ahn, Lee and Schmidt (2001), Pesaran (2006), etc, which has broad applications in social sciences. For example, in the income studies, Y_{it} represents the income of individual i at age t, $\mathbf{x}_{it}$ is a vector of observable characteristics that are associated with income. Here $\mathbf{b}_i$ represents a vector of unmeasured skills, such as innate ability, motivation, and hardworking; $\mathbf{f}_t$ is a vector of unobservable prices for the unmeasured skills, which is assumed to be time-varying.

Although OLS (ordinary least squares) produces a consistent estimator of $\boldsymbol{\beta}$, a more efficient estimation would be obtained by GLS (generalized least squares). The GLS method

depends on an estimator of $\mathbf{\Sigma}_\epsilon^{-1}$, the inverse of the covariance matrix of $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, ..., \varepsilon_{pt})'$, which should be consistent under the spectral norm. In a large panel model, p can be larger than T. As a result, the traditional GLS, which estimates $\mathbf{\Sigma}_\epsilon^{-1}$ based on the sample residual covariance, is infeasible.

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 $\hat{\varepsilon}_{it}$, obtained via first regressing Y_{it} on $\mathbf{x}_{it}$. We then apply the POET estimator to $T^{-1}\sum_{t=1}^T \widehat{\boldsymbol{\varepsilon}}_t\widehat{\boldsymbol{\varepsilon}}_t'$, as described in Section 2.1. By Theorem 3.2, the inverse of the resulting estimator is a consistent estimator of $\mathbf{\Sigma}_\epsilon^{-1}$ under the spectral norm. A slight difference lies in the fact that when we apply POET, $T^{-1}\sum_{t=1}^T \boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'$ is replaced with $T^{-1}\sum_{t=1}^T \widehat{\boldsymbol{\varepsilon}}_t\widehat{\boldsymbol{\varepsilon}}_t'$, which introduces an additional term $O_p(\sqrt{\frac{\log p}{T}})$ in the estimation error.

Example 4.4 (Testing for 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 returns of any financial asset equals to the excessive returns of the risk factors plus idiosyncratic noises that are only related to the asset itself. 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 and u_{it} is the idiosyncratic noise. To test the null hypothesis (1.2), one embeds the model into the multivariate linear model

$$\mathbf{y}_t = \boldsymbol{\mu} + \mathbf{B}\mathbf{f}_t + \mathbf{u}_t, \qquad t = 1, \cdots, T \tag{4.3}$$

and wishes to test $H_0 : \boldsymbol{\mu} = 0$. The F-test statistic involves the estimation of the covariance matrix $\mathbf{\Sigma}_u$, whose estimates are degenerate without regularization when $p \geq T$ even if $\mathbf{f}_t$ are observable risk factors. Therefore, in the literature (Sentana, 2009, and references therein), one focuses on the case $p \ll T$. 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 regularized technique, non-degenerate estimate $\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}}$ can be obtained and the F-test or likelihood-ratio test statistics can be employed even when $p \gg T$.

5 Monte Carlo Experiments

In this section, we will examine the performance of the POET method in the finite sample. We will also demonstrate the effect 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 unobserved 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 trivariate normal distributions $\mathbf{b} \sim N_3(\boldsymbol{\mu}_B, \boldsymbol{\Sigma}_B)$, the idiosyncratic errors from $\mathbf{u}_t \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}_u)$, and the factor returns $\mathbf{f}_t$ follow an VAR(1) model. To make the simulation more realistic, model parameters are calibrated from the financial returns, as detailed in the following section.

5.1 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 Jan 1^{st}, 2009 to Dec 31^{st}, 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 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 calculate the sample mean vector $\boldsymbol{\mu}_B$ and sample covariance matrix $\boldsymbol{\Sigma}_B$ of the rows of $\tilde{\mathbf{B}}$, 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 $N_3(\boldsymbol{\mu}_B, \boldsymbol{\Sigma}_B)$.

Table 1: Mean and covariance matrix used to generate $\mathbf{b}$

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

3. Assume that the factors $\mathbf{f}_t$ follow the stationary vector autoregressive model, $\mathbf{f}_t = \boldsymbol{\mu} + \boldsymbol{\Phi}\mathbf{f}_{t-1} + \boldsymbol{\varepsilon}_t$, a VAR(1) model. Then obtain the multivariate least squares estimator for $\boldsymbol{\mu}$ and $\boldsymbol{\Phi}$, and estimate $\boldsymbol{\Sigma}_\epsilon$, by autoregressing the columns of $\tilde{\mathbf{F}}$. Note that all eigenvalues of $\boldsymbol{\Phi}$ 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 below,

$$\text{cov}(\mathbf{f}_t) = \boldsymbol{\Phi}\text{cov}(\mathbf{f}_t)\boldsymbol{\Phi}' + \boldsymbol{\Sigma}_\epsilon.$$

The estimated parameters are depicted in Table 2. We then obtain the data generating process of $\mathbf{f}_t$.

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

4. For each value of p, we generate a sparse covariance matrix $\boldsymbol{\Sigma}_u$ of the form:

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

Here, $\boldsymbol{\Sigma}_0$ is the error correlation matrix, and $\mathbf{D}$ is the diagonal matrix of the standard deviations of the errors. We set $\mathbf{D} = \text{diag}(\sigma_1, ..., \sigma_p)$, 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 $\boldsymbol{\Sigma}_0$ are generated independently from a normal distribution, with mean and standard deviation equal to the sample mean and sample standard deviation of the sample correlations among the estimated residuals, conditional on their absolute values being no larger than 0.95. We then employ hard thresholding to make $\boldsymbol{\Sigma}_0$ be sparse, where the threshold is found as the smallest constant that provides the positive definiteness of $\boldsymbol{\Sigma}_0$. More precisely, start with threshold value 1, which gives $\boldsymbol{\Sigma}_0 = \mathbf{I}_p$ and then decrease the threshold values in grid until positive definiteness is violated.

5.2 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 $\{\mathbf{b}_i\}_{i=1}^{p} \sim N_3(\boldsymbol{\mu}_B, \boldsymbol{\Sigma}_B)$, and set $\mathbf{B} = (\mathbf{b}_1, ..., \mathbf{b}_p)'$.
2. Generate independently $\{\mathbf{u}_t\}_{t=1}^{T} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{u})$.
3. Generate $\{\mathbf{f}_t\}_{t=1}^{T}$ as a vector autoregressive sequence of the form $\mathbf{f}_t = \boldsymbol{\mu} + \Phi \mathbf{f}_{t-1} + \boldsymbol{\varepsilon}_t$.
4. Calculate $\{\mathbf{y}_t\}_{t=1}^{T}$ from $\mathbf{y}_t = \mathbf{B}\mathbf{f}_t + \mathbf{u}_t$.
5. Calculate $\widehat{\boldsymbol{\Lambda}}$ and $\widehat{\mathbf{F}}$ based on $\{\mathbf{y}_t\}_{t=1}^{T}$, using the PCA method described in Section 2.3.
6. Set $\omega_T = 2(K\sqrt{\frac{\log p}{T}} + \frac{K^3}{\sqrt{p}})$ with $K = 3$ to be the threshold for creating $\widehat{\boldsymbol{\Sigma}}_u^{\mathcal{T}}$. Calculate $\widehat{\boldsymbol{\Sigma}}^{\mathcal{T}}$ and $(\widehat{\boldsymbol{\Sigma}}^{\mathcal{T}})^{-1}$ using the POET method.
7. Calculate the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}_{sam}$.

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

5.3 Results

From the simulation results, reported in Figures 1-4, we observe that our estimator under the unobservable factor model performs just as well as the estimator in Fan et al. (2011) under the observable factor model, 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 1 and 2 that this cost vanishes for $p > 300$. To give a better insight of the impact of estimating the unknown factors for small p, a separate set of simulations is conducted for $p \leq 100$. As we can see from Figures 1 (bottom panel) and 2 (middle and bottom panels), the impact decreases quickly. In addition, when estimating $\boldsymbol{\Sigma}^{-1}$, it is hard to distinguish the estimators with known and unknown factors, whose performances are quite stable compared to the sample covariance

Figure 1: Averages (left panel) and standard deviations (right panel) of $\|\widehat{\mathbf{\Sigma}}^{\mathcal{T}} - \mathbf{\Sigma}\|_{\Sigma}$ with with known factors (solid red curve), unknown factors (solid blue curve), and $\|\widehat{\mathbf{\Sigma}}_{sam} - \mathbf{\Sigma}\|_{\Sigma}$ (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.

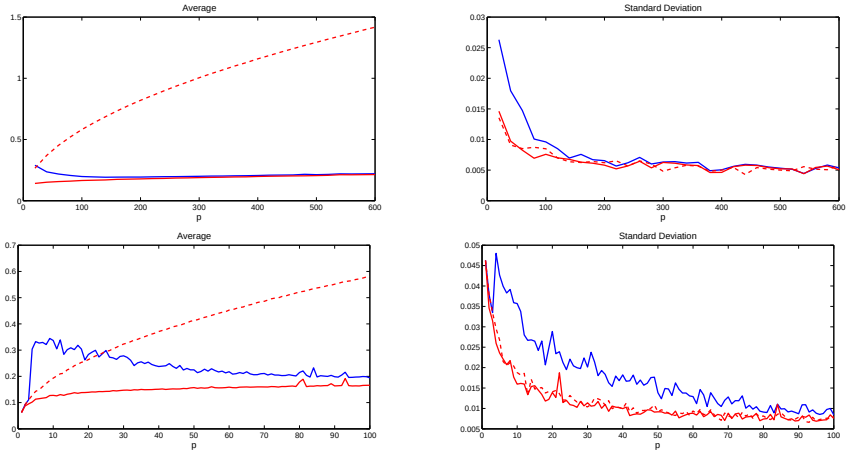


Figure 2: Averages (left panel) and standard deviations (right panel) of $\|(\widehat{\mathbf{\Sigma}}^{\mathcal{T}})^{-1} - \mathbf{\Sigma}^{-1}\|$ with known (solid red curve) and unknown (solid blue curve) factors and $\|(\widehat{\mathbf{\Sigma}}_{sam})^{-1} - \mathbf{\Sigma}^{-1}\|$ (dashed curve) 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.

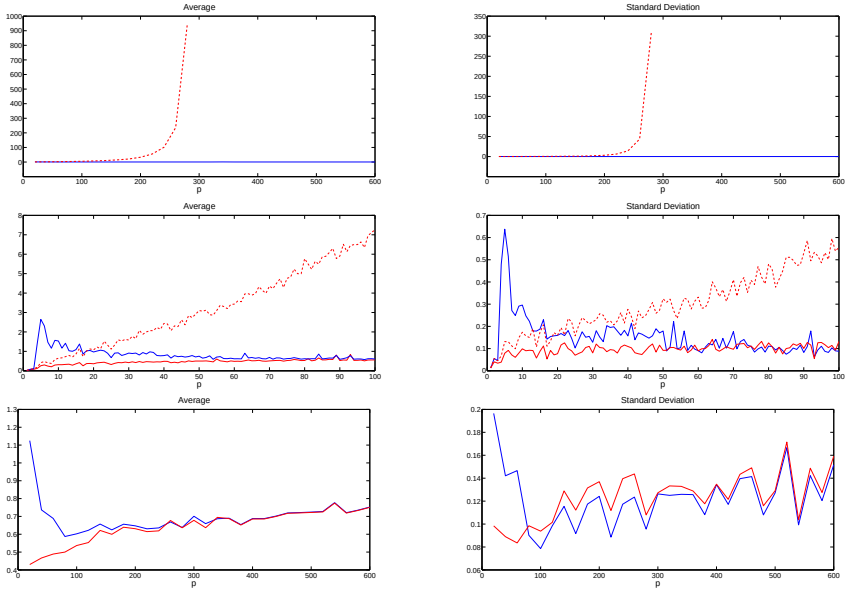
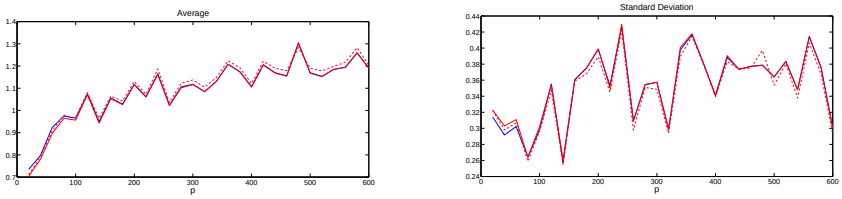
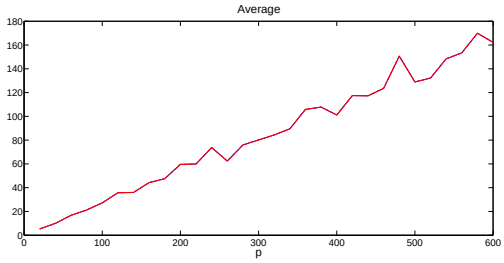


Figure 3: Averages (left panel) and standard deviations (right panel) $\|\widehat{\mathbf{\Sigma}}^{\mathcal{T}} - \mathbf{\Sigma}\|_{\max}$ with known (solid red curve) and unknown (solid blue curve) factors and $\|\widehat{\mathbf{\Sigma}}_{sam} - \mathbf{\Sigma}\|_{\max}$ (dashed curve) over 200 simulations, as a function of the dimensionality p. They are nearly indifferentable.



matrix. Also, the maximum absolute elementwise error (Figure 3) of our estimator performs very similarly to that of the sample covariance matrix, which coincides with our asymptotic result. Figure 4 shows that the performances of the three methods are indistinguishable in spectral norm, as expected.

Figure 4: Averages of $\|\widehat{\mathbf{\Sigma}}^{\mathcal{T}} - \mathbf{\Sigma}\|$ with known (solid red curve) and unknown (solid blue curve) factors and $\|\widehat{\mathbf{\Sigma}}_{sam} - \mathbf{\Sigma}\|$ (dashed curve) over 200 simulations, as a function of the dimensionality p. The three curves are plotted, but hardly distinguishable.



5.4 Portfolio allocation

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

Let $\widehat{\mathbf{\Sigma}}$ be a generic estimator of the covariance matrix of $\mathbf{y}_t$, 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 would be $R(\mathbf{w}) = \mathbf{w}'\mathbf{\Sigma}\mathbf{w}$ and $\hat{R}(\mathbf{w}) = \mathbf{w}'\widehat{\mathbf{\Sigma}}\mathbf{w}$, respectively. Now, define

$$\widehat{\mathbf{w}} = \text{argmin}_{\mathbf{w}'\mathbf{1}=1}\mathbf{w}'\widehat{\mathbf{\Sigma}}\mathbf{w},$$

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

For each fixed p, the population $\mathbf{\Sigma}$ was generated in the same way as described in Section 4.1, with a sparse but not diagonal error covariance. We use three different methods to estimate $\mathbf{\Sigma}$ and obtain $\widehat{\mathbf{w}}$: strict factor model $\widehat{\mathbf{\Sigma}}^{\text{diag}}$ (estimate $\mathbf{\Sigma}_u$ using a diagonal matrix), our POET estimator $\widehat{\mathbf{\Sigma}}^{\mathcal{T}}$, both are with unknown factors, and sample covariance $\widehat{\mathbf{\Sigma}}^{\text{sam}}$. We then calculate the corresponding actual and empirical risks.

It is interesting to examine the accuracy and the performance of the actual risk of our portfolio $\widehat{\mathbf{w}}$ in comparison to the oracle risk $R^* = \min_{\mathbf{w}'\mathbf{1}=1} \mathbf{w}'\mathbf{\Sigma}\mathbf{w}$, which is the theoretical risk of the portfolio we would have created if we knew the true covariance matrix $\mathbf{\Sigma}$. We thus compare the regret $R(\widehat{\mathbf{w}}) - R^*$, which is always nonnegative, for three estimators of $\widehat{\mathbf{\Sigma}}$. They are summarized by using the box plots over the 200 simulations. The results are reported in Figure 5. In practice, we are also concerned about the difference between the actual and empirical risk of the chosen portfolio $\widehat{\mathbf{w}}$. Hence, in Figure 6, we also compare the average difference $|R(\widehat{\mathbf{w}}) - \hat{R}(\widehat{\mathbf{w}})|$ over 200 simulations. When $\widehat{\mathbf{w}}$ is obtained based on the strict factor model, both the differences between actual and oracle risk, and between actual and empirical risk are persistently greater than the corresponding differences for the approximate factor estimator.

Figure 5: Box plots of regrets $R(\widehat{\mathbf{w}}) - R^*$ for $p = 80$ and 140. In each panel, the box plots from left to right correspond to $\widehat{\mathbf{w}}$ obtained using $\widehat{\mathbf{\Sigma}}$ based on approximate factor model, strict factor model, and sample covariance, respectively.

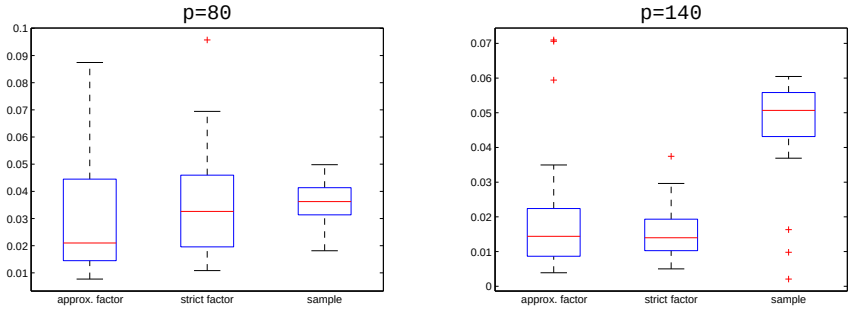
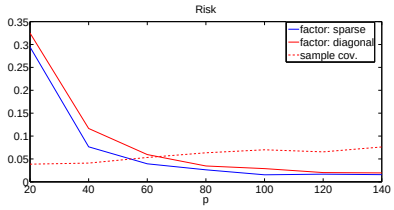


Figure 6: Error of risk estimation $|R(\widehat{\mathbf{w}}) - \hat{R}(\widehat{\mathbf{w}})|$ as p increases. $\widehat{\mathbf{w}}$ and $\hat{R}$ are obtained based on three estimators of $\widehat{\mathbf{\Sigma}}$. Blue line: approximate factor model $\widehat{\mathbf{\Sigma}}^{\mathcal{T}}$; red line: strict factor model $\widehat{\mathbf{\Sigma}}^{\text{diag}}$; dotted line: sample covariance $\widehat{\mathbf{\Sigma}}^{\text{sam}}$.



6 Real Data Example

We demonstrate an application of the approximate factor model on real data. The data was obtained from the CRSP database, and consists of $p = 50$ stocks and their annualized daily returns for the period Jan.1^{st}, 2010-Dec.31^{st}2010 ($T = 252$). The stocks are chosen from 5 different industry sectors, (more specifically, Consumer Goods-Textile & Apparel Clothing, Financial-Credit Services, Healthare-Hospitals, Services-Restaurants, Utilities-Water utilities), with 10 stocks from each sector.

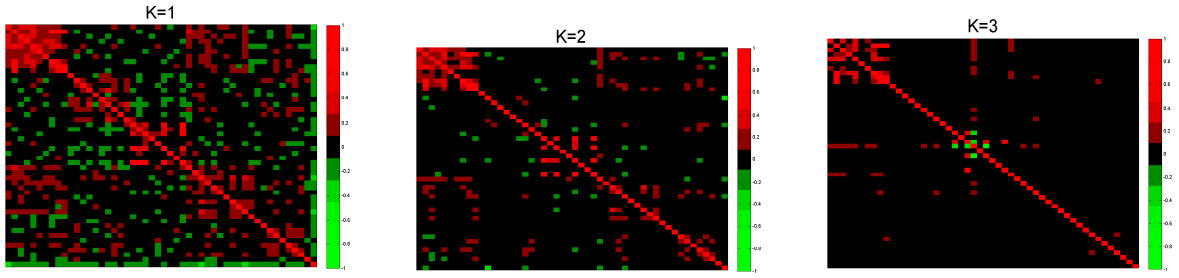
The largest eignevalues of the sample covariance equal $0.0102, 0.0045$ and 0.0039, while the rest are bounded by 0.0020. Hence we consider $K = 1, 2, 3$ factors respectively. The threshold has been chosen using a leave-one-out cross-validation procedure.

Figure 7 shows the heatmap of the thresholded error correlation matrix. We compare the level of sparsity (percentage of non-zero off-diagonal elements) for the diagonal 5 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 $K = 1$ and $K = 3$ show the same characteristics. These provide stark evidence that the strict factor model is not appropriate.

7 Conclusion

We study the problem of estimating a high dimensional covariance matrix with conditional sparsity. Realizing unconditional sparsity assumption is inappropriate in many appli-

Figure 7: Heatmap of thresholded error correlation matrix for number of factors $K = 1$, $K = 2$ and $K = 3$.



cations, we introduce a 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 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) and Fan et al. (2011), which arises from the effect of estimating the unobservable factors. As we can see, this effect vanishes as the dimensionality increases, as there are more information available about the common factors. 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. This fact was also shown in the asymptotic expansions obtained by Bai (2003).

The sparse covariance is estimated using the adaptive hard thresholding. Recently, Cai and Liu (2011) studied a more general thresholding function of Antoniadis and Fan (2001), which admits the form $\widehat{\sigma}_{ij}(\theta_{ij}) = s(\sigma_{ij})$, and also allows for soft-thresholding. It is easy to apply the more general thresholding here as well, and the rate of convergence of the resulting covariance matrix estimators should be straightforward to derive.

A Proofs for Section 2

Proof of Theorem 2.1

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

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

$$= \frac{1}{T}\mathbf{Y}\mathbf{Y}' - \widehat{\mathbf{\Lambda}}\widehat{\mathbf{\Lambda}}'.$$

where we used the normalization condition $\frac{1}{T}\widehat{\mathbf{F}}'\widehat{\mathbf{F}} = \mathbf{I}_K$.

If we show that $\widehat{\mathbf{\Lambda}}\widehat{\mathbf{\Lambda}}' = \sum_{i=1}^{K} \widehat{\lambda}_i \widehat{\boldsymbol{\xi}}_i \widehat{\boldsymbol{\xi}}_i'$, then from the decompositions of the sample covariance

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

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

We now show $\widehat{\mathbf{\Lambda}}\widehat{\mathbf{\Lambda}}' = \sum_{i=1}^{K} \widehat{\lambda}_i \widehat{\boldsymbol{\xi}}_i \widehat{\boldsymbol{\xi}}_i'$ indeed holds. Consider again the least squares problem (2.10) but with the following alternative normalization constraints:

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

Let $(\widetilde{\mathbf{\Lambda}}, \widetilde{\mathbf{F}})$ be the solution to the new optimization problem. Switching the roles of $\mathbf{B}$ and $\mathbf{F}$, then the solution of (2.12) is $\widetilde{\mathbf{\Lambda}} = (\hat{\boldsymbol{\xi}}_1, \cdots, \hat{\boldsymbol{\xi}}_K)$ and $\widetilde{\mathbf{F}} = p^{-1}\mathbf{Y}'\widetilde{\mathbf{\Lambda}}$. In addition,

$$T^{-1}\widetilde{\mathbf{F}}'\widetilde{\mathbf{F}} = \text{diag}(\widehat{\lambda}_1, \cdots, \widehat{\lambda}_K).$$

From $\widehat{\mathbf{\Lambda}}\widehat{\mathbf{F}} = \widetilde{\mathbf{\Lambda}}\widetilde{\mathbf{F}}$, it follows that

$$\begin{aligned}
\widehat{\mathbf{\Lambda}}\widehat{\mathbf{\Lambda}}' &= \frac{1}{T}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{F}}'\widehat{\mathbf{F}}\widehat{\mathbf{\Lambda}} \\
&= \frac{1}{T}\widetilde{\mathbf{\Lambda}}\widetilde{\mathbf{F}}'\widetilde{\mathbf{F}}\widetilde{\mathbf{\Lambda}}' \\
&= \sum_{i=1}^{K} \widehat{\lambda}_i \widehat{\boldsymbol{\xi}}_i \widehat{\boldsymbol{\xi}}_i'.
\end{aligned}$$

Q.E.D.

B Proofs for Section 3

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

B.1 Preliminary lemmas

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

Lemma B.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*

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

Lemma B.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* $s > 0$,

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

Lemma B.3. *Under the assumptions of Theorem 3.1,*
(i) $\max_{i,j\leq K} |\frac{1}{T}\sum_{t=1}^T f_{it}f_{jt} - Ef_{it}f_{jt}| = O_p(\sqrt{(\log K)/T})$.
(ii) $\max_{i,j\leq p} |\frac{1}{T}\sum_{t=1}^T u_{it}u_{jt} - Eu_{it}u_{jt}| = O_p(\sqrt{(\log p)/T})$
(iii) $\max_{i\leq K,j\leq p} |\frac{1}{T}\sum_{t=1}^T f_{it}u_{jt}| = O_p(\sqrt{(\log p)/T})$

Proof. The proof follows from the Bernstein inequality for weakly dependent data (Merlevède et al. (2009, Theorem 1). See Lemmas A.3 and B.1 of Fan, Liao and Mincheva (2011).

Lemma B.4. *Let* $\widehat{\lambda}_K$ *denote the* K*th largest eigenvalue of* $\widehat{\mathbf{\Sigma}} = \frac{1}{T}\sum_{t=1}^T \mathbf{y}_t\mathbf{y}_t'$, *then* $\widehat{\lambda}_K > C_1 p$ *with probability approaching one for some* $C_1 > 0$.

Proof. First of all, by Proposition 2.1, under Assumption 3.5, the Kth largest eigenvalue λ_K of $\mathbf{\Sigma}$ satisfies:

$$\begin{aligned} \lambda_K &\geq \|\mathbf{b}_K\| - |\lambda_K - \|\mathbf{b}_K\|| \geq p\lambda_{\min}(\Omega) - \|\mathbf{\Sigma}_u\| \\ &\geq p\lambda_{\min}(\Omega)/2 \end{aligned}$$

for sufficiently large p. Using Weyl's theorem, we need only to prove that $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\| = o_p(p)$. Without loss of generality, we prove the result under the identifiability condition (2.7). Using model (1.2),

$$\widehat{\mathbf{\Sigma}} = T^{-1}\sum_{t=1}^T (\mathbf{B}\mathbf{f}_t + \mathbf{u}_t)(\mathbf{B}\mathbf{f}_t + \mathbf{u}_t)'.$$

Using this and (1.3), $\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}$ can be decomposed as the sum of the four terms:

$$\begin{aligned}
\mathbf{D}_1 &= (T^{-1}\mathbf{B}\sum_{t=1}^{T}\mathbf{f}_t\mathbf{f}_t' - \mathbf{I}_K)\mathbf{B}', \\
\mathbf{D}_2 &= T^{-1}\sum_{t=1}^{T}(\mathbf{u}_t\mathbf{u}_t' - \mathbf{\Sigma}_u), \\
\mathbf{D}_3 &= \mathbf{B}T^{-1}\sum_{t=1}^{T}\mathbf{f}_t\mathbf{u}_t', \qquad \mathbf{D}_4 = \mathbf{D}_3'
\end{aligned}$$

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

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

First of all, by Lemma B.3,

$$\|T^{-1}\sum_{t=1}^{T}\mathbf{f}_t\mathbf{f}_t' - \mathbf{I}_K\| \le K\|T^{-1}\sum_{t=1}^{T}\mathbf{f}_t\mathbf{f}_t' - \mathbf{I}_K\|_{\max} = O_p(K\sqrt{(\log K)/T}),$$

which is $o_p(p)$ if $K\log p = o(T)$. Consequently, by Assumption 3.5, we have

$$\|\mathbf{D}_1\| \le O_p(K\sqrt{(\log K)/T})\|\mathbf{B}\mathbf{B}'\| = O_p(Kp\sqrt{(\log K)/T}).$$

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

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

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

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

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

Q.E.D.

Lemma B.5 (Theorem 2.1, Fan, Liao and Mincheva, 2011)**.** *Suppose there exists a positive sequence* $a_T \to 0$ *such that* $\max_{i\le p}\frac{1}{T}\sum_{t=1}^{T}(u_{it} - \hat{u}_{it})^2 = O_p(a_T^2)$*. In addition, assume* $\max_{i,t}|u_{it} - \hat{u}_{it}| = o_p(1)$*, and* $(\log p)^{6/\gamma-1} = o(T)$*. Then under Assumption 3.1,* $\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}}$ *defined*

as in (2.13) with $\omega_T = C(\sqrt{(\log p)/T} + a_T)$ *for sufficiently large* $C > 0$ *satisfies: (i)*

$$\|\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}} - \mathbf{\Sigma}_u\| = O_p(m_T\omega_T).$$

(ii) If $m_T\omega_T = o(1)$*, then* $\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}}$ *is positive definite with probability approaching one, and*

$$\|(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1} - \mathbf{\Sigma}_u^{-1}\| = O_p(m_T\omega_T).$$

B.2 Proof of Theorem 3.3

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

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

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 $\widehat{\mathbf{f}}_t = (\hat{f}_{1t}, ..., \hat{f}_{Kt})'$.

Lemma B.6. *(i)* $\max_{i\le K}\frac{1}{T}\sum_{t=1}^{T}(\frac{1}{T}\sum_{s=1}^{T}\hat{f}_{is}E(\mathbf{u}_s'\mathbf{u}_t)/p)^2 = O_p(T^{-1})$,
(ii) $\max_{i\le K}\frac{1}{T}\sum_{t=1}^{T}(\frac{1}{T}\sum_{s=1}^{T}\hat{f}_{is}\zeta_{st})^2 = O_p(p^{-1})$,
(iii) $\max_{i\le K}\frac{1}{T}\sum_{t=1}^{T}(\frac{1}{T}\sum_{s=1}^{T}\hat{f}_{is}\eta_{st})^2 = O_p(K^2p^{-1})$,
(iv) $\max_{i\le K}\frac{1}{T}\sum_{t=1}^{T}(\frac{1}{T}\sum_{s=1}^{T}\hat{f}_{is}\xi_{st})^2 = O_p(K^2p^{-1})$.

Proof. (i) We have $\forall i \le K$, $\sum_{s=1}^{T}\hat{f}_{is}^2 = T$. By the Cauchy-Schwarz inequality,

$$\begin{aligned}
\max_{i\le K}\frac{1}{T}\sum_{t=1}^{T}(\frac{1}{T}\sum_{s=1}^{T}\hat{f}_{is}E(\mathbf{u}_s'\mathbf{u}_t)/p)^2 &\le \frac{1}{T}\sum_{t=1}^{T}\frac{1}{T}\sum_{s=1}^{T}(E\mathbf{u}_s'\mathbf{u}_t/p)^2 \\
&\le \max_{t\le T}\frac{1}{T}\sum_{s=1}^{T}(E\mathbf{u}_s'\mathbf{u}_t/p)^2 \\
&\le \max_{s,t}|E\mathbf{u}_s'\mathbf{u}_t/p|\max_{t\le T}\frac{1}{T}\sum_{s=1}^{T}|E\mathbf{u}_s'\mathbf{u}_t/p|
\end{aligned}$$

By Assumption 3.3, $\max_{t\le T}\sum_{s=1}^{T}|E\mathbf{u}_s'\mathbf{u}_t/p| = O(1)$, which then yields the result.

(ii) By the Cauchy-Schwarz inequality,

$$\begin{aligned}
\max_{i\le K}\frac{1}{T}\sum_{t=1}^{T}(\frac{1}{T}\sum_{s=1}^{T}\hat{f}_{is}\zeta_{st})^2 &= \max_{i}\frac{1}{T^3}\sum_{s=1}^{T}\sum_{l=1}^{T}\hat{f}_{is}\hat{f}_{il}(\sum_{t=1}^{T}\zeta_{st}\zeta_{lt}) \\
&\le \max_{i}\frac{1}{T^3}\sqrt{\sum_{sl}(\hat{f}_{is}\hat{f}_{il})^2\sum_{sl}(\sum_{t=1}^{T}\zeta_{st}\zeta_{lt})^2}
\end{aligned}$$

$$\leq \max_i \frac{1}{T^3}\sum_{s=1}^{T}\hat{f}_{is}^2\sqrt{\sum_{sl}(\sum_{t=1}^{T}\zeta_{st}\zeta_{lt})^2}$$
$$= \frac{1}{T^2}\sqrt{\sum_{s=1}^{T}\sum_{l=1}^{T}(\sum_{t=1}^{T}\zeta_{st}\zeta_{lt})^2}.$$

Note that $E(\sum_{s=1}^{T}\sum_{l=1}^{T}(\sum_{t=1}^{T}\zeta_{st}\zeta_{lt})^2) = T^2E(\sum_{t=1}^{T}\zeta_{st}\zeta_{lt})^2 \leq T^4\max_{st}E|\zeta_{st}|^4$. 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

$$E\frac{1}{T}\sum_{t=1}^{T}\|\sum_{i=1}^{p}\mathbf{b}_i u_{it}\|^2 = E\|\sum_{i=1}^{p}\mathbf{b}_i u_{it}\|^2 = O(pK).$$

Therefore, by the Cauchy-Schwarz inequality,

$$\begin{aligned}\max_{i\leq K}\frac{1}{T}\sum_{t=1}^{T}(\frac{1}{T}\sum_{s=1}^{T}\hat{f}_{is}\eta_{st})^2 &\leq \max_i\|\frac{1}{T}\sum_{s=1}^{T}\hat{f}_{is}\mathbf{f}_s'\|^2\frac{1}{T}\sum_{t=1}^{T}\|\sum_{j=1}^{p}\mathbf{b}_j u_{jt}\frac{1}{p}\|^2\\ &\leq \max_i\frac{1}{Tp^2}\sum_{t=1}^{T}\|\sum_{j=1}^{p}\mathbf{b}_j u_{jt}\|^2\left(\frac{1}{T}\sum_{s=1}^{T}\hat{f}_{is}^2\frac{1}{T}\sum_{s=1}^{T}\|\mathbf{f}_s\|^2\right)\\ &= O_p\left(\frac{K^2}{p}\right).\end{aligned}$$

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

$$\begin{aligned}\max_{i\leq K}\frac{1}{T}\sum_{t=1}^{T}(\frac{1}{T}\sum_{s=1}^{T}\hat{f}_{is}\xi_{st})^2 &= \max_i\frac{1}{T}\sum_{t=1}^{T}\left|\frac{1}{T}\sum_{s=1}^{T}\mathbf{f}_t'\sum_{j=1}^{p}\mathbf{b}_j u_{js}\frac{1}{p}\hat{f}_{is}\right|^2\\ &\leq \max_i\frac{1}{T}\sum_{t=1}^{T}\|\mathbf{f}_t\|^2\cdot\left\|\frac{1}{T}\sum_{s=1}^{T}\sum_{j=1}^{p}\mathbf{b}_j u_{js}\frac{1}{p}\hat{f}_{is}\right\|^2\\ &\leq O_p(K)\max_i\frac{1}{T}\sum_{s=1}^{T}\left\|\sum_{j=1}^{p}\mathbf{b}_j u_{js}\frac{1}{p}\right\|^2\cdot\frac{1}{T}\sum_{s=1}^{T}\hat{f}_{is}^2\\ &\leq O_p\left(\frac{K^2}{p}\right),\end{aligned}$$

where the third line follows from the Cauchy-Schwarz inequality. Q.E.D.

Lemma B.7. *(i)* $\max_{t\leq T}\|\frac{1}{Tp}\sum_{s=1}^{T}\widehat{\mathbf{f}}_s E(\mathbf{u}_s'\mathbf{u}_t)\| = O_p(\sqrt{K/T})$,
(ii) $\max_{t\leq T}\|\frac{1}{T}\sum_{s=1}^{T}\widehat{\mathbf{f}}_s\zeta_{st}\| = O_p(\sqrt{K}T^{1/4}/\sqrt{p})$,
(iii) $\max_{t\leq T}\|\frac{1}{T}\sum_{s=1}^{T}\widehat{\mathbf{f}}_s\eta_{st}\| = O_p(K^{3/2}T^{1/4}/\sqrt{p})$,

(iv) $\max_{t\leq T}\|\frac{1}{T}\sum_{s=1}^{T}\widehat{\mathbf{f}}_s\xi_{st}\| = O_p(K^{3/2}T^{1/4}/\sqrt{p})$.

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

$$\begin{aligned}\max_{t\leq T}\|\frac{1}{Tp}\sum_{s=1}^{T}\widehat{\mathbf{f}}_s E(\mathbf{u}_s'\mathbf{u}_t)\| &\leq \max_{t\leq T}\sqrt{\frac{1}{T}\sum_{s=1}^{T}\|\widehat{\mathbf{f}}_s\|^2\frac{1}{T}\sum_{s=1}^{T}(E\mathbf{u}_s'\mathbf{u}_t/p)^2}\\ &\leq \sqrt{K}\max_{t\leq T}\sqrt{\frac{1}{T}\sum_{s=1}^{T}(E\mathbf{u}_s'\mathbf{u}_t/p)^2}\\ &\leq \sqrt{K}\max_{s,t}\sqrt{|E\mathbf{u}_s'\mathbf{u}_t/p|}\max_{t\leq T}\sqrt{\frac{1}{T}\sum_{s=1}^{T}|E\mathbf{u}_s'\mathbf{u}_t/p|}.\end{aligned}$$

The result then follows from Assumption 3.3.

(ii) By the Cauchy-Schwarz inequality,

$$\max_{t\leq T}\|\frac{1}{T}\sum_{s=1}^{T}\widehat{\mathbf{f}}_s\zeta_{st}\| \leq \max_{t\leq T}\frac{1}{T}\sqrt{\sum_{s=1}^{T}\|\widehat{\mathbf{f}}_s\|^2\sum_{s=1}^{T}\zeta_{st}^2} \leq \sqrt{K\max_t\frac{1}{T}\sum_{s=1}^{T}\zeta_{st}^2}.$$

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

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

$$\max_{t\leq T}\left\|\frac{1}{T}\sum_{s=1}^{T}\widehat{\mathbf{f}}_s\eta_{st}\right\| \leq \left\|\frac{1}{T}\sum_{s=1}^{T}\widehat{\mathbf{f}}_s\mathbf{f}_s'\right\|\max_t\left\|\frac{1}{p}\sum_{i=1}^{p}\mathbf{b}_i u_{it}\right\| = o_p\left(\frac{K^{3/2}T^{1/4}}{\sqrt{p}}\right).$$

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

$$\left\|\frac{1}{T}\sum_{s=1}^{T}\sum_{i=1}^{p}\mathbf{b}_i u_{is}\frac{1}{p}\widehat{\mathbf{f}}_s\right\| = O_p\left(\frac{K}{\sqrt{p}}\right).$$

In addition, since $E\|K^{-2}\mathbf{f}_t\|^4 < M$, $\max_{t\leq T}\|\mathbf{f}_t\| = O_p(T^{1/4}K^{1/2})$. It follows that

$$\max_{t\leq T}\|\frac{1}{T}\sum_{s=1}^{T}\widehat{\mathbf{f}}_s\xi_{st}\| \leq \max_{t\leq T}\|\mathbf{f}_t\| \cdot \left\|\frac{1}{T}\sum_{s=1}^{T}\sum_{i=1}^{p}\mathbf{b}_i u_{is}\frac{1}{p}\widehat{\mathbf{f}}_s\right\| = O_p(\frac{K^{3/2}T^{1/4}}{\sqrt{p}}).$$

Q.E.D.

Lemma B.8. *(i)* $\max_{i\leq K}\frac{1}{T}\sum_{t=1}^{T}(\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t)_i^2 = O_p(1/T + K^2/p)$.
(ii) $\frac{1}{T}\sum_{t=1}^{T}\|\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t\|^2 = O_p(K/T + K^3/p)$.
(iii) $\max_{t\leq T}\|\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t\| = O_p(\sqrt{K/T} + K^{3/2}T^{1/4}/\sqrt{p})$.

Proof. (i) By (B.2), the assumption that the entries of $\mathbf{V}$ are bounded away from zero, and the fact that $(a+b+c+d)^2 \leq 4(a^2+b^2+c^2+d^2)$, there exists a constant $C > 0$,

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

Each of the four terms on the right hand side above are bounded in Lemma B.6, which then yields the desired result.

(ii) It follows from part (i) and that

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

Part (iii) is implied by (B.2) and Lemma B.7. Q.E.D.

Lemma B.9. *There exist positive constants* c_1, c_2 *such that with probability approaching one,* $c_1 < \lambda_{\min}(\mathbf{H}'\mathbf{H}) < \lambda_{\max}(\mathbf{H}'\mathbf{H}) < c_2$.

Proof. We first show that $\|\mathbf{H}\| = O_p(1)$. In fact, $\lambda_{\max}(\mathbf{V}) = \lambda_{\max}(\frac{1}{pT}\sum_{t=1}^{T}\mathbf{y}_t\mathbf{y}_t') = O_p(1)$, and

$$\|\mathbf{F}\| = \lambda_{\max}^{1/2}(\mathbf{F}\mathbf{F}') = \lambda_{\max}^{1/2}(\sum_{t=1}^{T}\mathbf{f}_t\mathbf{f}_t') = O_p(\sqrt{T}).$$

In addition, $\|\widehat{\mathbf{F}}\| = \sqrt{T}$. It then follows from the definition of $\mathbf{H}$ that $\|\mathbf{H}\| = O_p(1)$.

Now

$$\begin{aligned}\widehat{\mathbf{F}}'\mathbf{F}\mathbf{H}/T &= (\widehat{\mathbf{F}}-\mathbf{F}\mathbf{H})'\mathbf{F}\mathbf{H}/T+\mathbf{H}'\mathbf{F}'\mathbf{F}\mathbf{H}/T \\ &= (\widehat{\mathbf{F}}-\mathbf{F}\mathbf{H})'\mathbf{F}\mathbf{H}/T+\mathbf{H}'\mathbf{H}+\mathbf{H}'(\mathbf{F}'\mathbf{F}/T-\mathbf{I}_K)\mathbf{H}. \qquad \text{(B.3)}\end{aligned}$$

On the other hand, using $\widehat{\mathbf{F}}'\widehat{\mathbf{F}}/T=\mathbf{I}_K$, we have

$$\begin{aligned}\widehat{\mathbf{F}}'\mathbf{F}\mathbf{H}/T &= \widehat{\mathbf{F}}'(\mathbf{F}\mathbf{H}-\widehat{\mathbf{F}})/T+\widehat{\mathbf{F}}'\widehat{\mathbf{F}}/T \\ &= \widehat{\mathbf{F}}'(\mathbf{F}\mathbf{H}-\widehat{\mathbf{F}})/T+\mathbf{I}_K. \qquad \text{(B.4)}\end{aligned}$$

Thus combining (B.3) and (B.4) gives us

$$\mathbf{H}'\mathbf{H}=\mathbf{I}_K+\mathbf{C}, \qquad \text{(B.5)}$$

where

$$\mathbf{C}=\widehat{\mathbf{F}}'(\mathbf{F}\mathbf{H}-\widehat{\mathbf{F}})/T-(\widehat{\mathbf{F}}-\mathbf{F}\mathbf{H})'\mathbf{F}\mathbf{H}/T-\mathbf{H}'(\mathbf{F}'\mathbf{F}/T-\mathbf{I}_K)\mathbf{H}.$$

It follows from Lemma B.8 and the triangular inequality that

$$\begin{aligned}\|\mathbf{C}\| &\leq \|\mathbf{F}\mathbf{H}-\widehat{\mathbf{F}}\|\|\mathbf{F}\|/T(1+\|\mathbf{H}\|)+\|\mathbf{H}\|^2\|\widehat{\mathbf{F}}'\widehat{\mathbf{F}}/T-\mathbf{I}_K\| \\ &\leq \sqrt{\frac{1}{T}\sum_{t=1}^{T}\|\mathbf{H}\mathbf{f}_t-\widehat{\mathbf{f}}_t\|^2}\frac{\|\mathbf{F}\|}{\sqrt{T}}O_p(1)+O_p(1)\|\widehat{\mathbf{F}}'\widehat{\mathbf{F}}/T-\mathbf{I}_K\| \\ &= O_p(\sqrt{\frac{K^3}{p}}+\sqrt{\frac{K^2\log K}{T}})=o_p(1).\end{aligned}$$

It then follows from Weyl's Theorem that $\lambda_{\min}(\mathbf{H}'\mathbf{H})>1/2$ with probability approaching one.

Q.E.D.

Proof of Theorem 3.3

The second part of this theorem was proved in Lemma B.8. We now derive the convergence rate of $\max_{i\leq p}\|\widehat{\mathbf{b}}_i-(\mathbf{H}')^{-1}\mathbf{b}_i\|$.

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

$$\widehat{\mathbf{b}}_i-(\mathbf{H}')^{-1}\mathbf{b}_i=\frac{1}{T}\sum_{t=1}^{T}\mathbf{H}\mathbf{f}_tu_{it}+\frac{1}{T}\sum_{t=1}^{T}\widehat{\mathbf{f}}_t(\mathbf{f}_t-\mathbf{H}^{-1}\widehat{\mathbf{f}}_t)'\mathbf{b}_i+\frac{1}{T}\sum_{t=1}^{T}(\widehat{\mathbf{f}}_t-\mathbf{H}\mathbf{f}_t)u_{it} \qquad \text{(B.6)}$$

We bound the three terms on the right hand side respectively. It follows from Lemmas B.3

and B.9 that

$$\max_{i\leq p}\left\|\frac{1}{T}\sum_{t=1}^{T}\mathbf{H}\mathbf{f}_t u_{it}\right\| \leq \|\mathbf{H}\|\max_i\sqrt{\sum_{k=1}^{K}(\frac{1}{T}\sum_{t=1}^{T}f_{kt}u_{it})^2} = O_p\left(\sqrt{\frac{K\log p}{T}}\right). \tag{B.7}$$

For the second term, $\max_i\|(\mathbf{H}')^{-1}\mathbf{b}_i\| = O(\sqrt{K})$. Therefore, the Cauchy-Schwarz inequality and Lemma B.8 imply

$$\begin{aligned}\max_{i\leq p}\|\frac{1}{T}\sum_{t=1}^{T}\widehat{\mathbf{f}}_t(\mathbf{f}_t-\mathbf{H}^{-1}\widehat{\mathbf{f}}_t)'\mathbf{b}_i\| &= \max_{i\leq p}\|\frac{1}{T}\sum_{t=1}^{T}\widehat{\mathbf{f}}_t(\mathbf{H}\mathbf{f}_t-\widehat{\mathbf{f}}_t)'(\mathbf{H}')^{-1}\mathbf{b}_i\| \\ &\leq \max_i\|(\mathbf{H}')^{-1}\mathbf{b}_i\|\sqrt{\frac{1}{T}\sum_{t=1}^{T}\|\widehat{\mathbf{f}}_t\|^2\frac{1}{T}\sum_{t=1}^{T}\|\mathbf{H}\mathbf{f}_t-\widehat{\mathbf{f}}_t\|^2} \\ &= O_p\left(\frac{K\sqrt{K}}{\sqrt{T}}+\frac{K^2\sqrt{K}}{\sqrt{p}}\right). \end{aligned}\tag{B.8}$$

Finally, as $\max_i\frac{1}{T}\sum_{t=1}^{T}u_{it}^2 \leq \max_i|\frac{1}{T}\sum_{t=1}^{T}u_{it}^2 - Eu_{it}^2| + \max_i Eu_{it}^2 < M$. Still by the Cauchy-Schwarz inequality,

$$\begin{aligned}\max_i\|\frac{1}{T}\sum_{t=1}^{T}(\widehat{\mathbf{f}}_t-\mathbf{H}\mathbf{f}_t)u_{it}\| &\leq \sqrt{\frac{1}{T}\sum_{t=1}^{T}\|\mathbf{H}\mathbf{f}_t-\widehat{\mathbf{f}}_t\|^2\max_i\frac{1}{T}\sum_t u_{it}^2} \\ &= O_p(\frac{\sqrt{K}}{\sqrt{T}}+\frac{K^{3/2}}{\sqrt{p}}). \end{aligned}\tag{B.9}$$

The result follows from combining (B.6)-(B.9). Q.E.D.

Proof of Corollary 3.1

Under Assumption 3.3, it can be shown by Bonferroni's method that

$$\max_{t\leq T}\|\mathbf{f}_t\| = O_p(\sqrt{K}(\log T)^{1/r_2}). \tag{B.10}$$

By Theorem 3.3, and $\max\{\|\mathbf{H}\|,\|\mathbf{H}^{-1}\|\} = O_p(1)$, $\forall i,t$,

$$\begin{aligned}\|\widehat{\mathbf{b}}_i'\widehat{\mathbf{f}}_t-\mathbf{b}_i'\mathbf{f}_t\| &\leq \|\widehat{\mathbf{b}}_i-(\mathbf{H}')^{-1}\mathbf{b}_i\|\|\widehat{\mathbf{f}}_t-\mathbf{H}\mathbf{f}_t\|+\|(\mathbf{H}')^{-1}\mathbf{b}_i\|\|\widehat{\mathbf{f}}_t-\mathbf{H}\mathbf{f}_t\| \\ &\quad +\|\widehat{\mathbf{b}}_i-(\mathbf{H}')^{-1}\mathbf{b}_i\|\|\mathbf{H}\mathbf{f}_t\| \\ &= O_p\left(\frac{\delta_T}{\sqrt{K}m_p}\delta_T^*\right)+O_p\left(\sqrt{K}\delta_T^*\right)+O_p\left(\frac{\delta_T}{\sqrt{K}m_p}\sqrt{K}(\log T)^{1/r_2}\right) \\ &= O_p\left(\sqrt{K}\delta_T^*+\frac{\delta_T(\log T)^{1/r_2}}{m_p}\right). \end{aligned}$$

B.3 Proof of Theorem 3.1

Lemma B.10.

$$\max_{i\le p}\frac{1}{T}\sum_{t=1}^{T}|u_{it}-\hat{u}_{it}|^2=O_p\left(\frac{K^2\log p+K^4}{T}+\frac{K^6}{p}\right),$$

$$\max_{i,t}|u_{it}-\hat{u}_{it}|=O_p\left(\sqrt{K}\delta_T^*+\frac{\delta_T(\log T)^{1/r_2}}{m_p}\right)=o_p(1).$$

Proof. We have,

$$u_{it}-\hat{u}_{it}=\mathbf{b}_i'\mathbf{H}^{-1}(\widehat{\mathbf{f}}_t-\mathbf{H}\mathbf{f}_t)+(\widehat{\mathbf{b}}_i'-\mathbf{b}_i'\mathbf{H}^{-1})(\widehat{\mathbf{f}}_t-\mathbf{H}\mathbf{f}_t)+(\widehat{\mathbf{b}}_i'-\mathbf{b}_i'\mathbf{H}^{-1})\mathbf{H}\mathbf{f}_t. \tag{B.11}$$

Therefore, using the inequality $(a+b+c)^2\le 4a^2+4b^2+4c^2$, we have:

$$\begin{aligned}
\max_{i\le T}\frac{1}{T}\sum_{t=1}^{T}(u_{it}-\hat{u}_{it})^2 \;\le\;& 4\max_{i\le T}\mathbf{b}_i'\mathbf{H}^{-1}\frac{1}{T}\sum_{t=1}^{T}(\widehat{\mathbf{f}}_t-\mathbf{H}\mathbf{f}_t)(\widehat{\mathbf{f}}_t-\mathbf{H}\mathbf{f}_t)'(\mathbf{H}')^{-1}\mathbf{b}_i\\
&+4\max_{i\le T}(\widehat{\mathbf{b}}_i'-\mathbf{b}_i'\mathbf{H}^{-1})\frac{1}{T}\sum_{t=1}^{T}(\widehat{\mathbf{f}}_t-\mathbf{H}\mathbf{f}_t)(\widehat{\mathbf{f}}_t-\mathbf{H}\mathbf{f}_t)'(\widehat{\mathbf{b}}_i'-\mathbf{b}_i'\mathbf{H}^{-1})'\\
&+4\max_{i\le T}(\widehat{\mathbf{b}}_i'-\mathbf{b}_i'\mathbf{H}^{-1})\mathbf{H}\frac{1}{T}\sum_{t=1}^{T}\mathbf{f}_t\mathbf{f}_t'\mathbf{H}'(\widehat{\mathbf{b}}_i'-\mathbf{b}_i'\mathbf{H}^{-1})'\\
\le\;& 4\max_i\|\mathbf{b}_i'\mathbf{H}^{-1}\|^2\frac{1}{T}\sum_{t=1}^{T}\|\widehat{\mathbf{f}}_t-\mathbf{H}\mathbf{f}_t\|^2\\
&+4\max_i\|\widehat{\mathbf{b}}_i'-\mathbf{b}_i'\mathbf{H}^{-1}\|^2\frac{1}{T}\sum_{t=1}^{T}\|\widehat{\mathbf{f}}_t-\mathbf{H}\mathbf{f}_t\|^2\\
&+4\max_i\|\widehat{\mathbf{b}}_i'-\mathbf{b}_i'\mathbf{H}^{-1}\|^2\frac{1}{T}\sum_{t=1}^{T}\|\mathbf{H}\mathbf{f}_t\mathbf{f}_t'\mathbf{H}'\|_F,
\end{aligned}$$

It then follows from Theorem 3.3 and Lemma B.8 that

$$\max_{i\le T}\frac{1}{T}\sum_{t=1}^{T}(u_{it}-\hat{u}_{it})^2=O_p(K^6/p+(K^2\log p+K^4)/T).$$

By (B.10), $\max_{t\le T}\|\mathbf{f}_t\|=O_p(\sqrt{K}(\log T)^{1/r_2})$. Hence part (ii) follows from Theorem 3.3. Q.E.D.

Proof of Theorem 3.1 The theorem follows immediately from Lemma B.5 and Lemma B.10.

B.4 Proof of Theorem 3.2

Define $\mathbf{D}_T$ and $\mathbf{C}_T$ as:

$$\mathbf{D}_T = \mathbf{I}_K - \mathbf{H}\mathbf{H}', \qquad \mathbf{C}_T = \widehat{\mathbf{\Lambda}} - \mathbf{B}\mathbf{H}^{-1}.$$

Lemma B.11. $\|\mathbf{D}_T\|_F = O_p(K\sqrt{\log K}/\sqrt{T} + K^2/\sqrt{p})$.

Proof. Applying the triangular inequality gives:

$$\|\mathbf{D}_T\|_F \leq \|\mathbf{H}\mathbf{H}' - \widehat{\mathrm{cov}}(\mathbf{H}\mathbf{f}_t)\|_F + \|\widehat{\mathrm{cov}}(\mathbf{H}\mathbf{f}_t) - \mathbf{I}_K\|_F \tag{B.12}$$

By Lemmas B.3 and B.9, the first term in (B.12) is

$$\|\mathbf{H}\mathbf{H}' - \widehat{\mathrm{cov}}(\mathbf{H}\mathbf{f}_t)\|_F \leq \|\mathbf{H}\|^2\|\mathbf{I}_K - \widehat{\mathrm{cov}}(\mathbf{f}_t)\|_F = O_p\left(K\sqrt{\frac{\log K}{T}}\right).$$

The second term of (B.12) can be bounded, by the Cauchy-Schwarz inequalities and Lemma B.8, as follows:

$$\begin{aligned}
& \left\|\frac{1}{T}\sum_{t=1}^{T}\mathbf{H}\mathbf{f}_t(\mathbf{H}\mathbf{f}_t)' - \frac{1}{T}\sum_{t=1}^{T}\widehat{\mathbf{f}}_t\widehat{\mathbf{f}}_t'\right\|_F \\
\leq & \left\|\frac{1}{T}\sum_{t}(\mathbf{H}\mathbf{f}_t - \widehat{\mathbf{f}}_t)(\mathbf{H}\mathbf{f}_t)'\right\|_F + \left\|\frac{1}{T}\sum_{t}\widehat{\mathbf{f}}_t(\widehat{\mathbf{f}}_t' - (\mathbf{H}\mathbf{f}_t)')\right\|_F \\
\leq & \sqrt{\frac{1}{T}\sum_{t}\|\mathbf{H}\mathbf{f}_t - \widehat{\mathbf{f}}_t\|^2\frac{1}{T}\sum_{t}\|\mathbf{H}\mathbf{f}_t\|^2} + \sqrt{\frac{1}{T}\sum_{t}\|\mathbf{H}\mathbf{f}_t - \widehat{\mathbf{f}}_t\|^2\frac{1}{T}\sum_{t}\|\widehat{\mathbf{f}}_t\|^2} \\
= & O_p\left(\frac{K}{\sqrt{T}} + \frac{K^2}{\sqrt{p}}\right).
\end{aligned}$$

Q.E.D.

Lemma B.12. $\|\mathbf{B}\mathbf{H}^{-1}\mathbf{D}_T(\mathbf{B}\mathbf{H}^{-1})'\|_\Sigma^2 = O(K^2(\log K)/(pT) + K^4/p^2)$.

Proof. First of all, note that

$$\lambda_{\max}((\mathbf{H}\mathrm{cov}(\mathbf{f}_t)\mathbf{H}')^{-1}) = \lambda_{\max}((\mathbf{H}\mathbf{H}')^{-1}),$$

which is bounded away from infinity by Lemma B.9. Hence the same argument of the proof of Theorem 2 in Fan, Fan and Lv (2008), with $\mathbf{B}$ and $\mathbf{f}$ replaced with $\mathbf{B}\mathbf{H}^{-1}$ and $\mathbf{H}\mathbf{f}_t$, yields

$$\|(\mathbf{B}\mathbf{H}^{-1})'\mathbf{\Sigma}^{-1}\mathbf{B}\mathbf{H}^{-1}\| \leq 2\|(\mathbf{H}\mathrm{cov}(\mathbf{f}_t)\mathbf{H}')^{-1}\| = O_p(1).$$

The lemma then follows from Lemma B.11 and the fact that

$$\begin{aligned}\|\mathbf{B}\mathbf{H}^{-1}\mathbf{D}_T(\mathbf{B}\mathbf{H}^{-1})'\|_\Sigma^2 &\leq p^{-1}\text{tr}[(\mathbf{D}_T(\mathbf{B}\mathbf{H}^{-1})'\mathbf{\Sigma}^{-1}\mathbf{B}\mathbf{H}^{-1})^2] \\ &\leq p^{-1}\|\mathbf{D}_T\|_F^2\|(\mathbf{B}\mathbf{H}^{-1})'\mathbf{\Sigma}^{-1}\mathbf{B}\mathbf{H}^{-1}\|^2.\end{aligned}$$

Q.E.D.

Lemma B.13. *Let* $\mathbf{E} = (\mathbf{u}_1, ..., \mathbf{u}_T)$.
(i) $\|T^{-1}\mathbf{E}\mathbf{F}\mathbf{H}'\|_F^2 = O_p(pK(\log p)/T)$.
(ii) $\|T^{-1}\mathbf{E}(\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H}')\|_F^2 = O_p(pK/T + K^3)$.
(iii) $\|\mathbf{B}\mathbf{H}^{-1}\mathbf{C}_T{}'\|_\Sigma^2 = O_p(K^3/p + K(\log p)/T)$.

Proof. (i) $\|T^{-1}\mathbf{E}\mathbf{F}\mathbf{H}'\|_F^2 \leq \|T^{-1}\mathbf{E}\mathbf{F}\|_F^2\|\mathbf{H}\|^2$. The result follows from $\|\mathbf{H}\| = O_p(1)$ and that

$$\|T^{-1}\mathbf{E}\mathbf{F}\|_F^2 \leq pK \max_{i\leq p, j\leq K}(\frac{1}{T}\sum_{t=1}^T f_{jt}u_{it})^2 = O_p(Kp(\log p)/T).$$

(ii) We have, by the Cauchy-Schwarz inequality and Lemma B.8,

$$\begin{aligned}\|\frac{1}{T}\mathbf{E}(\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H}')\|_F^2 &= \|\frac{1}{T}\sum_{t=1}^T \mathbf{u}_t(\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t)'\|_F^2 \\ &\leq (pK)\max_{i\leq p}\frac{1}{T}\sum_{t=1}^T u_{it}^2 \max_{j\leq K}\frac{1}{T}\sum_{t=1}^T(\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t)_j^2 \\ &= O_p(\frac{pK}{T} + K^3). \qquad \text{(B.13)}\end{aligned}$$

(iii) First all, it is easy to see that, for any $K \times p$ matrix $\mathbf{A}$,

$$\begin{aligned}p\|\mathbf{B}\mathbf{H}^{-1}\mathbf{A}\|_\Sigma^2 &= \text{tr}(\mathbf{H}^{-1}\mathbf{A}\mathbf{\Sigma}^{-1}\mathbf{A}'\mathbf{H}'^{-1}\mathbf{B}'\mathbf{\Sigma}^{-1}\mathbf{B}) \\ &\leq \|\mathbf{H}^{-1}\|^2\|\mathbf{B}'\mathbf{\Sigma}^{-1}\mathbf{B}\|\|\mathbf{\Sigma}^{-1}\|\|\mathbf{A}\|_F^2 \\ &= O_p(\|\mathbf{A}\|_F^2). \qquad \text{(B.14)}\end{aligned}$$

In addition, we have the decomposition:

$$\mathbf{C}_T = -\frac{1}{T}\mathbf{E}\mathbf{F}\mathbf{H}' + \frac{1}{T}\mathbf{B}\mathbf{H}^{-1}(\mathbf{H}\mathbf{F}' - \widehat{\mathbf{F}}')\widehat{\mathbf{F}} + \frac{1}{T}\mathbf{E}(\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H}'). \qquad \text{(B.15)}$$

Therefore,

$$p\|\mathbf{B}\mathbf{H}^{-1}(-\frac{1}{T}\mathbf{E}\mathbf{F}\mathbf{H}' + \frac{1}{T}\mathbf{E}(\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H}'))'\|_\Sigma^2 = O_p(\frac{pK\log p}{T} + K^3). \qquad \text{(B.16)}$$

In addition, using $\|\mathbf{B}'\mathbf{\Sigma}^{-1}\mathbf{B}\| = O(1)$, and $\|\widehat{\mathbf{F}}\|^2 = \lambda_{\max}(\sum_{t=1}^T \widehat{\mathbf{f}}_t\widehat{\mathbf{f}}_t') = T$, we obtain

$$\begin{aligned}
&p\|\mathbf{B}\mathbf{H}^{-1}(\frac{1}{T}\mathbf{B}\mathbf{H}^{-1}(\mathbf{H}\mathbf{F}' - \widehat{\mathbf{F}}')\widehat{\mathbf{F}})'\|_\Sigma^2 \\
&\leq \frac{1}{T^2}\|\mathbf{H}^{-1}\|^4\|\mathbf{B}'\mathbf{\Sigma}^{-1}\mathbf{B}\|^2\|\widehat{\mathbf{F}}\|^2\|\mathbf{H}\mathbf{F}' - \widehat{\mathbf{F}}'\|_F^2 \\
&= O_p(\frac{K}{T} + \frac{K^3}{p}).
\end{aligned} \tag{B.17}$$

The result then follows from (B.15)-(B.17). Q.E.D.

Lemma B.14.

$$\|\mathbf{C}_T\mathbf{C}_T{}'\|_\Sigma^2 = O_p(pK^2(\log p)^2/T^2 + K^4(\log p)/T + K^6/p).$$

Proof. Let $\mathbf{A}_1 = -\frac{1}{T}\mathbf{E}\mathbf{F}\mathbf{H}'$, $\mathbf{A}_2 = \frac{1}{T}\mathbf{B}\mathbf{H}^{-1}(\mathbf{H}\mathbf{F}' - \widehat{\mathbf{F}}')\widehat{\mathbf{F}}$, $\mathbf{A}_3 = \frac{1}{T}\mathbf{E}(\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H}')$. By (B.15), $\mathbf{C}_T = \mathbf{A}_1 + \mathbf{A}_2 + \mathbf{A}_3$. In addition,

$$\mathbf{C}_T\mathbf{C}_T{}' = B_{11} + (B_{12} + B_{12}' + B_{13} + B_{13}' + B_{23} + B_{23}') + B_{22} + B_{33},$$

where $B_{ij} = \mathbf{A}_i\mathbf{A}_j'$. We bound each term B_{ij} as follws. Straightforward calculation yields:

$$\begin{aligned}
\|B_{11}\|_\Sigma^2 &\leq p^{-1}\|\mathbf{\Sigma}^{-1}\|^2\left\|\frac{1}{T}\mathbf{E}\mathbf{F}\mathbf{H}'\right\|_F^4 = O_p(\frac{pK^2(\log p)^2}{T^2}) \\
\|B_{13}\|_\Sigma^2 &\leq p^{-1}\|\mathbf{\Sigma}^{-1}\|^2\left\|\frac{1}{T}\mathbf{E}\mathbf{F}\mathbf{H}'\right\|_F^2\left\|\frac{1}{T}\mathbf{E}(\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H}')\right\|_F^2 \\
&= O_p(\frac{pK^2\log p}{T^2} + \frac{K^4\log p}{T}), \\
\|B_{33}\|_\Sigma^2 &\leq p^{-1}\|\mathbf{\Sigma}^{-1}\|^2\left\|\frac{1}{T}\mathbf{E}(\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H}')\right\|_F^4 = O_p(\frac{pK^2}{T^2} + \frac{K^6}{p}).
\end{aligned}$$

By the facts that $\|\mathbf{B}'\mathbf{\Sigma}^{-1}\mathbf{B}\| = O(1)$, and $\|\widehat{\mathbf{F}}\|^2 = T$, we have

$$\begin{aligned}
\|B_{12}\|_\Sigma^2 &\leq p^{-1}\|\mathbf{\Sigma}^{-1}\|\|\mathbf{B}'\mathbf{\Sigma}^{-1}\mathbf{B}\|\left\|\frac{1}{T}\mathbf{E}\mathbf{F}\mathbf{H}'\right\|_F^2\left\|\frac{1}{T}\mathbf{E}(\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H}')\right\|_F^2\|\mathbf{H}^{-1}\|^2 \\
&= O_p(\frac{pK^2\log p}{T^2} + \frac{K^4\log p}{T}), \\
\|B_{22}\|_\Sigma^2 &\leq p^{-1}\|\mathbf{B}'\mathbf{\Sigma}^{-1}\mathbf{B}\|^2\left\|\frac{1}{T}\mathbf{H}^{-1}(\mathbf{H}\mathbf{F}' - \widehat{\mathbf{F}}')\widehat{\mathbf{F}}\right\|_F^4 \\
&= O_p(\frac{K^2}{pT^2} + \frac{K^6}{p^3}), \\
\|B_{23}\|_\Sigma^2 &\leq p^{-1}\|\mathbf{B}'\mathbf{\Sigma}^{-1}\mathbf{B}\|\|\mathbf{\Sigma}^{-1}\|\|\mathbf{H}^{-1}\|^2\left\|\frac{1}{T}\mathbf{E}(\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H}')\right\|_F^2
\end{aligned}$$

$$\times \left\| \frac{1}{T}(\mathbf{H}\mathbf{F}' - \widehat{\mathbf{F}}')\widehat{\mathbf{F}} \right\|_F^2$$
$$= O_p(\frac{K^2}{T^2} + \frac{K^6}{p^2} + \frac{K^4}{pT}).$$

Combining these results yields the lemma. Q.E.D.

Proof of Theorem 3.2 (i)

By Lemmas B.12-B.14,

$$\begin{aligned} \|\widehat{\mathbf{\Sigma}}^{\mathcal{T}} - \mathbf{\Sigma}\|_\Sigma^2 &\leq C[\|\mathbf{B}\mathbf{H}^{-1}\mathbf{D}_T(\mathbf{B}\mathbf{H}^{-1})'\|_\Sigma^2 + \|\mathbf{B}\mathbf{H}^{-1}\mathbf{C}_T{}'\|_\Sigma^2 \\ &\quad + \|\mathbf{C}_T\mathbf{C}_T{}'\|_\Sigma^2] + \|\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}} - \mathbf{\Sigma}_u\|_\Sigma^2 \\ &= \|\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}} - \mathbf{\Sigma}_u\|_\Sigma^2 + O_p(\frac{pK^2(\log p)^2}{T^2} + \frac{K^4 \log p}{T} + \frac{K^6}{p}). \end{aligned}$$

The theorem follows directly from Theorem 3.1. Q.E.D.

Lemma B.15. *(i)* $\lambda_{\min}((\mathbf{H}\mathbf{H}')^{-1} + (\mathbf{B}\mathbf{H}^{-1})'\mathbf{\Sigma}_u^{-1}\mathbf{B}\mathbf{H}^{-1}) \geq cp$ *for some* $c > 0$.
(ii) $\|\mathbf{I}_K - (\mathbf{H}\mathbf{H}')^{-1}\| = O_p(K\sqrt{(\log K)/T} + K^2/\sqrt{p})$.
(iii) $\|\widehat{\mathbf{\Lambda}}'(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1}\widehat{\mathbf{\Lambda}} - (\mathbf{B}\mathbf{H}^{-1})'\mathbf{\Sigma}_u^{-1}\mathbf{B}\mathbf{H}^{-1}\| = O_p(pm_pK\sqrt{(\log p)/T} + \sqrt{p}m_pK^3 + pm_pK^2/\sqrt{T})$.

Proof. (i) We have,

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

Part (ii) follows from Lemma B.11 and Lemma B.1. For part (iii), by Theorem 3.3,

$$\|\widehat{\mathbf{\Lambda}} - \mathbf{B}\mathbf{H}^{-1}\|_F^2 = \sum_{i=1}^{p} \|\widehat{\mathbf{b}}_i - (\mathbf{H}')^{-1}\mathbf{b}_i\|^2 = O_p(K^5 + pK\log p/T + pK^3/T). \tag{B.18}$$

Since $\|(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1}\| = O_p(1)$, $\|\mathbf{B}\| = O(\sqrt{p})$, and

$$\begin{aligned} \|\widehat{\mathbf{\Lambda}}'(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1}\widehat{\mathbf{\Lambda}} - (\mathbf{B}\mathbf{H}^{-1})'\mathbf{\Sigma}_u^{-1}\mathbf{B}\mathbf{H}^{-1}\| &\leq \|(\widehat{\mathbf{\Lambda}} - \mathbf{B}\mathbf{H}^{-1})'(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1}(\widehat{\mathbf{\Lambda}} - \mathbf{B}\mathbf{H}^{-1})\| \\ &\quad + 2\|(\widehat{\mathbf{\Lambda}} - \mathbf{B}\mathbf{H}^{-1})'(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1}\mathbf{B}\mathbf{H}^{-1}\| \\ &\quad + \|\mathbf{B}\mathbf{H}^{-1}((\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1} - \mathbf{\Sigma}_u^{-1})\mathbf{B}\mathbf{H}^{-1}\|. \end{aligned} \tag{B.19}$$

The desired result then follows from Theorem 3.1 and (B.18). Q.E.D.

Proof of Theorem 3.2: $\|(\widehat{\mathbf{\Sigma}}^{\mathcal{T}})^{-1} - \mathbf{\Sigma}^{-1}\|$.

Using the Sherman-Morrison-Woodbury formula, we have

$$\|\widehat{\mathbf{\Sigma}}^{\mathcal{T}-1} - \mathbf{\Sigma}^{-1}\| \leq \sum_{i=1}^{6} L_i,$$

where

$$\begin{aligned}
L_1 &= \|(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1} - \mathbf{\Sigma}_u^{-1}\| \\
L_2 &= \|((\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1} - \mathbf{\Sigma}_u^{-1})\widehat{\mathbf{\Lambda}}[\mathbf{I}_K + \widehat{\mathbf{\Lambda}}'(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1}\widehat{\mathbf{\Lambda}}]^{-1}\widehat{\mathbf{\Lambda}}'(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1}\| \\
L_3 &= \|((\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1} - \mathbf{\Sigma}_u^{-1})\widehat{\mathbf{\Lambda}}[\mathbf{I}_K + \widehat{\mathbf{\Lambda}}'(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1}\widehat{\mathbf{\Lambda}}]^{-1}\widehat{\mathbf{\Lambda}}'\mathbf{\Sigma}_u^{-1}\| \\
L_4 &= \|\mathbf{\Sigma}_u^{-1}(\widehat{\mathbf{\Lambda}} - \mathbf{B}\mathbf{H}^{-1})[\mathbf{I}_K + \widehat{\mathbf{\Lambda}}'(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1}\widehat{\mathbf{\Lambda}}]^{-1}\widehat{\mathbf{\Lambda}}'\mathbf{\Sigma}_u^{-1}\| \\
L_5 &= \|\mathbf{\Sigma}_u^{-1}(\widehat{\mathbf{\Lambda}} - \mathbf{B}\mathbf{H}^{-1})[\mathbf{I}_K + \widehat{\mathbf{\Lambda}}'(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1}\widehat{\mathbf{\Lambda}}]^{-1}(\mathbf{H}')^{-1}\mathbf{B}'\mathbf{\Sigma}_u^{-1}\| \\
L_6 &= \|\mathbf{\Sigma}_u^{-1}\mathbf{B}\mathbf{H}^{-1}([\mathbf{I}_K + \widehat{\mathbf{\Lambda}}'(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1}\widehat{\mathbf{\Lambda}}]^{-1} \\
&\quad -[\mathrm{cov}(\mathbf{H}\mathbf{f})^{-1} + (\mathbf{H}')^{-1}\mathbf{B}'\mathbf{\Sigma}_u^{-1}\mathbf{B}\mathbf{H}^{-1}]^{-1})(\mathbf{H}')^{-1}\mathbf{B}'\mathbf{\Sigma}_u^{-1}\|. \qquad \text{(B.20)}
\end{aligned}$$

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

$$L_2 \leq \|(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1} - \mathbf{\Sigma}_u^{-1}\| \cdot \|\widehat{\mathbf{\Lambda}}\mathbf{G}\widehat{\mathbf{\Lambda}}'\| \cdot \|(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1}\|.$$

Note that Theorem 3.1 implies $\|(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1}\| = O_p(1)$. Lemma B.15 implies

$$\|\mathbf{G}\| = O_p(p^{-1}). \qquad \text{(B.21)}$$

This shows that $L_2 = O_p(L_1)$. Similarly $L_3 = O_p(L_1)$.

In addition, by (B.18),

$$L_4 \leq \|\mathbf{\Sigma}_u^{-1}(\widehat{\mathbf{\Lambda}} - \mathbf{B}\mathbf{H}^{-1})\|\|\mathbf{G}\|\|\widehat{\mathbf{\Lambda}}'\mathbf{\Sigma}_u^{-1}\| = O_p(\sqrt{\frac{K^5}{p}} + \sqrt{\frac{K\log p + K^3}{T}}).$$

Similarly $L_5 = O_p(L_4)$. Finally, let

$$\mathbf{G}_1 = [(\mathbf{H}\mathbf{H}')^{-1} + (\mathbf{B}\mathbf{H}^{-1})'\mathbf{\Sigma}_u^{-1}\mathbf{B}\mathbf{H}^{-1}]^{-1}.$$

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

$$\begin{aligned}
\|\mathbf{G} - \mathbf{G}_1\| &= \|\mathbf{G}(\mathbf{G}^{-1} - \mathbf{G}_1^{-1})\mathbf{G}_1\| \\
&\leq O_p(p^{-2})\|(\mathbf{H}\mathbf{H}')^{-1} - \mathbf{I}_K\|
\end{aligned}$$

$$
\begin{aligned}
& +O_p(p^{-2})\|(\mathbf{B}\mathbf{H}^{-1})'\mathbf{\Sigma}_u^{-1}\mathbf{B}\mathbf{H}^{-1} - \widehat{\mathbf{\Lambda}}'(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1}\widehat{\mathbf{\Lambda}}\| \\
= \; & O_p\left(\frac{m_T K\sqrt{\log p} + m_p K^3}{p\sqrt{T}} + \frac{m_T K^2}{p^{3/2}}\right).
\end{aligned}
$$

Consequently,

$$
L_6 \leq \|\mathbf{\Sigma}_u^{-1}\mathbf{B}\mathbf{H}^{-1}\|^2\|\mathbf{G} - \mathbf{G}_1\| = O_p(\frac{m_T K\sqrt{\log p} + m_p K^2}{\sqrt{T}} + \frac{m_T K^3}{\sqrt{p}}).
$$

Adding up L_1-L_6 gives the result.

Q.E.D.

Proof of Theorem 3.2: $\|\widehat{\mathbf{\Sigma}}^{\mathcal{T}} - \mathbf{\Sigma}\|_{\max}$

Let $\mathbf{e}_{i,p}$ denote a p-dimensional unit vector with one on the ith element and rest zero.

Lemma B.16. *(i)* $\|\frac{1}{T}(\mathbf{H}\mathbf{F}' - \widehat{\mathbf{F}})\mathbf{E}'\|_{\max} = O_p(1/\sqrt{T} + K/\sqrt{p})$.
(ii) $\|\widehat{\mathbf{\Lambda}} - \mathbf{B}\mathbf{H}^{-1}\|_{\max} = O_p(K^2/\sqrt{p} + \sqrt{K(\log p)/T} + K/\sqrt{T})$.
(iii) $\|\mathbf{I}_K - \mathbf{H}\mathbf{H}'\|_{\max} = O_p(K\sqrt{\log K/T} + K/\sqrt{p})$.

Proof. (i) By the Cauchy-Schwarz inequality and Lemma B.8,

$$
\begin{aligned}
\|\frac{1}{T}(\mathbf{H}\mathbf{F}' - \widehat{\mathbf{F}})\mathbf{E}'\|_{\max} &= \|\frac{1}{T}\sum_{t=1}^{T}(\mathbf{H}\mathbf{f}_t - \widehat{\mathbf{f}}_t)\mathbf{u}_t'\|_{\max} \leq \max_{i,j}\frac{1}{T}|\sum_{t=1}^{T}(\mathbf{H}\mathbf{f}_t - \widehat{\mathbf{f}}_t)_i u_{jt}| \\
&\leq \max_{ij}\sqrt{\frac{1}{T}\sum_{t=1}^{T}(\mathbf{H}\mathbf{f}_t - \widehat{\mathbf{f}}_t)_i^2\frac{1}{T}\sum_{t=1}^{T}u_{jt}^2} \\
&\leq O_p(\frac{1}{\sqrt{T}} + \frac{K}{\sqrt{p}})\sqrt{\max_j\frac{1}{T}\sum_{t=1}^{T}u_{jt}^2}. \qquad \text{(B.22)}
\end{aligned}
$$

Lemma B.3 implies $\max_{i\leq p}\frac{1}{T}\sum_t u_{it}^2 = O_p(1)$, which yields the result.

(ii) We have

$$
\begin{aligned}
\|\widehat{\mathbf{\Lambda}} - \mathbf{B}\mathbf{H}^{-1}\|_{\max} \leq \; & \|\frac{1}{T}\mathbf{E}\mathbf{F}\mathbf{H}'\|_{\max} + \|\frac{1}{T}\mathbf{B}\mathbf{H}^{-1}(\mathbf{H}\mathbf{F}' - \widehat{\mathbf{F}}')\widehat{\mathbf{F}}\|_{\max} \\
& + \|\frac{1}{T}\mathbf{E}(\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H}')\|_{\max}.
\end{aligned}
$$

By Lemmas B.3, B.9,

$$
\|\frac{1}{T}\mathbf{E}\mathbf{F}\mathbf{H}'\|_{\max} = \max_{i,j}|\mathbf{e}_{i,p}'\frac{1}{T}\mathbf{E}\mathbf{F}\mathbf{H}'\mathbf{e}_{j,K}| \leq \max_{i,j}\|\mathbf{e}_{i,p}'\frac{1}{T}\mathbf{E}\mathbf{F}\|\|\mathbf{H}'\mathbf{e}_{j,K}\|
$$

$$
\begin{aligned}
&\leq O_p(\max_i \|\mathbf{e}'_{i,p}\frac{1}{T}\mathbf{EF}\|) \leq O_p(\sqrt{K})\max_{i,j}|\frac{1}{T}\sum_{t=1}^{T} f_{jt}u_{it}| \\
&= O_p(\sqrt{\frac{K\log p}{T}}).
\end{aligned}
$$

By Lemma B.8 and the Cauchy-Schwarz inequality,

$$
\begin{aligned}
\|\frac{1}{T}\mathbf{BH}^{-1}(\mathbf{HF}'-\widehat{\mathbf{F}})\widehat{\mathbf{F}}'\|_{\max} &= \max_{ij}|\mathbf{e}'_{i,p}\frac{1}{T}\mathbf{BH}^{-1}(\mathbf{HF}'-\widehat{\mathbf{F}}')\widehat{\mathbf{F}}\mathbf{e}_{j,K}| \\
&\leq \max_{ij}\|\mathbf{e}'_{i,p}\mathbf{BH}^{-1}\|\|\frac{1}{T}\sum_{t=1}^{T}(\mathbf{Hf}_t-\widehat{\mathbf{f}}_t)\hat{f}_{jt}\| \\
&\leq O_p(\sqrt{K})\max_j\sqrt{\frac{1}{T}\sum_{t=1}^{T}\|\mathbf{Hf}_t-\widehat{\mathbf{f}}_t\|^2\frac{1}{T}\sum_{t=1}^{T}\hat{f}_{jt}^2} \\
&= O_p(\frac{K}{\sqrt{T}}+\frac{K^2}{\sqrt{p}}).
\end{aligned}
$$

Still by Lemma B.8 and the Cauchy-Schwarz inequality,

$$
\begin{aligned}
\|\frac{1}{T}\mathbf{E}(\widehat{\mathbf{F}}-\mathbf{HF}')'\|_{\max} &= \max_{ij}|\frac{1}{T}\sum_{t=1}^{T}u_{it}(\widehat{\mathbf{f}}_t-\mathbf{Hf}_t)_j| \\
&\leq \max_{ij}\sqrt{\frac{1}{T}\sum_{t=1}^{T}u_{it}^2\frac{1}{T}\sum_{t=1}^{T}(\widehat{\mathbf{f}}_t-\mathbf{Hf}_t)_j^2} \\
&= O_p(\frac{1}{\sqrt{T}}+\frac{K}{\sqrt{p}}).
\end{aligned}
$$

Therefore we have the desired result.

(iii) By the triangular inequality,

$$
\|\mathbf{I}_K-\mathbf{HH}'\|_{\max} \leq \|\mathbf{I}_K-\widehat{\mathrm{cov}}(\mathbf{Hf}_t)\|_{\max}+\|\widehat{\mathrm{cov}}(\mathbf{Hf}_t)-\mathbf{HH}'\|_{\max}.
$$

In the proof of Lemma B.11, $\|\widehat{\mathrm{cov}}(\mathbf{Hf}_t)-\mathbf{HH}'\|_{\max}=O_p(K\sqrt{\log K/T})$, where we used the fact that $\|.\|_{\max}$ is dominated by $\|.\|_F$. In addition,

$$
\begin{aligned}
\|\mathbf{I}_K-\widehat{\mathrm{cov}}(\mathbf{Hf}_t)\|_{\max} &\leq \max_{ij}\sqrt{\frac{1}{T}\sum_t|\mathbf{Hf}_t-\widehat{\mathbf{f}}_t|_i^2\frac{1}{T}\sum_t(\mathbf{Hf}_t)_j^2} \\
&\quad+\max_{ij}\sqrt{\frac{1}{T}\sum_t|\mathbf{Hf}_t-\widehat{\mathbf{f}}_t|_i^2\frac{1}{T}\sum_t\hat{f}_{jt}^2} \\
&= O_p(\frac{1}{\sqrt{T}}+\frac{K}{\sqrt{p}}). \qquad \text{(B.23)}
\end{aligned}
$$

Q.E.D.

Completion of the Proof of Theorem 3.2 We first bound

$$\begin{aligned}\mathbf{A} &\equiv \|\widehat{\mathbf{\Lambda}}\widehat{\mathbf{\Lambda}}' - (\mathbf{B}\mathbf{H}^{-1})\mathbf{H}\mathbf{H}'(\mathbf{B}\mathbf{H}^{-1})'\|_{\max} \\ &\leq \|2\mathbf{C}_T(\mathbf{B}\mathbf{H}^{-1})'\|_{\max} + \|\mathbf{B}\mathbf{H}^{-1}\mathbf{D}_T(\mathbf{B}\mathbf{H}^{-1})'\|_{\max} + \|\mathbf{C}_T\mathbf{C}_T'\|_{\max} \\ &\quad + \|2\mathbf{B}\mathbf{H}^{-1}\mathbf{D}_T\mathbf{C}_T'\|_{\max}.\end{aligned}$$

We bound the terms on the right hand side. By Theorem 3.3,

$$\begin{aligned}\|\mathbf{C}_T(\mathbf{H}^{-1})'\mathbf{B}'\|_{\max} &= \max_{ij}|\mathbf{e}_{i,p}'\mathbf{C}_T(\mathbf{H}^{-1})'\mathbf{B}'\mathbf{e}_{j,p}| \leq \max_{ij}\|\widehat{\mathbf{b}}_i - (\mathbf{H}')^{-1}\mathbf{b}_i\|\|\mathbf{H}^{-1}\|\|\mathbf{b}_j\| \\ &= O_p(\frac{\delta_T}{m_p}).\end{aligned}$$

By Lemma B.16(iii), $\|\mathbf{D}_T\|_{\max} = O_p(K\sqrt{\log K/T} + K/\sqrt{p})$. Hence

$$\begin{aligned}\|\mathbf{B}\mathbf{H}^{-1}\mathbf{D}_T(\mathbf{B}\mathbf{H}^{-1})'\|_{\max} &\leq \max_{ij}\|\mathbf{e}_{i,p}'\mathbf{B}\|\|\mathbf{D}_T\|\|\mathbf{e}_{j,p}'\mathbf{B}\| = O_p(K^2)\|\mathbf{D}_T\|_{\max} \\ &= O_p(K^3\sqrt{\frac{\log K}{T}} + \frac{K^3}{\sqrt{p}}).\end{aligned}$$

Since $\|\mathbf{e}_{i,p}'\mathbf{C}_T\| = \|\widehat{\mathbf{b}}_i - (\mathbf{H}')^{-1}\mathbf{b}_i\|$,

$$\begin{aligned}\|\mathbf{C}_T\mathbf{H}\mathbf{H}'\mathbf{C}_T'\|_{\max} &\leq \max_i\|\mathbf{e}_{i,p}'\mathbf{C}_T\|^2\|\mathbf{H}\mathbf{H}'\| = O_p(\max_{i\leq p}\|\widehat{\mathbf{b}}_i - (\mathbf{H}')^{-1}\mathbf{b}_i\|^2) \\ &= O_p(\frac{\delta_T^2}{m_p^2 K}).\end{aligned}$$

$$\begin{aligned}\|2\mathbf{B}\mathbf{H}^{-1}\mathbf{D}_T\mathbf{C}_T'\|_{\max} &\leq \max_{i,j}\|\mathbf{e}_{i,p}'\mathbf{B}\|\|\mathbf{D}_T\|\|\mathbf{C}_T'\mathbf{e}_{j,p}\| \\ &= o_p(\frac{\delta_T}{m_p}).\end{aligned}$$

Therefore,

$$\mathbf{A} = O_p(K^3\sqrt{\frac{\log K}{T}} + \frac{\delta_T}{m_p}).$$

Finally,

$$\|\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}} - \mathbf{\Sigma}_u\|_{\max} = O_p\left(\frac{\delta_T}{m_p}\right).$$

Therefore,

$$\|\widehat{\mathbf{\Sigma}}^{\mathcal{T}} - \mathbf{\Sigma}\|_{\max} = O_p(K^3\sqrt{\frac{\log K}{T}} + \frac{\delta_T}{m_p}).$$

Q.E.D.

References

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

Amini, A. A. and Wainwright, M. J. (2009). High-dimensional analysis of semidefinite relaxations for sparse principal components. *Annals of Statistics*, **37**, 28772921.

Antoniadis, A. and Fan, J. (2001). Regularized wavelet approximations. *J. Amer. Statist. Assoc.* **96**, 939-967.

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

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

Bai, J. and Ng, S.(2008). Large dimensional factor analysis. *Foundations and trends in econometrics*. **3** 89-163.

Bai, J. and Shi, S.(2011). Estimating high dimensional covariance matrices and its applications. *Annals of Economics and Finance*. **12** 199-215.

Bickel, P. and Levina, E. (2004). Some theory for Fisher's linear discriminant function, "naive Bayes", and some alternatives when there are many more variables than observations *Bernoulli*. **10** 989-1010.

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

Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.* **106**, 672-684.

Cai, T. and Zhou, H. (2010). Optimal rates of convergence for sparse covariance matrix estimation. *Manuscript*. University of Pennsylvania.

Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure and mean-variance analyssi in large asset markets. *Econometrica*. **51** 1305-1324.

Doz, C., Giannone, D. and Reichlin, L. (2006). A two-step estimator for large approximate dynamic factor models based on Kalman filtering. *Manuscript*. Universite de Cergy-Pontoise.

dAspremont, A., Bach, F. and El Ghaoui, L. (2008). Optimal solutions for sparse principal component analysis. *Journal of Machine Learning Research*, **9**, 12691294.

Davis, C. and Kahan, W. (1970). The rotation of eigenvectors by a perturbation III. *SIAM Journal on Numerical Analysis*, **7**, 146.

Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *Jour Ameri. Statist. Assoc.*, **102**, 93-103.

Efron, B. (2010). Correlated z-values and the accuracy of large-scale statistical estimates. *Jour Ameri. Statist. Assoc.*, **105**, 1042-1055.

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

Fan, J., Fan, Y. and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *J. Econometrics*.

Fan, J., Han, X., and Gu, W.(2012). Control of the false discovery rate under arbitrary covariance dependence (with discussion). *Journal of American Statistical Association*, to appear.

Fan, J., Liao, Y. and Mincheva, M. (2011). High dimensional covariance matrix estimation in approximate factor models. *Ann. Statist*. To appear.

Fan, J., Zhang, J., and Yu, K. (2008). Asset Allocation and Risk Assessment with Gross Exposure Constraints for Vast Portfolios. *Manuscript*.

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

Hallin, M. and Liška, R. (2007). Determining the number of factors in the general dynamic factor model. *J. Amer. Statist. Assoc.* **102**, 603-617.

Harding, M. (2009). Structural estimation of high-dimensional factor models. *Manuscript* Stanford University.

Hastie, T.J., Tibshirani, R. and Friedman, J. (2009). *The elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed). *Springer*, New York.

James, W. and Stein, C. (1961). Estimation with quadratic loss, in *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1** 361-379. Univ. California Press. Berkeley.

Johnstone, I.M. and Lu, A.Y. (2009). On consistency and sparsity for principal components analysis in high dimensions.*Jour. Ameri. Statist. Assoc.*, **104**, 682-693.

Jung, S. and Marron, J.S. (2009). PCA consistency in high dimension, low sample size context. *Ann. Statist.*, **37**, 4104-4130.

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

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37** 4254-4278.

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

Ma, Z. (2011). Sparse principal components analysis and iterative thresholllding. *Manuscript.*

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

Merlevède, F., Peligrad, M. and Rio, E. (2009). A Bernstein type inequality and moderate deviations for weakly dependent sequences. *Manuscript*. Université Paris Est.

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics*. **92**, 10041016.

Pesaran, M.H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica.* **74**, 967-1012.

Ross, S.A. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, **13**, 341-360.

Rothman, A., Levina, E. and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *J. Amer. Statist. Assoc.* **104** 177-186.

Sentana, E. (2009). The econometrics of mean-variance efficiency tests: a survey *Econometrics Jour.*, **12**, C65C101.

Shen, H. and Huang, J. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *J. Multivariate Analysis* **99**, 1015 1034.

Leek, J.T. and Storey, J.D. (2008). A general framework for multiple testing dependence. *Proc. Natl. Acad. Sci.*, **105**, 18718-19723.

Sharpe, W. (1964). Capital asset prices: A theory of market equilibrium under conditons of risks. *Journal of Finance*, **19**, 425-442.

Stock, J. and Watson, M. (2002). Forecasting using principal components from a large number of predictors. *J. Amer. Statist. Assoc.* **97**, 1167-1179.

Wang, P. (2010). Large dimensional factor models with a multi-level factor structure: identification, estimation and inference. *Manuscript*. Hong Kong University of Science and Technology.

Witten, D.M., Tibshirani, R. and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, **10**, 515-534.

Xiong, H., Goulding, E.H., Carlson, E.J., Tecott, L.H., McCulloch, C.E. and Sen, S. (2011). A Flexible Estimating Equations Approach for Mapping Function-Valued Traits. Genetics, **189**, 305–316.

Yap, J.S., Fan, J., and Wu, R. (2009). Nonparametric modeling of longitudinal covariance structure in functional mapping of quantitative trait loci. *Biometrics*, **65**, 1068-1077.

Zhang, Y. and El Ghoui, L. (2011) Large-scale sparse principal component analysis with application to text data. *NIPS*.