Submitted to the Annals of Statistics

HIGH DIMENSIONAL COVARIANCE MATRIX ESTIMATION IN APPROXIMATE FACTOR MODELS

BY JIANQING FAN, YUAN LIAO AND MARTINA MINCHEVA

Princeton University

The variance covariance matrix plays a central role in the inferential theories of high dimensional factor models in finance and economics. Popular regularization methods of directly exploiting sparsity are not directly applicable to many financial problems. Classical methods of estimating the covariance matrices are based on the strict factor models, assuming independent idiosyncratic components. This assumption, however, is restrictive in practical applications. By assuming sparse error covariance matrix, we allow for the presence of the cross-sectional correlation even after taking out common factors, and it enables us to combine the merits of both sparsity and factor structures. We estimate the sparse covariance using the adaptive thresholding technique as in Cai and Liu (2011), taking into account the fact that direct observations of the idiosyncratic components are unavailable. The impact of high dimensionality on the covariance matrix estimation based on the factor structure is then studied.

1. Introduction.

We consider a factor model defined as follows:

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

where y_{it} is the observed datum for the ith ($i = 1, ..., p$) asset at time $t = 1, ..., T$; $\mathbf{b}_i$ is a $K \times 1$ vector of factor loadings; $\mathbf{f}_t$ is a $K \times 1$ vector of common factors, and u_{it} is the idiosyncratic error component of y_{it}. Classical factor analysis assumes that both p and K are fixed, while T is allowed to grow. However, in the recent decades, both economic and financial applications have encountered very large data sets which contain high dimensional variables. For example, the World Bank has data for about two-hundred countries over forty years; in portfolio allocation, the number of stocks can be in thousands and be larger or of the same order of the sample size. In modeling housing prices in each zip code, the number of regions can be of order thousands, yet the sample size can be 240 months or twenty years. The covariance matrix of order several thousands is critical for understanding the co-movement housing prices indices over these zip codes.

Inferential theory of factor analysis relies on estimating $\mathbf{\Sigma}_u$, the variance covariance matrix of the error term, and $\mathbf{\Sigma}$, the variance-covariance matrix of $\mathbf{y}_t =$

AMS 2000 subject classifications: Primary 62H25 ; secondary 62F12, 62H12

Keywords and phrases: sparse estimation, thresholding, cross-sectional correlation, common factors, idiosyncratic, seemingly unrelated regression

$(y_{1t}, ..., y_{pt})'$. In the literature, $\mathbf{\Sigma} = \text{cov}(\mathbf{y}_t)$ was traditionally estimated by the sample covariance matrix of $\mathbf{y}_t$:

$$\widehat{\mathbf{\Sigma}}_{sam} = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{y}_t - \bar{\mathbf{y}})(\mathbf{y}_t - \bar{\mathbf{y}})',$$

which was always assumed to be pointwise root-T consistent. However, the sample covariance matrix is an inappropriate estimator in high dimensional settings. For example, when p is larger than T, $\widehat{\mathbf{\Sigma}}_{sam}$ becomes singular while $\mathbf{\Sigma}$ is always strictly positive definite. Even if $p < T$, Fan, Fan and Lv (2008) showed that this estimator has a very slow convergence rate under the Frobenius norm. Realizing the limitation of the sample covariance estimator in high dimensional factor models, Fan, Fan and Lv (2008) considered more refined estimation of $\mathbf{\Sigma}$, by incorporating the common factor structure. One of the key assumptions they made was the cross-sectional independence among the idiosyncratic components, which results $E\mathbf{u}_t\mathbf{u}_t'$ to be a diagonal matrix. The cross-sectional independence, however, is restrictive in many applications, as it rules out the *approximate factor structure* as in Chamberlain and Rothschild (1983). In this paper, we relax this assumption, and investigate the impact of the cross-sectional correlations among the idiosyncratic noises on the estimation of $\mathbf{\Sigma}$ and $\mathbf{\Sigma}_u$, when both p and T are allowed to diverge.

Sparsity is one of the commonly used assumptions in estimating high dimensional covariance matrices, which assumes that many entries of the off diagonal elements are zero, and the number of nonzero off-diagonal entries is restricted to grow slowly. While directly assuming $\mathbf{\Sigma}$ to be sparse is inappropriate in financial applications, it is more reasonable to assume that $\mathbf{\Sigma}_u$ is sparse. We estimate both $\mathbf{\Sigma}_u$ and $\mathbf{\Sigma}_u^{-1}$ using the thresholding method (Bickel and Levina (2008a), Cai and Liu (2011)) based on the estimated residuals in the factor model. It is assumed that the factors $\mathbf{f}_t$ are observable, as in Fama and French (1992), Fan, Fan and Lv (2008), and many other empirical applications in finance and economics. We derive the convergence rates of both estimated $\mathbf{\Sigma}$ and its inverse respectively under various norms which are to be defined later. We show that the estimated covariance matrices are still invertible even if $p > T$. In particular, when estimating $\mathbf{\Sigma}^{-1}$ and $\mathbf{\Sigma}_u^{-1}$, p is allowed to grow exponentially fast in T. In addition, we achieve better convergence rates than those in Fan, Fan and Lv (2008).

In recent years, various approaches have been proposed on estimating the large covariance matrix: Bickel and Levia (2008a, 2008b) constructed the estimators based on regularization and thresholding respectively. Rothman, Levina and Zhou (2009) considered thresholding of the sample covariance matrix with more general thresholding functions. Lam and Fan (2009) proposed penalized quasi-likelihood method to achieve both the consistency and sparsistency of the estimation. More

recently, Cai and Zhou (2010) derived the minimax rate for sparse matrix estimation, and showed that the thresholding estimator attains this optimal rate under the operator norm. Cai and Liu (2011) proposed a thresholding procedure which is adaptive to the variability of individual entries, and unveiled its improved rate of convergence.

The rest of the paper is organized as follows. Section 2 provides the asymptotic theory for estimating the error covariance matrix and its inverse. Section 3 considers estimating the covariance matrix of $\mathbf{y}_t$. Section 4 extends the results to the *seemingly unrelated regression* model, a set of linear equations with correlated error terms in which the covariates are different across equations. Section 5 reports the simulation results. Finally, Section 6 concludes with discussions. 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 $\|\mathbf{A}\|_F$, $\|\mathbf{A}\|$ and $\|\mathbf{A}\|_\infty$ the Frobenius norm, operator norm and elementwise norm of a matrix $\mathbf{A}$ respectively, defined respectively as $\|\mathbf{A}\|_F = \mathrm{tr}^{1/2}(\mathbf{A}'\mathbf{A})$, $\|\mathbf{A}\| = \lambda_{\max}^{1/2}(\mathbf{A}'\mathbf{A})$, and $\|\mathbf{A}\|_\infty = \max_{i,j}|A_{ij}|$. Note that, when $\mathbf{A}$ is a vector, $\|\mathbf{A}\|$ is equal to the Euclidean norm.

2. Estimation of Error Covariance Matrix .

2.1. *Sparsity and thresholding.* Consider the following approximate factor model, in which the cross-sectional correlation among the idiosyncratic error components is allowed:

$$y_{it} = \mathbf{b}_i'\mathbf{f}_t + u_{it}, \tag{2.1}$$

where $i = 1, .., , p$ and $t = 1, ..., T$; $\mathbf{b}_i$ is a $K \times 1$ vector of factor loadings; $\mathbf{f}_t$ is a $K \times 1$ vector of observable common factors, uncorrelated with u_{it}. Write

$$\mathbf{B} = (\mathbf{b}_1, ..., \mathbf{b}_p)', \quad \mathbf{y}_t = (y_{1t}, ..., y_{pt})', \quad \mathbf{u}_t = (u_{1t}, ..., u_{pt})',$$

then model (2.1) can be written in a more compact form:

$$\mathbf{y}_t = \mathbf{B}\mathbf{f}_t + \mathbf{u}_t, \tag{2.2}$$

with $E(\mathbf{u}_t|\mathbf{f}_t) = 0$.

In practical applications, p can be thought of as the number of assets or stocks, or number of regions in spatial and temporal problems such as home price indices or sales of drugs, which can increase with the sample size T, and in practice can be of the same order as, or even larger than T. For example, an asset pricing model may contain hundreds of assets while the sample size on daily returns is less than several hundreds. In the estimation of the optimal portfolio allocation, it was observed by

Fan, Fan and Lv (2008) that the effect of large p on the convergence rate can be quite severe. In contrast, the number of common factors, K, can be much smaller. For example, the rank theory of consumer demand systems implies no more than three factors (e.g., Gorman (1981) and Lewbel (1991)).

The error covariance matrix

$$\mathbf{\Sigma}_u = \mathrm{cov}(\mathbf{u}_t),$$

itself is of interest for the inferential theory of factor models. For example, the asymptotic covariance of the least square estimator of $\mathbf{B}$ depends on $\mathbf{\Sigma}_u^{-1}$, and in simulating home price indices over a certain time horizon for mortgage based securities, a good estimate of $\mathbf{\Sigma}_u$ is needed. When p is close to or larger than T, estimating $\mathbf{\Sigma}_u$ is very challenging. Therefore, following the literature of high dimensional covariance matrix estimation, we assume it is sparse, i.e., many of its off-diagonal entries are zeros. Specifically, let $\mathbf{\Sigma}_u = (\sigma_{ij})_{p\times p}$. Define

$$m_T = \max_{i\leq p} \sum_{j\leq p} I(\sigma_{ij} \neq 0). \tag{2.3}$$

The sparsity assumption puts an upper bound restriction on m_T. Specifically, we assume:

$$m_T^2 = o\left(\frac{T}{K^2 \log p}\right). \tag{2.4}$$

In this formulation, we even allow the number of factors K to be large, possibly growing with T.

A more general treatment (e.g., Bickel and Levina (2008a) and Cai and Liu (2011)) is to assume that the l_q norm of the row vectors of $\mathbf{\Sigma}_u$ are uniformly bounded across rows by a slowly growing sequence, for some $q \in [0, 1)$. In contrast, the assumption we make in this paper, i.e., $q = 0$, has a clearer economic interpretation. For example, the firm returns can be modeled by the factor model, where u_{it} represents a firm's individual shock at time t. Driven by the industry-specific components, these shocks are correlated among the firms in the same industry, but can be assumed to be uncorrelated across industries, since the industry-specific components are not pervasive for the whole economy (Connor and Korajczyk (1993)).

We estimate $\mathbf{\Sigma}_u$ using the thresholding technique first introduced by Bickel and Levina (2008a), and later extended by Rothman, Levina and Zhu (2009), and improved by Cai and Liu (2011), which is summarized as follows: Suppose we observe data $(\mathbf{X}_1, ..., \mathbf{X}_T)$ of a $p \times 1$ vector $\mathbf{X}$, which follows a multivariate Gaussian distribution $N(0, \mathbf{\Sigma}_X)$. The sample covariance matrix of $\mathbf{X}$ is thus given by:

$$\mathbf{S}_X = \frac{1}{T}\sum_{i=1}^{T}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})' = (s_{ij})_{p\times p}.$$

Define the thresholding operator by $\mathcal{T}_t(\mathbf{M}) = (M_{ij}I(|M_{ij}| \geq t))$ for any symmetric matrix $\mathbf{M}$. Then $\mathcal{T}_t$ preserves the symmetry of $\mathbf{M}$. Let $\widehat{\mathbf{\Sigma}}_X^{\mathcal{T}} = T_{\omega_T}(\mathbf{S}_X)$, where $\omega_T = O(\sqrt{\log p/T})$. Bickel and Levina (2008a) then showed that:

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

In the factor models, however, we do not observe the idiosyncratic components directly. Hence when estimating the error covariance matrix, we need to construct a sample covariance matrix based on the residuals $\hat{u}_{it}$ before thresholding. The residuals are obtained using the plug-in method, by estimating the factor loadings first. Let $\widehat{\mathbf{b}}_i$ be the ordinary least square (OLS) estimator of $\mathbf{b}_i$, and

$$\hat{u}_{it} = y_{it} - \widehat{\mathbf{b}}_i'\mathbf{f}_t.$$

Denote by $\widehat{\mathbf{u}}_t = (\hat{u}_{1t}, ..., \hat{u}_{pt})'$. We then construct the residual covariance matrix as:

$$\widehat{\mathbf{\Sigma}}_u = \frac{1}{T}\sum_{t=1}^{T} \widehat{\mathbf{u}}_t\widehat{\mathbf{u}}_t' = (\widehat{\sigma}_{ij}).$$

Note that the thresholding value $\omega_T = O(\sqrt{\log p/T})$ in Bickel and Levina (2008a) is in fact obtained from the rate of convergence of $\max_{ij}|s_{ij} - \mathbf{\Sigma}_{X,ij}|$. This rate changes when s_{ij} is obtained from the residual $\hat{u}_{ij}$, which is slower if the number of common factors K increases with T. Therefore, the thresholding value ω_T used in this paper is adjusted to account for the effect of the estimation of u_{it}.

2.2. *Asymptotic properties of the thresholding estimator.* Bickel and Levina (2008a) used a universal constant as the thresholding value. As pointed out by Rothman, Levina and Zhu (2009) and Cai and Liu (2011), when the variances of the entries of the sample covariance matrix vary over a wide range, it is more desirable to use thresholds that capture the variability of individual estimation. For this purpose, in this paper, we apply the adaptive thresholding estimator (Cai and Liu (2011)) to estimate the error covariance matrix, which is given by

$$\begin{aligned} \widehat{\mathbf{\Sigma}}_u^{\mathcal{T}} = (\widehat{\sigma}_{ij}^{\mathcal{T}}), \quad \widehat{\sigma}_{ij}^{\mathcal{T}} &= \widehat{\sigma}_{ij}I(|\widehat{\sigma}_{ij}| \geq \sqrt{\hat{\theta}_{ij}}\omega_T) \\ \hat{\theta}_{ij} &= \frac{1}{T}\sum_{t=1}^{T}\left(\hat{u}_{it}\hat{u}_{jt} - \widehat{\sigma}_{ij}\right)^2, \end{aligned} \tag{2.5}$$

for some ω_T to be specified later.

We impose the following assumption:

ASSUMPTION 2.1. *(i)* $(\mathbf{u}_1, ..., \mathbf{u}_T)$ *are independent and identically distributed 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 $r > 0$ *and* $b > 0$, *such that for any* $s > 0$ *and* $i \leq p$,

$$P(|u_{it}| > s) \leq \exp(-(s/b)^r). \tag{2.6}$$

Condition (ii) requires the nonsingularity of $\mathbf{\Sigma}_u$. Note that Cai and Liu (2011) allowed $\max_j \sigma_{jj}$ to diverge when direct observations are available. Condition (ii), however, requires that σ_{jj} should be uniformly bounded. In factor models, a uniform upper bound on the variance of u_{it} is needed when we estimate the covariance matrix of $\mathbf{y}_t$ later. This assumption is satisfied by most of the applications of factor models. Condition (iii) requires the distributions of $(u_{1t}, ..., u_{pt})$ to have exponential-type tails, which allows us to apply the large deviation theory to $\frac{1}{T}\sum_{t=1}^{T} u_{it}u_{jt}$.

Suppose there exists a positive sequence a_T such that

$$\max_{i \leq p} \frac{1}{T} \sum_{t=1}^{T} |u_{it} - \hat{u}_{it}|^2 = O_p(a_T^2). \tag{2.7}$$

The following theorem establishes the asymptotic properties of the thresholding estimator $\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}}$, based on observations with estimation errors.

THEOREM 2.1. *Let* $\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}}$ *be defined as in (2.5) with*

$$\omega_T = C \max\left\{\sqrt{\frac{\log p}{T}}, a_T\right\}$$

for some $C > 0$. *Assume* $\max_{i,t} |u_{it} - \hat{u}_{it}| = o_p(1)$, $a_T = o(1)$, *and* $(\log p)^{4/r-1} = o(T)$. *Then under Assumption 2.1,*
(i)

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

(ii) $\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}}$ *is positive definite, and*

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

Note that without thresholding, when $p > T$, the usual covariance matrix based on $\hat{u}_{ij}$ is singular. In contrast, after thresholding, the estimated error covariance matrix preserves the nonsingularity, and achieves a convergence rate that depends on the residual mean squared error. We will see in the next section that when the number of common factors K increases slowly, the convergence rate in Theorem 2.1 is close to the minimax optimal rate derived by Cai and Zhou (2010).

3. Estimation of Covariance Matrix Using Factors. We now investigate the estimation of the covariance matrix $\mathbf{\Sigma}$ in the approximate factor model:

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

where $\mathbf{\Sigma} = \mathrm{cov}(\mathbf{y}_t)$. This covariance matrix is particularly of interest in many applications of factor models as well as corresponding inferential theories.

Note that

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

By the Sherman-Morrison-Woodbury formula,

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

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

$$\widehat{\mathbf{b}}_i = \arg\min_{\mathbf{b}_i} \frac{1}{Tp} \sum_{t=1}^{T} \sum_{i=1}^{p} (y_{it} - \mathbf{b}_i'\mathbf{f}_t)^2.$$

The covariance matrix $\mathrm{cov}(\mathbf{f}_t)$ can be estimated by the sample covariance matrix

$$\widehat{\mathrm{cov}}(\mathbf{f}_t) = T^{-1}\mathbf{X}\mathbf{X}' - T^{-2}\mathbf{X}\mathbf{1}\mathbf{1}'\mathbf{X}',$$

where $\mathbf{X} = (\mathbf{f}_1, ..., \mathbf{f}_T)$, and $\mathbf{1}$ is a T-dimensional column vector of ones. Therefore, by employing the thresholding estimator $\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}}$ in (2.5), we obtain substitution estimators

$$\widehat{\mathbf{\Sigma}}^{\mathcal{T}} = \widehat{\mathbf{B}}\widehat{\mathrm{cov}}(\mathbf{f}_t)\widehat{\mathbf{B}}' + \widehat{\mathbf{\Sigma}}_u^{\mathcal{T}},$$

and

$$(\widehat{\mathbf{\Sigma}}^{\mathcal{T}})^{-1} = (\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1} - (\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1}\widehat{\mathbf{B}}[\widehat{\mathrm{cov}}(\mathbf{f}_t)^{-1} + \widehat{\mathbf{B}}'(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1}\widehat{\mathbf{B}}]^{-1}\widehat{\mathbf{B}}'(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1}.$$

Fan, Fan and Lv (2008) obtained an upper bound of $\|\widehat{\mathbf{\Sigma}}^{\mathcal{T}} - \mathbf{\Sigma}\|_F$ under the Frobenius norm when $\mathbf{\Sigma}_u$ is diagonal, i.e., there was no cross-sectional correlation among the idiosyncratic errors. In order for their upper bound to decrease to zero, $p^2 < T$ is required. Even with this restrictive assumption, they showed that the convergence rate is as the same as the usual sample covariance matrix of $\mathbf{y}_t$, though the latter does not take into account of the factor structure. Alternatively, they considered the entropy loss norm, proposed by James and Stein (1961):

$$\|\widehat{\mathbf{\Sigma}}^{\mathcal{T}} - \mathbf{\Sigma}\|_\Sigma = \left(p^{-1}\mathrm{tr}(\widehat{\mathbf{\Sigma}}^{\mathcal{T}}\mathbf{\Sigma}^{-1} - I)^2\right)^{1/2} = p^{-1/2}\|\mathbf{\Sigma}^{-1/2}(\widehat{\mathbf{\Sigma}}^{\mathcal{T}} - \mathbf{\Sigma})\mathbf{\Sigma}^{-1/2}\|_F.$$

Here the factor $p^{-1/2}$ is used for normalization, such that $\|\mathbf{\Sigma}\|_\Sigma = 1$. Under this norm, Fan, Fan and Lv (2008) showed that the substitution estimator has a better convergence rate than the usual sample covariance matrix. Note that the normalization factor $p^{-1/2}$ is essential in our high dimensional setting as it cancels out the diverging dimensionality introduced by p. Thanks to this normalization factor, the estimated covariance matrix $\widehat{\mathbf{\Sigma}}^{\mathcal{T}}$ is consistent even if $p > T$ under norm $\|.\|_\Sigma$.

The following assumptions are made.

ASSUMPTION 3.1. *(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.*

In addition to the conditions above, 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

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

The following strong mixing assumption enables us to apply the Bernstein's inequality in the technical proofs.

ASSUMPTION 3.2. *There exist positive constants* γ *and* C *such that for all* $t \in \mathbb{Z}^+$,

$$\alpha(t) \leq \exp(-Ct^\gamma).$$

In addition, we impose the following regularity conditions.

ASSUMPTION 3.3. *(i) There exists a constant* $M > 0$ *such that for all* i, k, $|b_{ik}| < M$, *and* $E\|\mathbf{f}_t\|^4 \leq K^2M$.
(ii) There exist $r_2 > 0$ *and* $b_2 > 0$ *such that for any* $s > 0$ *and* $i \leq K$,

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

Condition (ii) allows us to apply the Bernstein type inequality for weakly dependent data.

ASSUMPTION 3.4. *There exist constants* $C_1 > 0$ *and* $C_2 > 0$ *such that* $\lambda_{\min}(\mathbf{\Sigma}) > C_1$, *and* $C_1 \leq \lambda_{\min}(\mathrm{cov}(\mathbf{f}_t)) \leq C_2$.

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

Assumption 3.4 ensures that $\mathbf{\Sigma}$ and $\mathrm{cov}(\mathbf{f}_t)$ are not ill-conditioned, which is needed to derive the convergence rate of $\|(\widehat{\mathbf{\Sigma}}^{\mathcal{T}})^{-1} - \mathbf{\Sigma}^{-1}\|$ below. Assumption 3.5 requires that the factors should be pervasive, i.e., impact every individual time series (Harding (2009)). It was imposed by Fan, Fan and Lv (2008) only when they tried to establish the asymptotic normality of the covariance estimator. However, it turns out to be also helpful to obtain a good upper bound of $\|(\widehat{\mathbf{\Sigma}}^{\mathcal{T}})^{-1} - \mathbf{\Sigma}^{-1}\|$, as it ensures that $\lambda_{\max}((\mathbf{B}'\mathbf{\Sigma}^{-1}\mathbf{B})^{-1}) = O(p^{-1})$.

The first result in this Section is an application of Theorem 2.1.

THEOREM 3.1. *Suppose* $\max\{(\log p)^{2/\gamma+2/r_2+4/r-1}, K^4(\log p)^2\} = o(T)$. *Under Assumptions 2.1, 3.1-3.3, the adaptive thresholding estimator defined in (2.5) with* $\omega_T^2 = \frac{K^2 \log p}{T}$ *satisfies*

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

and

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

Remarks We briefly provide a description of those terms in the convergence rate above.

1. The term K appears as an effect of using the estimated residuals to construct the thresholding covariance estimator, which is typically small compared to p and T in many applications. For instance, the famous Fama-French-three-factor model shows that $K = 3$ factors are adequate for the US equity market. In an empirical study on asset returns, Bai and Ng (2002) used the monthly data which contains the returns of 4883 stocks for sixty months. For their data set, $T = 60$, $p = 4883$. Bai and Ng (2002) determined $K = 2$ common factors.
2. As in Bickel and Levina (2008a) and Cai and Liu (2011), m_T, the maximum number of nonzero components across the rows of Σ_u, also plays a role in the convergence rate. Note that when K is bounded, the convergence rate is the same as the minimax rate derived by Cai and Zhou (2010).

Combining with the estimated low-rank matrix $\mathbf{B}\mathrm{cov}(\mathbf{f}_t)\mathbf{B}'$, Theorem 3.1 implies the main theorem in this section:

THEOREM 3.2. *Suppose* $\max\{(\log p)^{2/\gamma+2/r_2+4/r-1}, K^4(\log p)^2\} = o(T)$. *Under Assumptions 2.1, 3.1-3.5, we have*

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

$$\|(\widehat{\mathbf{\Sigma}}^{\mathcal{T}})^{-1} - \mathbf{\Sigma}^{-1}\|^2 = O_p\left(\frac{m_T^2 K^2 \log p}{T}\right), \tag{3.1}$$

and

$$\|\widehat{\mathbf{\Sigma}}^{\mathcal{T}} - \mathbf{\Sigma}\|_\infty^2 = O_p\left(\frac{K^6 \log p}{T}\right). \tag{3.2}$$

Note that we have derived a better convergence rate of $(\widehat{\mathbf{\Sigma}}^{\mathcal{T}})^{-1}$ than that in Fan, Fan and Lv (2008). When the operator norm is considered, p is allowed to grow exponentially fast in T in order for $(\widehat{\mathbf{\Sigma}}^{\mathcal{T}})^{-1}$ to be consistent.

We have also derived the maximum elementwise estimation $\|\widehat{\mathbf{\Sigma}}^{\mathcal{T}} - \mathbf{\Sigma}\|_\infty$. This quantity appears in risk assessment as in Fan, Zhang and Yu (2008). For any portfolio with 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

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

where $\|\mathbf{w}\|_1$, the l_1 norm of $\mathbf{w}$, is the gross exposure of the portfolio.

4. Extension: Seemingly Unrelated Regression. A *seemingly unrelated regression* model (Kmenta and Gilbert (1970)) is a set of linear equations in which the disturbances are correlated across equations. Specifically, we have

$$y_{it} = \mathbf{b}_i'\mathbf{f}_{it} + u_{it}, \quad i \le p, t \le T, \tag{4.1}$$

where $\mathbf{b}_i$ and $\mathbf{f}_{it}$ are both $K_i \times 1$ vectors. The p linear equations (4.1) are related because their error terms u_{it} are correlated, i.e., the covariance matrix

$$\mathbf{\Sigma}_u = (Eu_{it}u_{jt})_{p\times p}$$

is not diagonal.

Model (4.1) allows each variable y_{it} to have its own factors. This is important for many applications. In financial applications, the returns of individual stock depend on common market factors and sector-specific factors. In housing price index modeling, housing price appreciations depend on both national factors and local economy. When $\mathbf{f}_{it} = \mathbf{f}_t$ for each $i \le p$, model (4.1) reduces to the approximate factor model (1.1) with common factors $\mathbf{f}_t$.

Under mild conditions, running OLS on each equation produces unbiased and consistent estimator of $\mathbf{b}_i$ separately. However, since OLS does not take into account the cross sectional correlation among the noises, it is not efficient. Instead, statisticians obtain the best linear unbiased estimator (BLUE) via generalized least square (GLS). Write

$$\mathbf{y}_i = (y_{i1}, ..., y_{iT})', T \times 1, \quad \mathbf{X}_i = (\mathbf{f}_{i1}, ..., \mathbf{f}_{iT})', T \times K_i, \quad i \le p,$$

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_p \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mathbf{X}_p \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_p \end{pmatrix}.$$

The GLS estimator of $\mathbf{B}$ is given by Zellner (1962):

$$\widehat{\mathbf{B}}_{GLS} = [\mathbf{X}'(\widehat{\mathbf{\Sigma}}_u^{-1} \otimes I_T)^{-1}\mathbf{X}]^{-1}[\mathbf{X}'(\widehat{\mathbf{\Sigma}}_u^{-1} \otimes I_T)^{-1}\mathbf{y}], \tag{4.2}$$

where I_T denotes a $T \times T$ identity matrix, $\otimes$ represents the Kronecker product operation, and $\widehat{\mathbf{\Sigma}}_u$ is a consistent estimator of $\mathbf{\Sigma}_u$.

In classical seemingly unrelated regression in which p does not grow with T, $\mathbf{\Sigma}_u$ is estimated by a two-stage procedure: (Kmenta and Gilbert (1970)): On the first stage, estimate $\mathbf{B}$ via OLS, and obtain residuals

$$\hat{u}_{it} = y_{it} - \widehat{\mathbf{b}}_i' \mathbf{f}_{it}. \tag{4.3}$$

On the second stage, estimate $\mathbf{\Sigma}_u$ by

$$\widehat{\mathbf{\Sigma}}_u = (\widehat{\sigma}_{ij}) = (\frac{1}{T}\sum_{t=1}^{T} \hat{u}_{it}\hat{u}_{jt})_{p\times p}. \tag{4.4}$$

In high dimensional seemingly unrelated regression in which $p > T$, however, $\widehat{\mathbf{\Sigma}}_u$ is not invertible, and hence the GLS estimator (4.2) is infeasible.

By the sparsity assumption of $\mathbf{\Sigma}_u$, we can deal with this singularity problem by using the adaptive thresholding estimator, and produce a consistent nonsingular estimator of $\mathbf{\Sigma}_u$. To pursue this goal, we impose the following assumptions:

ASSUMPTION 4.1. *For each $i \leq p$,*
(i) $\{\mathbf{f}_{it}\}_{t\geq 1}$ is stationary and ergodic.
(ii) $\{\mathbf{u}_t\}_{t\geq 1}$ and $\{\mathbf{f}_{it}\}_{t\geq 1}$ are independent.

ASSUMPTION 4.2. *There exists positive constants C and γ such that for each $i \leq p$, the strong mixing condition in Assumption 3.2 is satisfied by $(\mathbf{f}_{it}, \mathbf{u}_t)$.*

ASSUMPTION 4.3. *There exist constants $M > 0$ and $C > 0$ such that for all i, $E\|\mathbf{f}_{it}\|^4 < K_i^2 M$, and $\min_{i\leq p} \lambda_{\min}(\mathrm{cov}(\mathbf{f}_{it})) > C$.*

ASSUMPTION 4.4. *There exist $r_3 > 0$ and $b_3 > 0$ such that for any $s > 0$ and i, j,*

$$P(|f_{it,j}| > s) \leq \exp(-(s/b_3)^{r_3}).$$

These assumptions are similar to those made in Section 3, except that here they are imposed on the sector-specific factors. The main theorem in this section is a direct application of Theorem 2.1. It shows that the adaptive thresholding technique (2.5) produces a consistent nonsingular estimator of $\widehat{\mathbf{\Sigma}}_u$.

THEOREM 4.1. *Suppose* $\max\{(\log p)^{2/\gamma+2/r_3+4/r-1}, K^4(\log p)^2\} = o(T)$, *where* $K = \max_{i\leq p} K_i$. *Under Assumptions 2.1, 4.1-4.4, the adaptive thresholding estimator defined in (4.4) and (2.5) with* $\omega_T^2 = \frac{K^2 \log p}{T}$ *satisfies*

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

and

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

Therefore, in the case when $p > T$, Theorem 4.1 enables us to efficiently estimate $\mathbf{B}$ via feasible GLS:

$$\widehat{\mathbf{B}}_{GLS}^{\mathcal{T}} = [\mathbf{X}'((\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1} \otimes I_T)^{-1}\mathbf{X}]^{-1}[\mathbf{X}'((\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1} \otimes I_T)^{-1}\mathbf{y}].$$

5. Monte Carlo Experiments. In this section, we use simulation to demonstrate the rates of convergence of $\widehat{\mathbf{\Sigma}}^{\mathcal{T}}$ and $(\widehat{\mathbf{\Sigma}}^{\mathcal{T}})^{-1}$. The simulation model is a modified version of the Fama-French-three-factor model described in Fan, Fan, Lv (2008). We fix the number of factors, $K = 3$ and the length of time, $T = 500$, and let the dimensionality p gradually increase.

The Fama-French three-factor model (Fama and French (1992)) is given by

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

which models the excess return (real rate of return minus risk-free rate) of the *i*th stock of a portfolio, y_{it}, with respect to 3 factors. The first factor is the excess return of the whole stock market, and the weighted excess return on all NASDAQ, AMEX and NYSE stocks is a commonly used proxy. It extends the capital assets pricing model (CAPM) by adding two new factors- SMB ("small minus big" cap) and HML ("high minus low" book/price). These two were added to the model after the observation that two types of stocks - small caps, and high book value to price ratio, tend to outperform the stock market as a whole.

We separate this section into three parts, calibration, simulation and results. Similar to Section 5 of Fan, Fan and Lv (2008), in the calibration part we want to calculate realistic multivariate distributions from which we can generate the factor loadings $\mathbf{B}$, idiosyncratic noises $\{\mathbf{u}_t\}_{t=1}^T$ and the observable factors $\{\mathbf{f}_t\}_{t=1}^T$. The data was obtained from the data library of Kenneth French's website.

5.1. *Calibration.* To estimate the parameters in the Fama-French model, we will use the two-year daily data $(\tilde{\mathbf{y}}_t, \tilde{\mathbf{f}}_t)$ from Jan 1^{st}, 2009 to Dec 31^{st}, 2010 (T=500) of 30 industry portfolios.

1. Calculate the least squares estimator $\tilde{\mathbf{B}}$ of $\tilde{\mathbf{y}}_t = \mathbf{B}\tilde{\mathbf{f}}_t + \mathbf{u}_t$, and take the rows of $\tilde{\mathbf{B}}$, namely $\tilde{\mathbf{b}}_1 = (b_{11}, b_{12}, b_{13})$,...,$\tilde{\mathbf{b}}_{30} = (b_{30,1}, b_{30,2}, b_{30,3})$, to calculate the sample mean vector $\boldsymbol{\mu}_B$ and sample covariance matrix $\boldsymbol{\Sigma}_B$. We then create a mutlivariate normal distribution $N_3(\boldsymbol{\mu}_B, \boldsymbol{\Sigma}_B)$, from which the factor loadings $\{\mathbf{b}_i\}_{i=1}^p$ are drawn.

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

$\boldsymbol{\mu}_B$	$\boldsymbol{\Sigma}_B$		
1.0641	0.0475	0.0218	0.0488
0.1233	0.0218	0.0945	0.0215
-0.0119	0.0488	0.0215	0.1261

2. For each fixed p, create the sparse matrix $\boldsymbol{\Sigma}_u = \mathbf{D} + \mathbf{ss}' - \text{diag}\{s_1^2, ..., s_p^2\}$ in the following way. Let $\widehat{\mathbf{u}}_t = \tilde{\mathbf{y}}_t - \tilde{\mathbf{B}}\tilde{\mathbf{f}}_t$. For $i = 1, ..., 30$, let $\widehat{\sigma}_i$ denote the standard deviation of the residuals of the ith portfolio. We find $\min(\hat{\sigma}_i) = 0.3533$, $\max(\hat{\sigma}_i) = 1.5222$, and calculate the mean and the standard deviation of the $\widehat{\sigma}_i$'s, namely $\bar{\sigma} = 0.6055$ and $\sigma_{SD} = 0.2621$.
   Let $\mathbf{D} = \text{diag}\{\sigma_1^2, ..., \sigma_p^2\}$, where $\sigma_1, ..., \sigma_p$ are generated independently from the Gamma distribution $G(\alpha, \beta)$, with mean $\alpha\beta$ and standard deviation $\alpha^{1/2}\beta$. We match these values to $\bar{\sigma} = 0.6055$ and $\sigma_{SD} = 0.2621$, to get $\alpha = 5.6840$ and $\beta = 0.1503$. Further, we create a loop that only accepts the value of σ_i if it is between $\min(\hat{\sigma}_i) = 0.3533$ and $\max(\hat{\sigma}_i) = 1.5222$.
   Create $\mathbf{s} = (s_1, ..., s_p)'$ to be a sparse vector. We set each $s_i \sim N(0, 1)$ with probability $\frac{0.2}{\sqrt{p}\log p}$, and $s_i = 0$ otherwise. This leads to an average of $\frac{0.2\sqrt{p}}{\log p}$ nonzero elements per each row of the error covariance matrix.
   Create a loop that generates $\boldsymbol{\Sigma}_u$ multiple times until it is positive definite.
3. Assume the factors follow the vector autoregressive model (VAR(1)) model $\mathbf{f}_t = \boldsymbol{\mu} + \boldsymbol{\Phi}\mathbf{f}_{t-1} + \boldsymbol{\varepsilon}_t$ for some 3×3 matrix $\boldsymbol{\Phi}$, where $\boldsymbol{\varepsilon}_t$'s are i.i.d. $N_3(0, \boldsymbol{\Sigma}_\epsilon)$. We estimate $\boldsymbol{\Phi}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_\epsilon$ from the data, and obtain $\text{cov}(\mathbf{f}_t)$.

TABLE 2
Parameters of $\mathbf{f}_t$ *generating process*

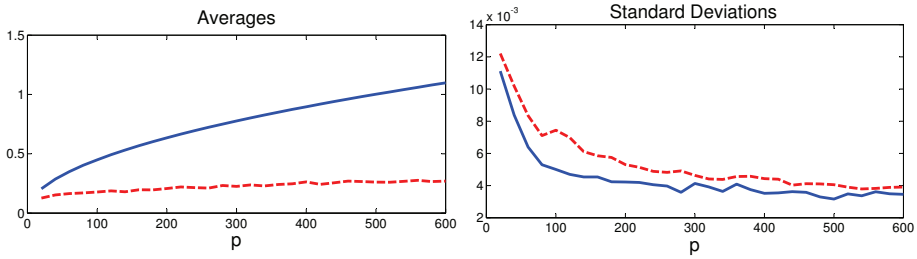
$\boldsymbol{\mu}$	$\text{cov}(\mathbf{f}_t)$			$\boldsymbol{\Phi}$		
0.1074	2.2540	0.2735	0.9197	-0.1149	0.0024	0.0776
0.0357	0.2735	0.3767	0.0430	0.0016	-0.0162	0.0387
0.0033	0.9197	0.0430	0.6822	-0.0399	0.0218	0.0351

5.2. *Simulation.* We keep $T = 500$ fixed, and gradually increase p from 20 to 600 in multiples of 20 to illustrate the rates of convergence when the number of variables diverges with respect to the sample size. The number of simulations for each fixed p is 200. Specifically, at each simulation, we do the following:

1. Generate $\{\mathbf{b}_i\}_{i=1}^p$ independently from $N_3(\boldsymbol{\mu}_B, \boldsymbol{\Sigma}_B)$, and set $\mathbf{B} = (\mathbf{b}_1, ..., \mathbf{b}_p)'$.
2. Generate $\{\mathbf{u}_t\}_{t=1}^T$ independently from $N_p(0, \boldsymbol{\Sigma}_u)$.
3. Generate $\{\mathbf{f}_t\}_{t=1}^T$ from the VAR(1) model $\mathbf{f}_t = \boldsymbol{\mu} + \boldsymbol{\Phi}\mathbf{f}_{t-1} + \boldsymbol{\varepsilon}_t$.
4. Calculate $\mathbf{y}_t = \mathbf{B}\mathbf{f}_t + \mathbf{u}_t$ for $t = 1, ..., T$.
5. Set $\omega_T = 0.10K\sqrt{\log p/T}$ to obtain the thresholding estimator (2.5) $\widehat{\boldsymbol{\Sigma}}_u^{\mathcal{T}}$ and compute the relevant estimators.

We graph the distance of $\widehat{\boldsymbol{\Sigma}}^{\mathcal{T}}$ and $\widehat{\boldsymbol{\Sigma}}_{sam}$ to $\boldsymbol{\Sigma}$, under the entropy-loss norm $\|.\|_\Sigma$ and the L-infinity norm $\|.\|_\infty$. We also graph the distance of the inverses $(\widehat{\boldsymbol{\Sigma}}^{\mathcal{T}})^{-1}$ and $\widehat{\boldsymbol{\Sigma}}_{sam}^{-1}$ to $\boldsymbol{\Sigma}^{-1}$ under the operator norm. Note that we graph that only for p from 20 to 300. Since $T = 500$, for $p > 500$ the sample covariance matrix is singular. Also, for p close to 500, $\widehat{\boldsymbol{\Sigma}}_{sam}$ is nearly singular, which leads to abnormally large values of the operator norm. We also record the standard deviations of these norms.

FIG 1. *Averages and standard deviations of* $\|\widehat{\boldsymbol{\Sigma}}^{\mathcal{T}} - \boldsymbol{\Sigma}\|_\Sigma$ *(dashed curve) and* $\|\widehat{\boldsymbol{\Sigma}}_{sam} - \boldsymbol{\Sigma}\|_\Sigma$ *(solid curve) over 200 iterations, as a function of the dimensionality p.*



5.3. *Results.* In Figures 1-3, the dashed curves correspond to $\widehat{\boldsymbol{\Sigma}}^{\mathcal{T}}$ and the solid curves correspond to the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}_{sam}$. Figure 1 and 2 presents the averages and standard deviations of the estimation error of both of these matrices with respect to the Σ-norm and infinity norm, respectively. Figure 3 presents the averages and sandard deviations of estimation errors of the inverses with respect to the operator norm. Based on the simulation results, we observe the following:

1. The standard deviations of the norms are negligible when compared to their corresponding averages.

FIG 2. *Averages and standard deviations of* $\|\widehat{\mathbf{\Sigma}}^{\mathcal{T}} - \mathbf{\Sigma}\|_\infty$ *(dashed curve) and* $\|\widehat{\mathbf{\Sigma}}_{sam} - \mathbf{\Sigma}\|_\infty$ *(solid curve) over 200 iterations, as a function of the dimensionality p.*

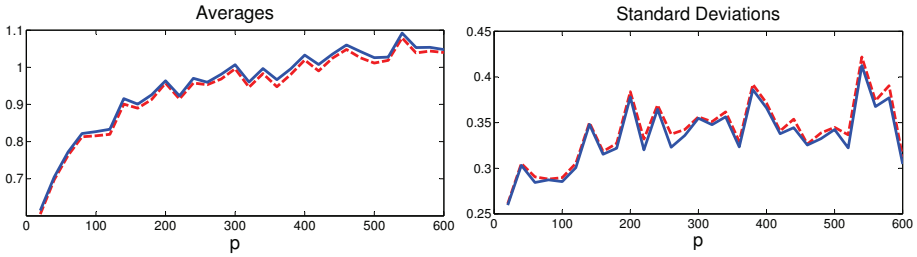
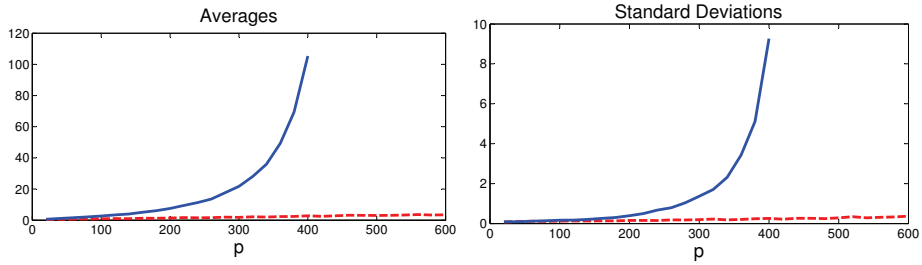


FIG 3. *Averages and standard deviations of* $\|(\widehat{\mathbf{\Sigma}}^{\mathcal{T}})^{-1} - \mathbf{\Sigma}^{-1}\|$ *(dashed curve) and* $\|\widehat{\mathbf{\Sigma}}_{sam}^{-1} - \mathbf{\Sigma}^{-1}\|$ *(solid curve) over 200 iterations, as a function of the dimensionality p.*



2. Under the $\|.\|_\Sigma$, our estimate of the covariance matrix of $\mathbf{y}$, $\widehat{\mathbf{\Sigma}}^{\mathcal{T}}$ performs much better than the sample covaraince matrix $\widehat{\mathbf{\Sigma}}_{sam}$. Note that, in the proof of Theorem 2 in Fan, Fan, Lv(2008), it was shown that:

$$\|\widehat{\mathbf{\Sigma}}_{sam} - \mathbf{\Sigma}\|_\Sigma^2 = O_p\left(\frac{K^3}{Tp}\right) + O_p\left(\frac{p}{T}\right) + O_p\left(\frac{K^{3/2}}{T}\right). \tag{5.1}$$

For a small fixed value of K, such as $K = 3$, the dominating term in (5.1) is $O\left(\frac{p}{T}\right)$. From Theorem 3.2, and given that $m_T = o(p^{1/4})$, the dominating term in the convergence of $\|\widehat{\mathbf{\Sigma}}^{\mathcal{T}} - \mathbf{\Sigma}\|_\Sigma^2$ is $O_p\left(\frac{p}{T^2} + \frac{m_T^2 \log p}{T}\right)$. So, we would expect out estimator to perform better, and the simulation results are consistent with the theory.

3. Under the infinity norm, both estimators perform roughly the same. This is to be expected, given that the thresholding affects mainly the elements of the covariance matrix that are closest to 0, and the infinity norm depicts the magnitude of the largest elementwise absolute error.
4. Under the operator norm, the inverse of our estimator, $(\widehat{\mathbf{\Sigma}}^{\mathcal{T}})^{-1}$ also outperforms significantly the inverse of the sample covariance matrix.
5. Finally, when $p > 500$, the thresholding estimators $\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}}$ and $\widehat{\mathbf{\Sigma}}^{\mathcal{T}}$ are still nonsingular.

In conclusion, even after imposing less restrictive assumptions on the error covariance matrix, we still reach an estimator $\widehat{\mathbf{\Sigma}}^{\mathcal{T}}$ that significantly outperforms the standard sample covariance matrix.

6. Conclusions and Discussions. We studied the rate of convergence of high dimensional covariance matrix of approximate factor models under various norms. By assuming sparse error covariance matrix, we allow for the presence of the cross-sectional correlation even after taking out common factors. Since direct observations of the noises are not available, we constructed the error sample covariance matrix first based on the estimation residuals, and then estimate the error covariance matrix using the adaptive thresholding method. We then constructed the covariance matrix of $\mathbf{y}_t$ using the factor model, assuming that the factors follow a stationary and ergodic process, but can be weakly-dependent. It was shown that after thresholding, the estimated covariance matrices are still invertible even if $p > T$, and the rate of convergence of $(\widehat{\mathbf{\Sigma}}^{\mathcal{T}})^{-1}$ and $(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1}$ is of order $O_p(Km_T\sqrt{\log p/T})$, where K comes from the impact of estimating the unobservable noise terms. This demonstrates when estimating the inverse covariance matrix, p is allowed to be much larger than T.

The rate of convergence in Theorem 2.1 reflects the impact of unobservable idiosyncratic components on the thresholding method. Generally, whether it is the minimax rate when direct observations are not available but have to be estimated is an interesting question, which is left as a research direction in the future.

Moreover, this paper uses the hard-thresholding technique, which takes the form of $\widehat{\sigma}_{ij}(\sigma_{ij}) = \sigma_{ij}I(|\sigma_{ij}| > \theta_{ij})$ for some pre-determined threshold θ_{ij}. Recently, Rothman et al (2009) and Cai and Liu (2011) studied a more general thresholding function $\widehat{\sigma}_{ij}(\theta_{ij}) = s(\sigma_{ij})$ introduced in Antoniadis and Fan (2001) for covariance matrix estimation, which also allows for soft-thresholdings, i.e., $s(.)$ is a continuous function. 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 the same.

Finally, we considered the case when common factors are observable, as in Fama and French (1992). In some applications, the common factors are unobservable

and need to be estimated (Bai (2003)). In that case, it is still possible to consistently estimate the covariance matrices using similar techniques as those in this paper. However, the impact of high dimensionality on the rate of convergence also comes from the estimation error of the unobservable factors. We plan to address this problem in a separate paper.

APPENDIX A: PROOFS FOR SECTION 2

A.1. Lemmas. The following lemmas are useful to be proved first, in which we consider the operator norm $\|\mathbf{A}\|^2 = \lambda_{\max}(\mathbf{A}'\mathbf{A})$.

LEMMA A.1. *Suppose that* $\mathbf{A}$ *and* $\mathbf{B}$ *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*

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

PROOF. Suppose both $\mathbf{A}$ and $\mathbf{B}$ are $m \times m$. For any $\mathbf{v} \in \mathbb{R}^m$ such that $\|\mathbf{v}\| = 1$, $|\mathbf{v}'(\mathbf{A} - \mathbf{B})\mathbf{v}| \leq \|\mathbf{v}\|^2 \cdot \|\mathbf{A} - \mathbf{B}\| = o_p(c_T)$. Hence for all large T, $\mathbf{v}'\mathbf{A}\mathbf{v} \geq \mathbf{v}'\mathbf{B}\mathbf{v} - 0.5c_T \geq \lambda_{\min}(\mathbf{B}) - 0.5c_T > 0.5c_T$. Hence, $\lambda_{\min}(\mathbf{A}) \geq 0.5c_T$. In addition,

$$\begin{aligned} \|\mathbf{A}^{-1} - \mathbf{B}^{-1}\| &= \|\mathbf{A}^{-1}(\mathbf{B} - \mathbf{A})\mathbf{B}^{-1}\| \\ &\leq \lambda_{\min}(\mathbf{A})^{-1}\|\mathbf{A} - \mathbf{B}\|\lambda_{\min}(\mathbf{B})^{-1} \\ &= O_p(c_T^{-2})\|\mathbf{A} - \mathbf{B}\|. \end{aligned}$$

Q.E.D.

LEMMA A.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* $\forall s > 0$,

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

Then with $r_3 = r_1 r_2/(r_1 + r_2)$, *for some* $b_3 > 0$, *and any* $s > 0$,

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

PROOF. We have, for any $s > 0$, $M = (s b_2^{r_2/r_1}/b_1)^{r_1/(r_1+r_2)}$, $b = b_1 b_2$, and $r = r_1 r_2/(r_1 + r_2)$,

$$\begin{aligned} P(|Z_1 Z_2| > s) &\leq P(M|Z_1| > s) + P(|Z_2| > M) \\ &\leq \exp(1 - (s/b_1 M)^{r_1}) + \exp(1 - (M/b_2)^{r_2}) \\ &= 2\exp(1 - (s/b)^r). \end{aligned}$$

Pick up an $r_3 \in (0, r)$, and $b_3 > \max\{(r_3/r)^{1/r}b, (1+\log 2)^{1/r}b\}$, then it can be shown that $F(s) = (s/b)^r - (s/b_3)^{r_3}$ is increasing when $s > b_3$. Hence $F(s) > F(b_3) > \log 2$ when $s > b_3$, which implies when $s > b_3$,

$$P(|Z_1 Z_2| > s) \leq 2\exp(1 - (s/b)^r) \leq \exp(1 - (s/b_3)^{r_3}).$$

When $s \leq b_3$,

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

Q.E.D.

LEMMA A.3. *Under Assumption 2.1, $a_T = o(1)$ and $\log p = o(T)$,*
(i)

$$\max_{i,j\leq p} |\frac{1}{T}\sum_{t=1}^{T} u_{it}u_{jt} - \sigma_{ij}| = O_p(\sqrt{\frac{\log p}{T}}).$$

(ii)

$$\max_{i,j\leq p} |\widehat{\sigma}_{ij} - \sigma_{ij}| = O_p(\max\{\sqrt{\frac{\log p}{T}}, a_T\}),$$

PROOF. (i) By Assumption 2.1(iii) and Lemma A.2, $u_{it}u_{jt}$ satisfies the exponential tail condition. Thus by Theorem 1 of Merlevède (2009), there exist constants C_1, C_2, C_3, C_4 and $C_5 > 0$ that only depend on b and r such that for any $i, j \leq p$, and $\gamma = r/4$ (where r is defined in Assumption 2.1(iii)),

$$P(|\frac{1}{T}\sum_{t=1}^{T} u_{it}u_{jt} - \sigma_{ij}| \geq s) \leq T\exp\left(-\frac{(Ts)^\gamma}{C_1}\right) + \exp\left(-\frac{T^2 s^2}{C_2(1+TC_3)}\right)$$
$$+ \exp\left(-\frac{(Ts)^2}{C_4 T}\exp\left(\frac{(Ts)^{\gamma(1-\gamma)}}{C_5(\log Ts)^\gamma}\right)\right).$$

Using Bonferroni's method, we have

$$P(\max_{i,j\leq p} |\frac{1}{T}\sum_{t=1}^{T} u_{it}u_{jt} - \sigma_{ij}| > s) \leq p^2 \max_{i,j\leq p} P(|\frac{1}{T}\sum_{t=1}^{T} u_{it}u_{jt} - \sigma_{ij}| > s).$$

Since $(\log p)^{4/r-1} = o(T)$, as long as $s > \sqrt{(\log p)/T}$, for all large T,

$$p^2 T\exp\left(-\frac{(Ts)^\gamma}{C_1}\right) + p^2\exp\left(-\frac{(Ts)^2}{C_4 T}\exp\left(\frac{(Ts)^{r(1-r)}}{C_5(\log Ts)^r}\right)\right) = o(1).$$

In addition, as long as $s^2T > 6C_2C_3 \log p$, for all large T,

$$p^2 \exp\left(-\frac{T^2 s^2}{C_2(1+TC_3)}\right) = o(1),$$

which implies the desired result.

(ii) We have, by part (i) and the triangular inequality,

$$\max_{i,j\leq p} |\widehat{\sigma}_{ij} - \sigma_{ij}| \leq O_p(\sqrt{\frac{\log p}{T}}) + \max_{i,j\leq p} |\frac{1}{T}\sum_{t=1}^{T}(\hat{u}_{it}\hat{u}_{jt} - u_{it}u_{jt})|.$$

We now show that $A \equiv \max_{i,j\leq p} |\frac{1}{T}\sum_{t=1}^{T}(\hat{u}_{it}\hat{u}_{jt} - u_{it}u_{jt})| = O_p(a_T)$. By the triangular and Cauchy Schwarz inequalities,

$$\begin{aligned}
A &\leq \max_{i,j\leq p} |\frac{1}{T}\sum_{t=1}^{T}(\hat{u}_{it} - u_{it})(\hat{u}_{jt} - u_{jt})| + 2\max_{i,j\leq p} |\frac{1}{T}\sum_{t=1}^{T} u_{it}(\hat{u}_{jt} - u_{jt})| \\
&\leq \max_{i\leq p} \frac{1}{T}\sum_{t=1}^{T}(\hat{u}_{it} - u_{it})^2 + 2\sqrt{\max_{i\leq p} \frac{1}{T}\sum_{t=1}^{T} u_{it}^2}\sqrt{\max_{i\leq p} \frac{1}{T}\sum_{t=1}^{T}(\hat{u}_{it} - u_{it})^2} \\
&= O_p(a_T^2) + 2\sqrt{o_p(1) + \max_{i\leq p}\sigma_{ii}}\sqrt{a_T^2} \\
&= O_p(a_T).
\end{aligned}$$

Hence the desired result follows. Q.E.D.

LEMMA A.4. *There exist $C_1, C_2 > 0$ such that with probability approaching one,*

$$C_1 \leq \min_{i,j} \hat{\theta}_{ij} \leq \max_{i,j} \hat{\theta}_{ij} \leq C_2.$$

PROOF. (i) For any i, j, by adding and subtracting terms, we have

$$\begin{aligned}
\hat{\theta}_{i,j} &= \frac{1}{T}\sum_t (\hat{u}_{it}\hat{u}_{jt} - \frac{1}{T}\sum_l \hat{u}_{il}\hat{u}_{jl})^2 \\
&\leq \frac{2}{T}\sum_t (\hat{u}_{it}\hat{u}_{jt} - \sigma_{ij})^2 + 2\max_{i,j}(\sigma_{ij} - \frac{1}{T}\sum_l \hat{u}_{il}\hat{u}_{jl})^2 \\
&\leq \frac{2}{T}\sum_t (\hat{u}_{it}\hat{u}_{jt} - \sigma_{ij})^2 + o_p(1),
\end{aligned}$$

where the term $o_p(1)$ does not depend on i, j, by Lemma A.3. Still by adding and subtracting terms,

$$\sum_t (\hat{u}_{it}\hat{u}_{jt} - \sigma_{ij})^2$$

$$\begin{aligned}
&\leq 4\sum_t(\hat{u}_{it}-u_{it})^2\hat{u}_{jt}^2+4\sum_t(\hat{u}_{jt}-u_{jt})^2u_{it}^2+2\sum_t(u_{it}u_{jt}-\sigma_{ij})^2\\
&\leq 4\max_{it}|\hat{u}_{it}-u_{it}|^2(2\max_j\sum_t(\hat{u}_{jt}-u_{jt})^2+3\max_j\sum_t u_{jt}^2)+2\sum_t(u_{it}u_{jt}-\sigma_{ij})^2\\
&= o_p(1)(o_p(Ta_T^2)+\max_j\sum_t u_{jt}^2))+2\sum_t(u_{it}u_{jt}-\sigma_{ij})^2.
\end{aligned}$$

Since $(\mathbf{u}_t)_{t\geq 1}$ are i.i.d. random vectors with exponential tail on each component, the same arguments as those in the proof of Lemma 2 in Cai and Liu (2011) imply that

$$\max_{i,j}|\frac{1}{T}\sum_t(u_{it}u_{jt}-\sigma_{ij})^2-\mathrm{var}(u_{it}u_{jt})|=o_p(1),$$

and $\mathrm{var}(u_{it}u_{jt})$ is bounded away from both zero and infinity. Therefore, $\frac{1}{T}\sum_t(u_{it}u_{jt}-\sigma_{ij})^2$ is bounded away from zero and infinity with probability approaching one. In addition, by Lemma A.3(i), with probability approaching one,

$$\max_j\frac{1}{T}\sum_t u_{jt}^2\leq o_p(1)+\max_j\sigma_{jj}\leq 2\max_j\sigma_{jj}.$$

In summary, $\max_{ij}\hat{\theta}_{ij}$ is bounded away from infinity with probability approaching one.

(ii) By adding and subtracting terms, we obtain

$$\begin{aligned}
&\sum_t(u_{it}u_{jt}-\sigma_{ij})^2\\
&\leq 4\sum_t(u_{it}u_{jt}-\hat{u}_{it}\hat{u}_{jt})^2+4\sum_t(\hat{u}_{it}\hat{u}_{jt}-\frac{1}{T}\sum_l\hat{u}_{il}\hat{u}_{jl})^2\\
&\quad +4\sum_t(\sigma_{ij}-\frac{1}{T}\sum_l\hat{u}_{il}\hat{u}_{jl})^2\\
&\leq 8\sum_t u_{it}^2(u_{jt}-\hat{u}_{jt})^2+8\sum_t\hat{u}_{jt}^2(u_{it}-\hat{u}_{it})^2+4T\hat{\theta}_{ij}+o_p(T)\\
&\leq 16\max_{it}|\hat{u}_{it}-u_{it}|^2(\max_j\sum_t(\hat{u}_{jt}-u_{jt})^2+\max_j\sum_t u_{jt}^2)+4T\hat{\theta}_{ij}+o_p(T),
\end{aligned}$$

where $o_p(T)$ does not depend on i,j due to Lemma A.3. As is demonstrated in part (i),

$$16\max_{it}|\hat{u}_{it}-u_{it}|^2(\max_j\sum_t(\hat{u}_{jt}-u_{jt})^2+\max_j\sum_t u_{jt}^2)=o_p(T),$$

and

$$\frac{1}{T}\sum_t(u_{it}u_{jt}-\sigma_{ij})^2\geq C$$

uniformly in i,j for some $C>0$ with probability approaching one. This establishes the result. Q.E.D.

A.2. Proof of Theorem 2.1.

PROOF. (i) For the operator norm, the triangular inequality still holds:

$$\|\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}} - \mathbf{\Sigma}_u\| \leq \|\mathbf{\Sigma}_u^{\mathcal{T}} - \mathbf{\Sigma}_u\| + \|\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}} - \mathbf{\Sigma}_u^{\mathcal{T}}\|,$$

where

$$\mathbf{\Sigma}_u^{\mathcal{T}} = (\sigma_{ij}^{\mathcal{T}}), \quad \sigma_{ij}^{\mathcal{T}} = \sigma_{ij} I(|\sigma_{ij}| > \sqrt{\hat{\theta}_{ij}}\omega_T).$$

and $\omega_T = C\max(\sqrt{\log p/T}, a_T)$ for some $C > 0$. We bound $\|\mathbf{\Sigma}_u^{\mathcal{T}} - \mathbf{\Sigma}_u\|$ and $\|\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}} - \mathbf{\Sigma}_u^{\mathcal{T}}\|$ separately.

First of all, for symmetric matrix $\mathbf{A} = (a_{ij})$, $\|\mathbf{A}\| \leq \max_i \sum_{j=1}^p |a_{ij}|$. Therefore we have

$$\begin{aligned}\|\mathbf{\Sigma}_u^{\mathcal{T}} - \mathbf{\Sigma}_u\| &\leq \max_{i\leq p} \sum_{j=1}^p |\sigma_{ij}| I(|\sigma_{ij}| \leq \omega_T \hat{\theta}_{ij}^{1/2}) \\ &\leq \max_i \sum_{j:\sigma_{ij}\neq 0} \omega_T \hat{\theta}_{ij}^{1/2} = O_p(\omega_T m_T),\end{aligned}$$

where the last equality is due to $\hat{\theta}_{ij}$ is bounded above uniformly in i, j with probability approaching one, according to Lemma A.4.

On the other hand,

$$\begin{aligned}\|\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}} - \mathbf{\Sigma}_u^{\mathcal{T}}\| \leq\ & \max_{i\leq p} \sum_{j=1}^p |\sigma_{ij}| I(|\widehat{\sigma}_{ij}| \leq \omega_T \hat{\theta}_{ij}^{1/2}, |\sigma_{ij}| > \omega_T \hat{\theta}_{ij}^{1/2}) \\ &+ \max_{i\leq p} \sum_{j=1}^p |\sigma_{ij} - \widehat{\sigma}_{ij}| I(|\widehat{\sigma}_{ij}| > \omega_T \hat{\theta}_{ij}^{1/2}, |\sigma_{ij}| > \omega_T \hat{\theta}_{ij}^{1/2}) \\ &+ \max_{i\leq p} \sum_{j=1}^p |\widehat{\sigma}_{ij}| I(|\sigma_{ij}| \leq \omega_T \hat{\theta}_{ij}^{1/2}, |\widehat{\sigma}_{ij}| > \omega_T \hat{\theta}_{ij}^{1/2}).\end{aligned}$$

Since by Lemma A.4, $\hat{\theta}_{ij}^{1/2}$ is bounded away from both zero and infinity uniformly in i, j, all the three terms on the right hand side can be bounded in a similar way as in the proof of Theorem 1 in Bickel and Levina (2008a), corresponding to the case $q = 0$. Therefore the details are omitted, which are available from the authors. Here we only show a key different step in the proof, which is,

$$\max_{i\leq p} \sum_{j=1}^p I(|\widehat{\sigma}_{ij} - \sigma_{ij}| \geq (1-r)\omega_T \hat{\theta}_{ij}) = O_p(1). \tag{A.2}$$

for any $r \in (0,1)$. This implies that

$$\begin{aligned}
&\max_{i\le p}\sum_{j=1}^{p}|\widehat{\sigma}_{ij}-\sigma_{ij}|I(|\widehat{\sigma}_{ij}|\ge \omega_T\hat{\theta}_{ij},|\sigma_{ij}|\le r\omega_T\hat{\theta}_{ij})\\
&\le O_p(\omega_T)\max_{i\le p}\sum_{j=1}^{p}I(|\widehat{\sigma}_{ij}|\ge \omega_T\hat{\theta}_{ij},|\sigma_{ij}|\le r\omega_T\hat{\theta}_{ij})\\
&= O_p(\omega_T). \qquad \text{(A.3)}
\end{aligned}$$

To show (A.2), let $C_1 > 0$ be such that $P(\min_{ij}\hat{\theta}_{ij} \le C_1) = o(1)$, whose existence is guaranteed by Lemma A.4. Since $\max_{ij}|\widehat{\sigma}_{ij}-\sigma_{ij}| = O_p(\omega_T)$, for any $\epsilon, M > 0$, and sufficiently large $C > 0$,

$$\begin{aligned}
&P(\max_{i\le p}\sum_{j=1}^{p}I(|\widehat{\sigma}_{ij}-\sigma_{ij}|\ge(1-r)\omega_T\hat{\theta}_{ij})>M)\\
&\le P(\max_{ij}|\widehat{\sigma}_{ij}-\sigma_{ij}|\ge(1-r)\omega_T\hat{\theta}_{ij})\\
&\le P\left(\frac{\max_{ij}|\widehat{\sigma}_{ij}-\sigma_{ij}|}{\max\{\sqrt{(\log p)/T},a_T\}}\ge(1-r)CC_1\right)+o(1)<\epsilon,
\end{aligned}$$

which yields the result.

(ii) Since both $\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}}$ and $\mathbf{\Sigma}_u$ are symmetric and $\lambda_{\min}(\mathbf{\Sigma}_u) > C$ for some $C > 0$, the result follows immediately from Lemma A.1.

Q.E.D.

APPENDIX B: PROOFS FOR SECTION 3

B.1. Proof of Theorem 3.1.

LEMMA B.1. *(i)* $\max_{i\le p}\|\widehat{\mathbf{b}}_i-\mathbf{b}_i\| = O_p\left(\sqrt{\frac{K\log p}{T}}\right)$.
(ii) $\max_{i,j\le K}|\frac{1}{T}\sum_{t=1}^{T}f_{it}f_{jt}-Ef_{it}f_{jt}| = O_p(\sqrt{\frac{\log K}{T}})$.

PROOF. As $\widehat{\mathbf{b}}_i-\mathbf{b}_i = (\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}\mathbf{u}_i$, we have

$$\|\widehat{\mathbf{b}}_i-\mathbf{b}\|^2 = \mathbf{u}_i'\mathbf{X}'(\mathbf{X}\mathbf{X}')^{-2}\mathbf{X}\mathbf{u}_i.$$

Since $\{\mathbf{f}_t\}_{t\le T}$ is ergodic and stationary, $\|\frac{1}{T}\mathbf{X}\mathbf{X}'-E\mathbf{f}_t\mathbf{f}_t'\| = o_p(1)$. Then $\lambda_{\min}(\mathrm{cov}(\mathbf{f}_t)) > c > 0$ implies that the smallest eigenvalue of $\frac{1}{T}\mathbf{X}\mathbf{X}'$ is bounded away from zero with probability approaching one. Then

$$\|\widehat{\mathbf{b}}_i-\mathbf{b}\|^2 \le C\|\frac{1}{T}\mathbf{X}\mathbf{u}_i\|^2 = C\sum_{k=1}^{K}(\frac{1}{T}\sum_{t=1}^{T}f_{kt}u_{it})^2$$

$$\leq \quad O_p(K) \max_{k\leq K, i\leq p} (\frac{1}{T}\sum_{t=1}^{T} f_{kt}u_{it})^2.$$

Let $Z_{ki} = \frac{1}{T}\sum_{t=1}^{T} f_{kt}u_{it}$. We bound $|Z_{ki}|$ using the Bernstein type inequality. By Lemma A.2, and Assumptions 2.1(iii) and 3.3(ii), $f_{kt}u_{it}$ satisfies the exponential tail condition (2.6) for some $r_3 = \frac{rr_2}{2r+2r_2} \in (0,1)$, as well as the strong mixing condition. Hence by the Bernstein inequality for weakly dependent data in Merlevède (2009, Theorem 1), there exist $C_i > 0$, $i = 1, ..., 5$, for any $s > 0$

$$P(|Z_{ki}| > s) \leq T\exp\left(-\frac{(Ts)^{r_3}}{C_1}\right) + \exp\left(-\frac{T^2s^2}{C_2(1+TC_3)}\right)$$
$$+\exp\left(-\frac{(Ts)^2}{C_4T}\exp\left(\frac{(Ts)^{r_3(1-r_3)}}{C_5(\log Ts)^{r_3}}\right)\right).$$

The Bonferroni's method then implies that, if $(\log p)^{2/\gamma+2/r+2/r_2-1} = o(T)$,

$$\max_{i\leq p, k\leq K} |Z_{ki}| = O_p(\sqrt{\frac{\log p}{T}}).$$

It then yields the desired result.

(ii) Lemma A.2 implies that for any i and $j \leq K$, $f_{it}f_{jt}$ satisfies the exponential tail condition (2.6) with $r_4 = r_2/4 \in (0,1)$. Therefore, the result follows from Theorem 1 of Merlevède (2009) and the Bonferroni's method.

Q.E.D.

LEMMA B.2. *When $(\mathbf{f}_t)_{t\geq 1}$ are observable,*
(i) $\max_{i\leq p} \frac{1}{T}\sum_{t=1}^{T} |u_{it} - \hat{u}_{it}|^2 = O_p(\frac{K^2 \log p}{T})$,
(ii) $\max_{i,t} |u_{it} - \hat{u}_{it}| = o_p(1)$.

PROOF. (i) $\max_{i\leq p} \frac{1}{T}\sum_{t=1}^{T} |u_{it} - \hat{u}_{it}|^2 \leq \max_{i\leq p} \frac{1}{T}\sum_t \|\mathbf{f}_t\|^2 \|\widehat{\mathbf{b}}_i - \mathbf{b}_i\|^2$. Note that

$$\frac{1}{T}\sum_t \|\mathbf{f}_t\|^2 \leq K \max_{k\leq K} |\frac{1}{T}\sum_{t=1}^{T} f_{kt}^2 - Ef_{kt}^2| + K\max_{k\leq K} Ef_{kt}^2 = O_p(K).$$

The result then follows immediately from Lemma B.1.

(ii) By Assumption 3.3, $E\|K^{-1/2}\mathbf{f}_t\|^4 < M$. Hence Lemma D.2 in Kitamura et al (2004) yields $\max_{t\leq T} \|\mathbf{f}_t\| = o_p(T^{1/4}\sqrt{K})$. We then have

$$\max_{t\leq T, i\leq p} |u_{it} - \hat{u}_{it}| = \max_{t\leq T, i\leq p} |(\widehat{\mathbf{b}}_i - \mathbf{b}_i)'\mathbf{f}_t| = O_p(K\sqrt{\log p}T^{-1/4}) = o_p(1).$$

Proof of Theorem 3.1 Theorem 3.1 follows immediately from Theorem 2.1 and Lemma B.2. Q.E.D.

B.2. Proof of Theorem 3.2 Part (i). We follow similar lines of proof as in Fan, Fan and Lv (2008). Define

$$\mathbf{D}_T = \widehat{\text{cov}}(\mathbf{f}_t) - \text{cov}(\mathbf{f}_t), \qquad \mathbf{H} = \mathbf{X}'(\mathbf{X}\mathbf{X}')^{-1}\mathbf{X},$$

$$\mathbf{C}_T = \widehat{\mathbf{B}} - \mathbf{B}, \qquad \mathbf{E} = (\mathbf{u}_1, ..., \mathbf{u}_T).$$

With probability approaching one,

$$\begin{aligned} \|\widehat{\mathbf{\Sigma}}^{\mathcal{T}} - \mathbf{\Sigma}\|_\Sigma^2 \quad \leq \quad & C[\|\mathbf{B}\mathbf{D}_T\mathbf{B}'\|_\Sigma^2 + \|\mathbf{B}\widehat{\text{cov}}(\mathbf{f})\mathbf{C}_T{}'\|_\Sigma^2 \\ & + \|\mathbf{C}_T\widehat{\text{cov}}(\mathbf{f})\mathbf{C}_T{}'\|_\Sigma^2 + \|\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}} - \mathbf{\Sigma}_u\|_\Sigma^2]. \end{aligned} \tag{B.1}$$

We bound the terms on the right hand side in the following lemmas.

LEMMA B.3. *(i)* $\|\mathbf{D}_T\|_F^2 = O_p(\frac{K^2 \log K}{T})$.
(ii) $\|\mathbf{C}_T\|_F^2 = O_p(pK/T)$.

PROOF. (i) It follows immediately from Lemma B.1(ii) since

$$\mathbf{D}_T^2 \leq K^2 \left(\max_{i,j \leq K} |\frac{1}{T}\sum_{t=1}^T f_{it}f_{jt} - Ef_{it}f_{jt}|^2 + \max_{i,j \leq K} |\frac{1}{T}\sum_{t=1}^T f_{it}\frac{1}{T}\sum_{t=1}^T f_{jt} - Ef_{it}Ef_{jt}|^2 \right).$$

(ii) Note that $\mathbf{C}_T = \mathbf{E}\mathbf{X}'(\mathbf{X}\mathbf{X}')^{-1}$. By the facts that $E(u_{it}^2)$ is bounded uniformly in i, t, and $(\mathbf{u}_t)_{t=1}^T$ are i.i.d., we have

$$E[\|\mathbf{C}_T\|_F^2|\mathbf{X}] = \text{tr}((\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}E(\mathbf{E}'\mathbf{E}|\mathbf{X})\mathbf{X}'(\mathbf{X}\mathbf{X}')^{-1}) = O_p(Kp/T).$$

LEMMA B.4. $\|\mathbf{B}\mathbf{D}_T\mathbf{B}'\|_\Sigma^2 + \|\mathbf{B}\widehat{\text{cov}}(\mathbf{f}_t)\mathbf{C}_T'\|_\Sigma^2 = O_p(K/T + K^2 \log K/(Tp))$.

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

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

Hence $\|\mathbf{B}\mathbf{D}_T\mathbf{B}'\|_\Sigma^2 \leq p^{-1}\|\mathbf{D}_T\mathbf{B}'\mathbf{\Sigma}^{-1}\mathbf{B}\|_F^2 = O_p(p^{-1})\|\mathbf{D}_T\|_F^2 = K^2 \log K/(Tp)$.

On the other hand,

$$\|\mathbf{B}\widehat{\text{cov}}(\mathbf{f})\mathbf{C}_T{}'\|_\Sigma^2 \leq 8T^{-2}\|\mathbf{B}\mathbf{X}\mathbf{X}'\mathbf{C}_T{}'\|_\Sigma^2 + 8T^{-4}\|\mathbf{B}\mathbf{X}\mathbf{1}\mathbf{1}'\mathbf{X}'\mathbf{C}_T{}'\|_\Sigma^2.$$

Respectively, $\|\mathbf{B}\mathbf{X}\mathbf{X}'\mathbf{C}_T{}'\|_\Sigma^2 \leq p^{-1}\|\mathbf{X}\mathbf{X}'\mathbf{C}_T'\mathbf{\Sigma}^{-1}\|_F\|\mathbf{C}_T\mathbf{X}\mathbf{X}'\mathbf{B}'\mathbf{\Sigma}^{-1}\mathbf{B}\|_F = O_p(TK)$. Likewise, $\|\mathbf{B}\mathbf{X}\mathbf{1}\mathbf{1}'\mathbf{X}'\mathbf{C}_T'\|_\Sigma^2 = O_p(KT^3)$. This yields the result.

LEMMA B.5. $\|\mathbf{C}_T\widehat{\text{cov}}(\mathbf{f})\mathbf{C}_T{}'\|_\Sigma^2 = O_p(\frac{pK^2}{T^2})$.

PROOF. Straightforward calculation yields:

$$\begin{aligned} p\|\mathbf{C}_T\widehat{\text{cov}}(\mathbf{f})\mathbf{C}_T{}'\|_\Sigma^2 &= \text{tr}(\mathbf{C}_T\widehat{\text{cov}}(\mathbf{f})\mathbf{C}_T{}'\mathbf{\Sigma}^{-1}\mathbf{C}_T\widehat{\text{cov}}(\mathbf{f})\mathbf{C}_T{}'\mathbf{\Sigma}^{-1}) \\ &\leq \|\mathbf{E}\mathbf{X}'(\mathbf{X}\mathbf{X}')^{-1}\widehat{\text{cov}}(\mathbf{f})(\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}\mathbf{E}'\mathbf{\Sigma}^{-1}\|_F^2 \\ &= O_p(1)\|\mathbf{E}\mathbf{X}'(\mathbf{X}\mathbf{X}')^{-1}\|_F^4. \end{aligned}$$

We have $E(\mathbf{E}'\mathbf{E}|\mathbf{X}) = \text{tr}(\mathbf{\Sigma}_u)I_p$. Hence $E\|\mathbf{E}\mathbf{X}'(\mathbf{X}\mathbf{X}')^{-1}\|_F^2 = O(p)E\text{tr}((\mathbf{X}\mathbf{X}')^{-1}) = O(pK/T)$, which implies $\|\mathbf{E}\mathbf{X}'(\mathbf{X}\mathbf{X}')^{-1}\|_F^2 = O_p(pK/T)$, and yields the desired result.

Proof of Theorem 3.2 Part (i)

(a) By Theorem 3.1, we have

$$\begin{aligned} \|\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}} - \mathbf{\Sigma}_u\|_\Sigma &= p^{-1/2}\|\mathbf{\Sigma}^{-1/2}(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}} - \mathbf{\Sigma}_u)\mathbf{\Sigma}^{-1/2}\|_F \\ &\leq \|\mathbf{\Sigma}^{-1/2}(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}} - \mathbf{\Sigma}_u)\mathbf{\Sigma}^{-1/2}\| \\ &\leq \|\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}} - \mathbf{\Sigma}_u\| \cdot \lambda_{\max}(\mathbf{\Sigma}^{-1}) \\ &= O_p(Km_T\sqrt{\frac{\log p}{T}}). \end{aligned} \tag{B.2}$$

Therefore, (B.1) and Lemma B.4-B.5 yield

$$\|\widehat{\mathbf{\Sigma}}^{\mathcal{T}} - \mathbf{\Sigma}\|_\Sigma^2 = O_p(\frac{pK^2}{T^2} + \frac{K^2m_T^2\log p}{T}).$$

(b) For the infinity norm, it is straightforward to show that

$$\begin{aligned} \|\widehat{\mathbf{\Sigma}}^{\mathcal{T}} - \mathbf{\Sigma}|_\infty \leq\ & K^2\|\mathbf{B}\|_\infty\|\widehat{\text{cov}}(\mathbf{f}_t) - \text{cov}(\mathbf{f}_t)\|_\infty(\|\widehat{\mathbf{B}} - \mathbf{B}\|_\infty + \|\mathbf{B}\|_\infty) \\ &+ K^2\|\widehat{\mathbf{B}} - \mathbf{B}\|_\infty(\|\widehat{\text{cov}}(\mathbf{f}_t) - \text{cov}(\mathbf{f}_t)\|_\infty + \|\text{cov}(\mathbf{f}_t)\|_\infty)(\|\widehat{\mathbf{B}} - \mathbf{B}\|_\infty + \|\mathbf{B}\|_\infty) \\ &+ K^2\|\mathbf{B}\|_\infty\|\text{cov}(\mathbf{f}_t)\|_\infty\|\widehat{\mathbf{B}} - \mathbf{B}\|_\infty \\ &+ \|\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}} - \mathbf{\Sigma}_u|_\infty. \end{aligned} \tag{B.3}$$

In addition, we have

$$\|\frac{1}{T}\mathbf{E}\mathbf{X}'\|_\infty = \max_{i\leq K, j\leq p}|\frac{1}{T}\sum_{t=1}^T f_{it}u_{jt}| = O_p(\sqrt{\frac{\log p}{T}}),$$

and

$$\|\widehat{\text{cov}}(\mathbf{f}_t) - \text{cov}(\mathbf{f}_t)\|_\infty = O_p(\sqrt{\frac{\log K}{T}}).$$

In addition, $\|(\frac{1}{T}\mathbf{X}\mathbf{X}')^{-1}\|_\infty \leq \lambda_{\max}((\frac{1}{T}\mathbf{X}\mathbf{X}')^{-1}) \leq \lambda_{\min}^{-1}(\widehat{\text{cov}}(\mathbf{f}_t)) = O_p(1)$. Hence

$$\|\widehat{\mathbf{B}} - \mathbf{B}\|_\infty \leq K\|\frac{1}{T}\mathbf{E}\mathbf{X}'\|_\infty\|(\frac{1}{T}\mathbf{X}\mathbf{X}')^{-1}\|_\infty = O_p(K\sqrt{\frac{\log p}{T}}). \tag{B.4}$$

Inserting $\|\mathbf{B}\|_\infty = O(1)$, $\|\widehat{\text{cov}}(\mathbf{f}_t)\|_\infty$, $\|\mathbf{B} - \widehat{\mathbf{B}}\|_\infty$, and $\|\widehat{\text{cov}}(\mathbf{f}_t) - \text{cov}(\mathbf{f}_t)\|_\infty$ into (B.3) yields

$$\|\widehat{\mathbf{\Sigma}}^{\mathcal{T}} - \mathbf{\Sigma}\|_\infty \leq O_p(K^3\sqrt{\frac{\log p}{T}}) + \|\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}} - \mathbf{\Sigma}_u\|_\infty.$$

Moreover, the (i,j)th entry of $\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}} - \mathbf{\Sigma}_u$ is given by

$$\widehat{\sigma}_{ij}I(|\widehat{\sigma}_{ij}| \geq \omega_T\sqrt{\hat{\theta}_{ij}}) - \sigma_{ij} = \begin{cases} -\sigma_{ij}, & \text{if } |\widehat{\sigma}_{ij}| < \omega_T\sqrt{\hat{\theta}_{ij}} \\ \widehat{\sigma}_{ij} - \sigma_{ij}, & \text{o.w.} \end{cases}$$

When $|\widehat{\sigma}_{ij}| < \omega_T\sqrt{\hat{\theta}_{ij}}$, $|\sigma_{ij}| \leq |\sigma_{ij} - \widehat{\sigma}_{ij}| + |\widehat{\sigma}_{ij}| = O_p(\omega_T)$, by Lemmas A.3 and A.4. Hence $\|\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}} - \mathbf{\Sigma}_u\|_\infty = O_p(\omega_T)$, which yields the result. Q.E.D.

B.3. Proof of Theorem 3.2 Part (ii).

LEMMA B.6. *(i)* $\|[\widehat{\text{cov}}(\mathbf{f})^{-1} + \widehat{\mathbf{B}}'(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1}\widehat{\mathbf{B}}]^{-1}\| = O_p(p^{-1})$,
$\|[\text{cov}(\mathbf{f})^{-1} + \mathbf{B}'\mathbf{\Sigma}_u^{-1}\mathbf{B}]^{-1}\| = O(p^{-1})$.
(ii) $\lambda_{\min}(\mathbf{B}'\mathbf{\Sigma}_u^{-1}\mathbf{B}) \geq cp$ *for some* $c > 0$.

PROOF. By Assumption, $\lambda_{\min}(\text{cov}(\mathbf{f})^{-1} + \mathbf{B}'\mathbf{\Sigma}_u^{-1}\mathbf{B}) \geq \lambda_{\min}(\mathbf{B}'\mathbf{\Sigma}_u^{-1}\mathbf{B}) > cp$ for a constant $c > 0$. In addition,

$$\|\widehat{\mathbf{B}} - \mathbf{B}\| = O_p(\sqrt{pK/T}),$$

$$\|\text{cov}(\mathbf{f})^{-1} - \widehat{\text{cov}}(\mathbf{f})^{-1}\|_F = O_p(K/\sqrt{T}),$$

and that

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

The same argument of Fan, Fan and Lv (2008) (eq. 14) implies that $\|\mathbf{B}\|_F = O(\sqrt{p})$. Therefore,

$$\|\text{cov}(\mathbf{f})^{-1} + \mathbf{B}'\mathbf{\Sigma}_u^{-1}\mathbf{B} - (\widehat{\text{cov}}(\mathbf{f})^{-1} + \widehat{\mathbf{B}}'(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1}\widehat{\mathbf{B}})\| = o_p(p).$$

The results in (i) then follow from Lemma A.1.

(ii) Let $\mathbf{v}$ be the eigenvector of $\mathbf{B}'\mathbf{\Sigma}_u^{-1}\mathbf{B}$ corresponding to the smallest eigenvalue, and $\|\mathbf{v}\| = 1$. Then

$$\lambda_{\min}(\mathbf{B}'\mathbf{\Sigma}_u^{-1}\mathbf{B}) = \mathbf{v}'\mathbf{B}'\mathbf{\Sigma}_u^{-1}\mathbf{B}\mathbf{v} \geq \lambda_{\min}(\mathbf{\Sigma}_u^{-1})\mathbf{v}'\mathbf{B}'\mathbf{B}\mathbf{v} \geq c\lambda_{\min}(\mathbf{B}'\mathbf{B})$$

given that $\|\mathbf{\Sigma}_u\|$ is bounded from above. By Assumption 3.5 and Lemma A.1, the smallest eigenvalue of $\mathbf{B}'\mathbf{B} \geq c_1 p$ for some $c_1 > 0$, which completes the proof.

Q.E.D.

Using the Sherman-Morrison-Woodbury formula, we have

(B.5)

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

Theorem 3.1 implies $L_1 = O_p(Km_T\sqrt{\frac{\log p}{T}})$.

Let $G = [\widehat{\mathrm{cov}}(\mathbf{f})^{-1} + \widehat{\mathbf{B}}'(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1}\widehat{\mathbf{B}}]^{-1}$, then

(B.6)

$$L_2 \leq \|((\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1} - \mathbf{\Sigma}_u^{-1})(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{1/2}\| \cdot \|(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1/2}\widehat{\mathbf{B}}G\widehat{\mathbf{B}}'(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1/2}\| \cdot \|(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1/2}\|.$$

Note that $\|\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}}\| \leq \|\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}} - \mathbf{\Sigma}_u\| + \lambda_{\max}(\mathbf{\Sigma}_u) < C$ for a constant $C > 0$. By Lemma A.1 and Theorem 3.1, $\lambda_{\max}(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1/2} \leq C$.

In addition, the middle term in (B.6) can be treated in the same way as in the proof of (28) in Fan, Fan and Lv (2008):

$$(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1/2}\widehat{\mathbf{B}}G\widehat{\mathbf{B}}'(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1/2} = I - (\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{1/2}(\widehat{\mathbf{\Sigma}}^{\mathcal{T}})^{-1}(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{1/2} \leq I.$$

Hence the middle term is bounded by one. This shows that $L_2 = O_p(L_1)$. Similarly,

$$\begin{aligned}
L_3 &\leq \|((\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1} - \mathbf{\Sigma}_u^{-1})(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{1/2}\| \cdot \|(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1/2}\widehat{\mathbf{B}}G\widehat{\mathbf{B}}'(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1/2}\| \cdot \|\mathbf{\Sigma}_u^{-1}(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{1/2}\| \\
&= O_p(L_1).
\end{aligned}$$

Lemma B.6 shows that $\|G\| = O_p(p^{-1})$. Hence

$$L_4 \leq \|\mathbf{\Sigma}_u^{-1}(\widehat{\mathbf{B}} - \mathbf{B})\|\|G\|\|\widehat{\mathbf{B}}'\mathbf{\Sigma}_u^{-1}\| = O_p(\sqrt{\frac{pK}{T}}\frac{1}{p}\sqrt{p}) = O_p(\sqrt{\frac{K}{T}}).$$

Similarly $L_5 = O_p(\sqrt{K/T})$. Finally, let $G_1 = [\text{cov}(\mathbf{f})^{-1} + \mathbf{B}'\mathbf{\Sigma}_u^{-1}\mathbf{B}]^{-1}$, then

$$\begin{aligned}\|G - G_1\| &\leq \|G(G^{-1} - G_1^{-1})G_1\| \\ &\leq \|G\|\|G_1\|\|\text{cov}(\mathbf{f})^{-1} + \mathbf{B}'\mathbf{\Sigma}_u^{-1}\mathbf{B} - (\widehat{\text{cov}}(\mathbf{f})^{-1} + \widehat{\mathbf{B}}'(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1}\widehat{\mathbf{B}})\| \\ &\leq O_p(p^{-2})\|\text{cov}(\mathbf{f})^{-1} - \widehat{\text{cov}}(\mathbf{f})^{-1}\| + O_p(p^{-2})\|\mathbf{B}'\mathbf{\Sigma}_u^{-1}\mathbf{B} - \widehat{\mathbf{B}}'(\widehat{\mathbf{\Sigma}}_u^{\mathcal{T}})^{-1}\widehat{\mathbf{B}}\| \\ &= O_p(p^{-1}Km_T\sqrt{\frac{\log p}{T}}).\end{aligned}$$

Therefore $L_6 \leq \|\mathbf{\Sigma}_u^{-1}\mathbf{B}\|^2\|G - G_1\| = O_p(Km_T\sqrt{\frac{\log p}{T}})$. The proof is completed by combining $L_1 \sim L_6$. Q.E.D.

APPENDIX C: PROOFS FOR SECTION 4

The OLS is given by

$$\widehat{\mathbf{b}}_i = (\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{y}_i, i \leq p.$$

The same arguments in the proof of Lemma B.1 can yield

$$\max_{i \leq p}\|\widehat{\mathbf{b}}_i - \mathbf{b}_i\| = O_p\left(\sqrt{\frac{K\log p}{T}}\right),$$

which then implies the rate of

$$\max_{i \leq p}\frac{1}{T}\sum_{t=1}^{T}(u_{it} - \hat{u}_{it})^2 \leq \|\widehat{\mathbf{b}}_i - \mathbf{b}_i\|^2\frac{1}{T}\sum_{t=1}^{T}\|\mathbf{f}_{it}\|^2.$$

The result then follows from a straightforward application of Theorem 2.1.

REFERENCES

[1] ANTONIADIS, A. and FAN, J. (2001). Regularization of wavelet approximations. *J. Amer. Statist. Assoc.* **96** 939-967.

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

[3] BAI, J. and NG, S.(2002). Determining the number of factors in approximate factor models. *Econometrica*. **70** 191-221.

[4] BERNANKE, B. and BOIVIN, J. (2003). Monetary policy in a data rich environment. *Journal of Monetary Economics*. **50** 525-546.

[5] BICKEL, P. and LEVINA, E. (2008a). Covariance regularization by thresholding. *Ann. Statist.*. **36** 25772604.

[6] BICKEL, P. and LEVINA, E. (2008b). Regularized estimation of large covariance matrices. *Ann. Statist*. **36** 199-227.

[7] CAI, T. and LIU, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. To appear in *J. Amer. Statist. Assoc.*.

[8] CAI, T. and ZHOU, H. (2010). Optimal rates of convergence for sparse covariance matrix estimation. *Manuscript*. University of Pennsylvania.

[9] CHAMBERLAIN, G. and ROTHSCHILD, M. (1983). Arbitrage, factor structure and mean-variance analyssi in large asset markets. *Econometrica*. **51** 1305-1324.

[10] CONNOR, G. and KORAJCZYK, R. (1993). A Test for the number of factors in an approximate factor model. *Journal of Finance*. **48**, 1263-1291

[11] 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.

[12] FAMA, E. and FRENCH, K. (1992). The cross-section of expected stock returns. *Journal of Finance*. **47** 427-465.

[13] LAM, C. and FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.*. **37** 4254-4278.

[14] FAN, J., FAN, Y. and LV, J. (2008). High dimensional covariance matrix estimation using a factor model. *J. Econometrics*. **147** 186-197

[15] FAN, J., ZHANG, J. and YU, K. (2008). Asset allocation and risk assessment with gross exposure constraints for vast portfolios. *Manuscript*. Princeton University.

[16] FAVERO, C. and MARCELLINO, M. (2001). Large datasets, small models and monetary policy in Europe. IEP, Bocconi University, and CEPR.

[17] FORINI, 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.

[18] GORMAN, M. (1981). Some Engel curves, in *Essays in the Theory and Measurement of Consumer Behavior in Honor of Sir Richard Stone*, ed. by A. Deaton. New York: Cambridge University Press.

[19] HARDING, M. (2009). Structural estimation of high-dimensional factor models. *Manuscript* Stanford University

[20] 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.

[21] KMENTA, J. and GILBERT, R. (1970). Estimation of seemingly unrelated regressions with autoregressive disturbances, *J. Amer. Statist. Assoc.* **65** 186-196.

[22] KITAMURA, Y., TRIPATHI, G. and Ahn, H. (2004). Empirical likelihood-based inference in conditional moment restriction models. *Econometrica*. **72** 1667-1714.

[23] LEWBEL, A. (1991). The rank of demand systems: theory and nonparametric estimation. *Econometrica*. **59** 711-730.

[24] MERLEVÈDE, F., PELIGRAD, M. and RIO, E. (2009). A Bernstein type inequality and moderate deviations for weakly dependent sequences. *Manuscript*. Université Paris Est.

[25] ROTHMAN, A., LEVINA, E. and ZHU, J. (2009) Generalized thresholding of large covariance matrices. *J. Amer. Statist. Assoc.* **104** 177-186.

[26] 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.

[27] ZELLNER, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *J. Amer. Statist. Assoc.* **57** 348-368.

DEPARTMENT OF OPERATIONS RESEARCH AND FINANCIAL ENGINEERING
PRINCETON UNIVERSITY
PRINCETON, NJ 08544
E-MAIL: jqfan@princeton.edu; yuanliao@princeton.edu; mincheva@princeton.edu