HIGH-DIMENSIONAL STRUCTURED
COVARIANCE MATRIX ESTIMATION WITH
FINANCIAL APPLICATIONS

Martina Zhelcheva Mincheva

A DISSERTATION
PRESENTED TO THE FACULTY
OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE
BY THE DEPARTMENT OF
OPERATIONS RESEARCH AND FINANCIAL ENGINEERING
ADVISER: JIANQING FAN

JUNE 2014
Abstract

This thesis deals with high dimensional statistical inference, and more specifically with uncovering low dimensional structures in high dimensional systems. It focuses on large covariance and precision matrix estimation under various complexity constraints, such as low rank and conditional sparsity properties. I have also demonstrated their applications in portfolio allocation and risk minimization.

In recent years, high-dimensional settings have become one of the major focuses of statistics research. The interest has been driven by applications in finance, genetics, image processing, etc. and technological advances have made large data sets widely available for examination. In a data rich environment, the number of parameters can diverge at a rate faster than that of the sample size. A natural way of dealing with this challenge is establishing low dimensional patterns in the data, such as low-rank matrices and sparsity. The low-rank approach assumes that there is an underlying lower-dimensional subspace that drives the random processes. Sparsity implies that a large number of the coefficients to be estimated are equal to zero, and the number of non-zero entries is restricted to grow slowly.

My thesis is on high dimensional covariance matrix estimation with a conditional sparsity structure and fast diverging eigenvalues. I consider the high dimensional approximate factor model, in which the number of units grows possibly exponentially fast with the sample size. Classical methods of estimating the covariance matrices in factor models are based on the cross-sectional independence assumption among the idiosyncratic components. This assumption, however, is restrictive in practical applications. For example, returns depend on equity market risks, housing prices depend on economic health. By assuming the error covariance matrix to be sparse, I allow the presence of the cross-sectional correlation, and combine the merits of both
the sparsity modeling and the factor structure. In financial applications, the residual covariance represents idiosyncratic risk that can be diversified away, and so makes a smaller order contribution to portfolio risk, but in practice it can be important.

I study the impact of weakly dependent data with strong mixing conditions on estimation, and obtain asymptotically nonsingular estimators for the covariance matrices using various thresholding techniques. It is shown that the estimated covariance matrices are consistent under various norms. Both observable and unobservable factor cases are considered. The uniform rates of convergence for the factor loadings and the unobservable factors are also derived. This approach is simple and optimization free and it uses the data only through the sample covariance matrix.

The results of the thesis are separated into two chapters. Chapter 1 is based on the assumption that direct observations of the common factors are available. The true number of factors is also given. We take the classical definition of sparsity for the error covariance matrix, namely that many of its off-diagonal elements are zero. We estimate the sparse covariance using the adaptive thresholding technique as in Cai and Liu (2011), taking into account the fact that the true idiosyncratic components are unknown.

Chapter 2 explores a much more general setup. We introduce the Principal Orthogonal complement Thresholding (POET) method for covariance estimation through principal component analysis of the data. We provide mathematical insights for the connection between this method and the approximate factor model with sparsity for high-dimensional data. The common factors are unobservable, and their number is also unknown. It is shown that the impact of estimating the unknown factors vanishes as the dimensionality increases. We use a generalized notion of sparsity and a broader thresholding technique, namely the generalized shrinkage
function of Antoniadis and Fan (2001). Thus, 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. The advantages of this relative to other methods are presented through extensive simulation and real data studies. A case study for risk minimization demonstrates that portfolios created with the POET procedure significantly outperform their counterparts.
Acknowledgements

First and foremost, I would like to thank my advisor, Jianqing Fan, for his help and patience over the last five years. He has been my mentor both academically and personally, and it’s been an honor to be a part of his group. He is a true scholar with brilliant ideas, and needless to say, none of the results in this thesis would have existed without him. He has also taught me, on multiple occasions, that confidence is just as important as knowledge for academic success, and I am grateful for his insights.

Also, I have to thank Yuan Liao, who has been like a second advisor to me, and the best coauthor I could have hoped for. I have learnt so much from him, and I sincerely hope to have the chance to keep collaborating with him beyond the scope of this thesis.

And last, but not least, I am grateful to Marc Hallin, Han Liu, Oliver Linton, Lie Wang, Piotr Fryzlewicz, and the many more discussants of one of our papers. Thank you for providing insightful comments and bringing up issues that will contribute towards my future research agenda.
To Dilyana.
# Contents

Abstract ................................................................. iii  
Acknowledgements ..................................................... vi  
List of Figures ........................................................... xi  

1 Covariance Estimation in Approximate Factor Models with Observable Factors 1  
1.1 Introduction .......................................................... 1  
1.2 Estimation of Error Covariance Matrix .......................... 5  
1.2.1 Adaptive thresholding .......................................... 5  
1.2.2 Asymptotic properties of the thresholding estimator ....... 8  
1.3 Estimation of Covariance Matrix Using Factors ............... 11  
1.4 Extension: Seemingly Unrelated Regression .................... 16  

2 Covariance Estimation by Thresholding Principal Orthogonal Complements (a.k.a. in Approximate Factor Models with Unobservable Factors) 19  
2.1 Introduction .......................................................... 19  
2.2 Regularized Covariance Matrix via PCA ....................... 26  
2.2.1 High-dimensional PCA and factor model ................... 26  
2.2.2 POET ............................................................... 30  
2.2.3 Least squares point of view .................................. 32
5 Technical Proofs

5.1 Proofs for Section 1.2 ........................................ 89
   5.1.1 Lemmas ........................................ 89
   5.1.2 Proof of Theorem 1.2.1 ............................ 94

5.2 Proofs for Section 1.3 ....................................... 97
   5.2.1 Proof of Theorem 1.3.1 ............................. 97
   5.2.2 Proof of Theorem 1.3.2 Part (i) .................... 99
   5.2.3 Proof of Theorem 1.3.2 Part (ii) ................... 102

5.3 Proofs for Section 4 ........................................ 105

5.4 Estimating a sparse covariance with contaminated data ............ 105

5.5 Proofs for Section 2.2 ....................................... 108

5.6 Proofs for Section 2.3 ....................................... 110
   5.6.1 Preliminary lemmas ................................ 110
   5.6.2 Proof of Theorem 2.3.3 ............................. 112
   5.6.3 Proof of Theorem 2.3.1 ............................. 118
   5.6.4 Proof of Theorem 2.3.2 ............................. 118

Bibliography ................................................... 123
List of Figures

2.1 Minimum eigenvalue of $\hat{\Sigma}_T^{u,K}(C)$ as a function of $C$ for three choices of thresholding rules. The plot is based on the simulated data set in Section 6.2. ................................................................. 48

3.1 Averages and standard deviations of $\|\hat{\Sigma}^T - \Sigma\|_\Sigma$ (dashed curve) and $\|\hat{\Sigma}_y - \Sigma\|_\Sigma$ (solid curve) over $N = 200$ iterations, as a function of the dimensionality $p$. ................................................................. 59

3.2 Averages and standard deviations of $\|\hat{\Sigma}^T - \Sigma\|_\infty$ (dashed curve) and $\|\hat{\Sigma}_y - \Sigma\|_\infty$ (solid curve) over $N = 200$ iterations, as a function of the dimensionality $p$. ................................................................. 60

3.3 Averages and standard deviations of $\|\hat{\Sigma}_y^{-1} - \Sigma^{-1}\|$ (dashed curve) and $\|\hat{\Sigma}_y^{-1} - \Sigma^{-1}\|$ (solid curve) over $N = 200$ iterations, as a function of the dimensionality $p$. ................................................................. 61

3.4 Averages (left panel) and standard deviations (right panel) of the relative error $p^{-1/2}\|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - I_p\|_F$ with known factors ($\hat{\Sigma} = \hat{\Sigma}_{obs}$ solid red curve), POET ($\hat{\Sigma} = \hat{\Sigma}_R$ solid blue curve), and sample covariance ($\hat{\Sigma} = \hat{\Sigma}_{sam}$ dashed curve) over 200 simulations, as a function of the dimensionality $p$. Top panel: $p$ ranges in 20 to 600 with increment 20; bottom panel: $p$ ranges in 1 to 100 with increment 1. ........................... 67
3.5 Averages (left panel) and standard deviations (right panel) of $\|\hat{\Sigma}^{-1} - \Sigma^{-1}\|$ with known factors ($\hat{\Sigma} = \hat{\Sigma}_{\text{obs}}$ solid red curve), POET ($\hat{\Sigma} = \hat{\Sigma}_{\text{K}}$ solid blue curve), and sample covariance ($\hat{\Sigma} = \hat{\Sigma}_{\text{sam}}$ dashed curve) over 200 simulations, as a function of the dimensionality $p$. Top panel: $p$ ranges in 20 to 600 with increment 20; middle panel: $p$ ranges in 1 to 100 with increment 1; Bottom panel: the same as the top panel with dashed curve excluded. 68

3.6 Averages (left panel) and standard deviations (right panel) of $\|\hat{\Sigma} - \Sigma\|_{\text{max}}$ with known factors ($\hat{\Sigma} = \hat{\Sigma}_{\text{obs}}$ solid red curve), POET ($\hat{\Sigma} = \hat{\Sigma}_{\text{K}}$ solid blue curve), and sample covariance ($\hat{\Sigma} = \hat{\Sigma}_{\text{sam}}$ dashed curve) over 200 simulations, as a function of the dimensionality $p$. They are nearly indifferentiable. 69

3.7 Averages of $\|\hat{\Sigma} - \Sigma\|$ (left panel) and $\|\Sigma^{1/2}\hat{\Sigma}\Sigma^{-1/2} - I_p\|$ with known factors ($\hat{\Sigma} = \hat{\Sigma}_{\text{obs}}$ solid red curve), POET ($\hat{\Sigma} = \hat{\Sigma}_{\text{K}}$ solid blue curve), and sample covariance ($\hat{\Sigma} = \hat{\Sigma}_{\text{sam}}$ dashed curve) over 200 simulations, as a function of the dimensionality $p$. The three curves are hardly distinguishable on the left panel. 69

3.8 Robustness of $K$ as $p$ increases for various choices of $K$ (Design 1). Top left: $\|\hat{\Sigma}_{u,K}^T - \Sigma_u\|_F/\sqrt{p}$; top right: $\|(\hat{\Sigma}_{u,K}^T)^{-1} - \Sigma_u^{-1}\|_F/\sqrt{p}$; bottom left: $\|\hat{\Sigma} - \Sigma\|_\Sigma$; bottom right: $\|\hat{\Sigma}_K^{-1} - \Sigma^{-1}\|_F/\sqrt{p}$. $T$ is fixed to be 300. 71

3.9 Robustness of $K$ as $p$ increases for various choices of $K$ (Design 1, $T = 300$). Top left: $\|\hat{\Sigma}_{u,K}^T - \Sigma_u\|$; top right: $\|(\hat{\Sigma}_{u,K}^T)^{-1} - \Sigma_u^{-1}\|$; bottom left: $\|\hat{\Sigma}_K - \Sigma\|_\Sigma$; bottom right: $\|\hat{\Sigma}_K^{-1} - \Sigma^{-1}\|$. 72

3.10 Robustness of $K$ for various $p$. Design 2. 75
3.11 Box plots of regrets $R(\hat{w}) - R^*$ for $p = 40$, $p = 80$, $p = 120$ and $p = 140$. In each panel, the box plots from left to right correspond to $\hat{w}$ obtained using $\hat{\Sigma}$ based on approximate factor model, strict factor model, and sample covariance, respectively.

3.12 Estimation errors for risk assessments as a function of the portfolio size $p$. Left panel plots the average absolute error $|R(\hat{w}) - \hat{R}(\hat{w})|$ and right panel depicts the average relative error $|\hat{R}(\hat{w})/R(\hat{w}) - 1|$. Here, $\hat{w}$ and $\hat{R}$ are obtained based on three estimators of $\hat{\Sigma}$.

3.13 Heatmap of thresholded error correlation matrix for number of factors $K = 0$, $K = 1$, $K = 2$ and $K = 3$.

3.14 Risk of portfolios created with POET and SFM (strict factor model)
Chapter 1

Covariance Estimation in Approximate Factor Models with Observable Factors

This chapter has been joint work with Jianqing Fan and Yuan Liao. It has been published in the Annals of Statistics, for reference please look at Fan, Liao and Mincheva (2011).

1.1 Introduction

We consider a factor model defined as follows:

\[ y_{it} = b_i' f_t + u_{it}, \]  

(1.1.1)

where \( y_{it} \) is the observed datum for the \( i \)th \( (i = 1, ..., p) \) asset at time \( t = 1, ..., T \); \( b_i \) is a \( K \times 1 \) vector of factor loadings; \( f_t \) 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 $\Sigma_u$, the variance covariance matrix of the error term, and $\Sigma$, the variance-covariance matrix of $y_t = (y_{1t}, \ldots, y_{pt})'$. In the literature, $\Sigma = \text{cov}(y_t)$ was traditionally estimated by the sample covariance matrix of $y_t$:

$$
\Sigma_{\text{sam}} = \frac{1}{T-1} \sum_{t=1}^{T} (y_t - \bar{y})(y_t - \bar{y})',
$$

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 $T$, $\Sigma_{\text{sam}}$ becomes singular while $\Sigma$ 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 $\Sigma$, by incorporating the common factor structure. One of the key assumptions they made was the cross-sectional independence among the idiosyncratic components, which results $E u_t u_t'$ 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 text, we relax this assumption, and investigate the impact of the cross-sectional correlations of idiosyncratic noises on the estimation of $\Sigma$ and $\Sigma_u$, when both $p$ and $T$ are allowed to diverge. In particular, when estimating $\Sigma^{-1}$...
and $\Sigma_u^{-1}$, we allow $p$ to increase much faster than $T$, say, $\log p = O(T^\alpha)$, for some $\alpha \in (0,1)$.

Sparsity is one of the commonly used assumptions in the estimation of 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. In this text we assume that $\Sigma_u$ is sparse, and estimate both $\Sigma_u$ and $\Sigma_u^{-1}$ 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 $f_t$ 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 $\Sigma$ 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 the inverse matrices, $p$ is allowed to grow exponentially fast in $T$. In addition, we achieve better convergence rates than those in Fan, Fan and Lv (2008).

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 text is organized as follows. Section 2 provides the asymptotic theory for estimating the error covariance matrix and its inverse. Section 3 con-
siders estimating the covariance matrix of $y_t$. 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 text, we use $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ to denote the minimum and maximum eigenvalues of a matrix $A$. We also denote by $\|A\|_F$, $\|A\|$ and $\|A\|_\infty$ the Frobenius norm, operator norm and elementwise norm of a matrix $A$ respectively, defined respectively as $\|A\|_F = tr^{1/2}(A'A)$, $\|A\| = \lambda_{\max}^{1/2}(A'A)$, and $\|A\|_\infty = \max_{i,j} |A_{ij}|$. Note that, when $A$ is a vector, $\|A\|$ is equal to the Euclidean norm.
1.2 Estimation of Error Covariance Matrix

1.2.1 Adaptive thresholding

Consider the following approximate factor model, in which the cross-sectional correlation among the idiosyncratic error components is allowed:

\[ y_{it} = b_i'f_t + u_{it}, \quad (1.2.1) \]

where \( i = 1, \ldots, p \) and \( t = 1, \ldots, T; b_i \) is a \( K \times 1 \) vector of factor loadings; \( f_t \) is a \( K \times 1 \) vector of observable common factors, uncorrelated with \( u_{it} \). Write

\[ B = (b_1, \ldots, b_p)', \quad y_t = (y_{1t}, \ldots, y_{pt})', \quad u_t = (u_{1t}, \ldots, u_{pt})', \]

then model (1.2.1) can be written in a more compact form:

\[ y_t = Bf_t + u_t, \quad (1.2.2) \]

with \( E(u_t | f_t) = 0 \).

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

\[ \Sigma_u = \text{cov}(u_t), \]

itself is of interest for the inferential theory of factor models. For example, the asymptotic covariance of the least square estimator of \( B \) depends on \( \Sigma_u^{-1} \), and in simulating home price indices over a certain time horizon for mortgage based securities, a good estimate of \( \Sigma_u \) is needed. When \( p \) is close to or larger than \( T \), estimating \( \Sigma_u \) 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 \( \Sigma_u = (\sigma_{ij})_{p \times p} \). Define

\[ m_T = \max_{i \leq p} \sum_{j \leq p} I(\sigma_{ij} \neq 0). \tag{1.2.3} \]

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

\[ m_T^2 = o\left( \frac{T}{K^2 \log p} \right). \tag{1.2.4} \]

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 \( \Sigma_u \) are uniformly bounded across rows by a slowly growing sequence, for some \( q \in [0, 1) \). In contrast, the assumption we make in this text, i.e., \( q = 0 \), has 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 un-
correlated across industries, since the industry-specific components are not pervasive for the whole economy (Connor and Korajczyk (1993)).

We estimate \( \Sigma_u \) 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 \((X_1, \ldots, X_T)\) of a \( p \times 1 \) vector \( X \), which follows a multivariate Gaussian distribution \( N(0, \Sigma_X) \). The sample covariance matrix of \( X \) is thus given by:

\[
S_X = \frac{1}{T} \sum_{i=1}^{T} (X_i - \bar{X})(X_i - \bar{X})' = (s_{ij})_{p \times p}.
\]

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

\[
\| \hat{\Sigma}_X^T - \Sigma_X \| = O_p(\omega_T m_T).
\]

In the factor models, however, we do not observe the error term directly. Hence when estimating the error covariance matrix of a factor model, 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 \( \hat{b}_i \) be the ordinary least square (OLS) estimator of \( b_i \), and

\[
\hat{u}_{it} = y_{it} - \hat{b}'_i f_t.
\]

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

\[
\hat{\Sigma}_u = \frac{1}{T} \sum_{t=1}^{T} \hat{u}_t \hat{u}_t' = (\hat{\sigma}_{ij}).
\]
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 $\max_{ij} |s_{ij} - \Sigma_{X,ij}|$. This rate changes when $s_{ij}$ is replaced with the residual $\hat{u}_{ij}$, which will be slower if the number of common factors $K$ increases with $T$. Therefore, the thresholding value $\omega_T$ used in this text is adjusted to account for the effect of the estimation of the residuals.

1.2.2 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 text, we apply the adaptive thresholding estimator (Cai and Liu (2011)) to estimate the error covariance matrix, which is given by

$$
\hat{\Sigma}_u^T = (\hat{\sigma}_{ij}^T), \quad \hat{\sigma}_{ij}^T = \hat{\sigma}_{ij}I(|\hat{\sigma}_{ij}| \geq \sqrt{\hat{\theta}_{ij}\omega_T})
$$

$$
\hat{\theta}_{ij} = \frac{1}{T} \sum_{t=1}^{T} (\hat{u}_{it}\hat{u}_{jt} - \hat{\sigma}_{ij})^2, \quad (1.2.5)
$$

for some $\omega_T$ to be specified later.

We impose the following assumption:

**Assumption 1.2.1.** (i) $(u_1, ..., u_T)$ are independent and identically distributed with mean vector zero and covariance matrix $\Sigma_u$.

(ii) There exit constants $c_1, c_2 > 0$ such that $c_1 < \lambda_{\min}(\Sigma_u) \leq \lambda_{\max}(\Sigma_u) < c_2$.

(iii) There exist $r > 0$ and $b > 0$, such that for any $s > 0$ and $i \leq p$,

$$
P(|u_{it}| > s) \leq \exp(-(s/b)^r). \quad (1.2.6)
$$
Condition (iii) requires the distributions of \((u_1, ..., u_p)\) to have exponential-type tails, which allows us to apply the large deviation theory to \(\frac{1}{T} \sum_{t=1}^{T} u_{it} u_{jt}\). Condition (ii) requires the nonsingularity of \(\Sigma_u\). Note that Cai and Liu (2011) allowed \(\max_j \sigma_{jj}\) to diverge when direct observations are available. Condition (ii), however, requires that \(\sigma_{jj}\) 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 \(y_t\) later. This assumption is satisfied by most of the applications of factor models.

Suppose there exists a positive sequence \(a_T\) such that

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

(1.2.7)

The following theorem establishes the asymptotic properties of the thresholding estimator \(\hat{\Sigma}_u\), based on observations with estimation errors.

**Theorem 1.2.1.** Let \(\hat{\Sigma}_u^T\) be defined as in (1.2.5) 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 2.3.2,

(i)

\[
\|\hat{\Sigma}_u^T - \Sigma_u\| = O_p(m_T \omega_T),
\]

(ii) \(\hat{\Sigma}_u^T\) is positive definite, and

\[
\|(\hat{\Sigma}_u^T)^{-1} - \Sigma_u^{-1}\| = O_p(m_T \omega_T).
\]

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 averaged estimation error of the residual terms. We will see in the next section that when the number of common factors $K$ increases slowly, the convergence rate in Theorem 1.2.1 is close to the minimax optimal rate as in Cai and Zhou (2010).
1.3 Estimation of Covariance Matrix Using Factors

We now investigate the estimation of the covariance matrix $\Sigma$ in the approximate factor model:

$$y_t = Bf_t + u_t,$$

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

Note that

$$\Sigma = B\text{cov}(f_t)B' + \Sigma_u.$$

By the Sherman-Morrison-Woodbury formula,

$$\Sigma^{-1} = \Sigma_u^{-1} - \Sigma_u^{-1}B[\text{cov}(f_t)^{-1} + B'\Sigma_u^{-1}B]^{-1}B'\Sigma_u^{-1}.$$

When the factors are observable, one can estimate $B$ by the least squares method:

$$\hat{B} = (\hat{b}_1, ..., \hat{b}_p)'$$

where,

$$\hat{b}_i = \arg \min_{b_i} \frac{1}{Tp} \sum_{t=1}^{T} \sum_{i=1}^{p} (y_{it} - b_i'f_t)^2.$$

The covariance matrix $\text{cov}(f_t)$ can be estimated by the sample covariance matrix

$$\hat{\text{cov}}(f_t) = T^{-1}XX' - T^{-2}X11'X',$$

where $X = (f_1, ..., f_T)$, and $1$ is a $T$-dimensional column vector of ones. Therefore, by employing the thresholding estimator $\hat{\Sigma}_u^T$ in (1.2.5), we obtain substitution estimators

$$\hat{\Sigma}^T = B\hat{\text{cov}}(f_t)\hat{B}' + \hat{\Sigma}_u^T,$$
\[
(\hat{\Sigma}^T)^{-1} = (\hat{\Sigma}_u^T)^{-1} - (\hat{\Sigma}_u^T)^{-1}\hat{\mathcal{B}}[\hat{\mathcal{B}}^\prime (\hat{\Sigma}_u^T)^{-1}\hat{\mathcal{B}}]^{-1}\hat{\mathcal{B}}^\prime (\hat{\Sigma}_u^T)^{-1}.
\]

Fan, Fan and Lv (2008) obtained an upper bound of \(\|\hat{\Sigma}^T - \Sigma\|_F\) under the Frobenius norm when \(\Sigma_u\) 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 \(y_t\), 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):

\[
\|\hat{\Sigma}^T - \Sigma\|_\Sigma = \left(p^{-1}\text{tr}(\hat{\Sigma}^T\Sigma^{-1} - I)^2\right)^{1/2} = p^{-1/2}\|\Sigma^{-1/2}(\hat{\Sigma}^T - \Sigma)\Sigma^{-1/2}\|_F.
\]

Here the factor \(p^{-1/2}\) is used for normalization, such that \(\|\Sigma\|_\Sigma = 1\). 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 \(\hat{\Sigma}^T\) is consistent even if \(p > T\) under norm \(\|.\|_\Sigma\).

The following assumptions are made.

**Assumption 1.3.1.** (i) \(\{f_t\}_{t \geq 1}\) is stationary and ergodic.

(ii) \(\{u_t\}_{t \geq 1}\) and \(\{f_t\}_{t \geq 1}\) 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 \(\sigma\)-algebras generated by \(\{(f_t, u_t) : -\infty \leq t \leq 0\}\) and \(\{(f_t, u_t) : T \leq t \leq \infty\}\)
respectively. In addition, define the mixing coefficient

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

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

**Assumption 1.3.2.** There exist positive constants \( \gamma \) 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 1.3.3.** (i) There exists a constant \( M > 0 \) such that for all \( i,t \), \( Ey_{it}^2 < M, \|B\|_\infty < M, \) and \( E\|f_t\|^4 \leq K^2M. \)

(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 1.3.4.** There exists a constant \( C > 0 \) such that \( \lambda_{\min}(\Sigma) > C \), and \( \lambda_{\min}(\text{cov}(f_t)) > C \).

**Assumption 1.3.5.** \( \|p^{-1}B'B - \Omega\| = o(1) \) for some \( K \times K \) symmetric positive definite matrix \( \Omega \) such that \( \lambda_{\min}(\Omega) \) is bounded away from zero.

Assumption 1.3.4 ensures that \( \Sigma \) and \( \text{cov}(f_t) \) are not ill-conditioned, which is needed to derive the convergence rate of \( \|\hat{\Sigma}^T - \Sigma^{-1}\| \) below. Assumption 2.3.1 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 $\|\hat{\Sigma}^T - \Sigma^{-1}\|$, as it ensures that $\lambda_{\max}((B\Sigma^{-1}B)^{-1}) = O(p^{-1})$.

The first result in this Section is an application of Theorem 1.2.1.

**Theorem 1.3.1.** Suppose $\max\{(\log p)^{1/r + 4/r_2}, K^4(\log p)^2\} = o(T)$. Under Assumptions 2.3.2, 1.3.1-2.3.4, the adaptive thresholding estimator defined in (1.2.5) with $\omega^2_T = \frac{K^2 \log p}{T}$ satisfies

$$\|\hat{\Sigma}^T_u - \Sigma_u\| = O_p\left(mTK\sqrt{\frac{\log p}{T}}\right),$$

and

$$\|\hat{\Sigma}_u^{-1} - \Sigma_u^{-1}\| = O_p\left(mTK\sqrt{\frac{\log p}{T}}\right).$$

**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 $\Sigma_u$, 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 $B\text{cov}(f_t)B'$, Theorem 1.3.1 implies the main theorem in this section:

**Theorem 1.3.2.** Suppose $\max\{\log p^{4/r+4/\gamma-1}, K^4(\log p)^2\} = o(T)$. Under Assumptions 2.3.2, 1.3.1-2.3.1, we have

\[
\|\hat{\Sigma}^T - \Sigma\|_F^2 = O_p\left(\frac{pK^2}{T^2} + \frac{m^2K^2\log p}{T}\right),
\]

\[
\|\left(\hat{\Sigma}^T\right)^{-1} - \Sigma^{-1}\|^2 = O_p\left(\frac{m^2K^2\log p}{T}\right),
\]

(1.3.1)

and

\[
\|\hat{\Sigma}^T - \Sigma\|^2_\infty = O_p\left(\frac{K^6 \log p}{T}\right).
\]

(1.3.2)

Note that we have derived a better convergence rate of $(\hat{\Sigma}^T)^{-1}$ 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 $(\hat{\Sigma}^T)^{-1}$ to be consistent.

We have also derived the maximum elementwise estimation $\|\hat{\Sigma}^T - \Sigma\|_\infty$. This quantity appears in risk assessment as in Fan, Zhang and Yu (2008). For any portfolio with allocation vector $w$, the true portfolio variance and the estimated one are given by $w'\Sigma w$ and $w'\hat{\Sigma}^T w$ respectively. The estimation error is bounded by

\[
|w'\hat{\Sigma}^T w - w'\Sigma w| \leq \|\hat{\Sigma}^T - \Sigma\|_\infty \|w\|_1^2,
\]

where $\|w\|_1$, the $l_1$ norm of $w$, is the gross exposure of the portfolio.
1.4 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

\[ y_{it} = b_i'f_{it} + u_{it}, \quad i \leq p, t \leq T, \quad (1.4.1) \]

where \( b_i \) and \( f_{it} \) are both \( K_i \times 1 \) vectors. The \( p \) linear equations (1.4.1) are related because their error terms \( u_{it} \) are correlated, i.e., the covariance matrix

\[ \Sigma_u = (Eu_{it}u_{jt})_{p \times p} \]

is not diagonal.

Model (1.4.1) 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 \( f_{it} = f_i \) for each \( i \leq p \), model (1.4.1) reduces to the approximate factor model (1.1) with common factors \( f_t \).

Under mild conditions, running OLS on each equation produces unbiased and consistent estimator of \( b_i \) 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

\[
\begin{align*}
\mathbf{y}_i &= (y_{i1}, ..., y_{iT})', T \times 1, \\
\mathbf{X}_i &= (f_{i1}, ..., f_{iT})', T \times K_i, \quad i \leq p, \\
\mathbf{y} &= \begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} X_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & X_p \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} b_1 \\ \vdots \\ b_p \end{pmatrix}.
\end{align*}
\]
The GLS estimator of $\mathbf{B}$ is given by Zellner (1962):

$$
\hat{\mathbf{B}}_{GLS} = \left[ \mathbf{X}'(\hat{\mathbf{\Sigma}}_u^{-1} \otimes \mathbf{I}_T)^{-1}\mathbf{X} \right]^{-1}\left[ \mathbf{X}'(\hat{\mathbf{\Sigma}}_u^{-1} \otimes \mathbf{I}_T)^{-1}\mathbf{y} \right], \quad (1.4.2)
$$

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

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

$$
\hat{\mathbf{u}}_it = \mathbf{y}_it - \hat{\mathbf{b}}_i'\mathbf{f}_it. \quad (1.4.3)
$$

On the second stage, estimate $\Sigma_u$ by

$$
\hat{\Sigma}_u = (\hat{\sigma}_{ij}) = \left( \frac{1}{T} \sum_{t=1}^{T} \hat{\mathbf{u}}_it\hat{\mathbf{u}}_jt \right)_{p \times p}. \quad (1.4.4)
$$

In high dimensional seemingly unrelated regression in which $p > T$, however, $\hat{\Sigma}_u$ is not invertible, and hence the GLS estimator (1.4.2) is infeasible.

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

**Assumption 1.4.1.** For each $i \leq p$,

(i) $\{\mathbf{f}_it\}_{t \geq 1}$ is stationary and ergodic.

(ii) $\{\mathbf{u}_i\}_{t \geq 1}$ and $\{\mathbf{f}_it\}_{t \geq 1}$ are independent.

**Assumption 1.4.2.** There exists positive constants $C$ and $\gamma$ such that for each $i \leq p$, the strong mixing condition in Assumption 2.3.3 is satisfied by $(\mathbf{f}_it, \mathbf{u}_i)$.
**Assumption 1.4.3.** There exist constants $M$ and $C > 0$ such that for all $i, j, t$

(i) $Ey_{it}^2 < M, \|B\|_{\infty} < M$, and $E\|f_{it}\|^4 < K_i^2 M$.

(ii) $\lambda_{\min}(\Sigma) > C$, and $\min_{i \leq p} \lambda_{\min}(\text{cov}(f_{it})) > C$.

**Assumption 1.4.4.** 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 1.3.1. It shows that the adaptive thresholding technique (1.2.5) produces a consistent nonsingular estimator of $\hat{\Sigma}_u$.

**Theorem 1.4.1.** Let $K = \max_{i \leq p} K_i$; suppose $\max\{(\log p)^{4/r+4/r_3-1}, K^4(\log p)^2\} = o(T)$. Under Assumptions 2.3.2, 1.4.1-1.4.4, the adaptive thresholding estimator defined in (1.4.4) and (1.2.5) with $\omega_T^2 = \frac{K^2 \log p}{T}$ satisfies

$$\|\hat{\Sigma}_u^T - \Sigma_u\| = O_p \left( mTK \sqrt{\frac{\log p}{T}} \right),$$

and

$$\|(\hat{\Sigma}_u^T)^{-1} - \Sigma_u^{-1}\| = O_p \left( mTK \sqrt{\frac{\log p}{T}} \right).$$

Therefore, in the case when $p > T$, Theorem 1.4.1 enables us to efficiently estimate $B$ via feasible GLS:

$$\hat{B}_{GLS}^T = [X'((\hat{\Sigma}_u^T)^{-1} \otimes I_T)^{-1}X]^{-1}[X'((\hat{\Sigma}_u^T)^{-1} \otimes I_T)^{-1}y].$$
Chapter 2

Covariance Estimation by Thresholding Principal Orthogonal Complements (a.k.a. in Approximate Factor Models with Unobservable Factors)

This chapter has been joint work with Jianqing Fan and Yuan Liao. It has been published in the Journal of the Royal Statistical Society, Series B, for reference please look at Fan, Liao and Mincheva (2013).

2.1 Introduction

Information and technology make large data sets widely available for scientific discovery. Much statistical analysis of such high-dimensional data involves the estimation of a covariance matrix or its inverse (the precision matrix). Examples include portfolio
management and risk assessment (Fan, Fan and Lv, 2008), high-dimensional classification such as Fisher discriminant (Hastie, Tibshirani and Friedman, 2009), graphic models (Meinshausen and Bühlmann, 2006), statistical inference such as controlling false discoveries in multiple testing (Leek and Storey, 2008; Efron, 2010), finding quantitative trait loci based on longitudinal data (Yap, Fan, and Wu, 2009; Xiong et al. 2011), and testing the capital asset pricing model (Sentana, 2009), among others. See Section 5 for some of those applications. Yet, the dimensionality is often either comparable to the sample size or even larger. In such cases, the sample covariance is known to have poor performance (Johnstone, 2001), and some regularization is needed.

Realizing the importance of estimating large covariance matrices and the challenges brought by the high dimensionality, in recent years researchers have proposed various regularization techniques to consistently estimate $\Sigma$. One of the key assumptions is that the covariance matrix is sparse, namely, many entries are zero or nearly so (Bickel and Levina, 2008, Rothman et al, 2009, Lam and Fan 2009, Cai and Zhou, 2010, Cai and Liu, 2011). In many applications, however, the sparsity assumption directly on $\Sigma$ is not appropriate. For example, financial returns depend on the equity market risks, housing prices depend on the economic health, gene expressions can be stimulated by cytokines, among others. Due to the presence of common factors, it is unrealistic to assume that many outcomes are uncorrelated. An alternative method is to assume a factor model structure, as in Fan, Fan and Lv (2008). However, they restrict themselves to the strict factor models with known factors.

A natural extension is the conditional sparsity. Given the common factors, the outcomes are weakly correlated. In order to do so, we consider an approximate factor model, which has been frequently used in economic and financial studies (Chamberlain
Here $y_{it}$ is the observed response for the $i$th ($i = 1, ..., p$) individual at time $t = 1, ..., T$; $b_i$ is a vector of factor loadings; $f_t$ is a $K \times 1$ vector of common factors, and $u_{it}$ is the error term, usually called *idiosyncratic component*, uncorrelated with $f_t$. Both $p$ and $T$ diverge to infinity, while $K$ is assumed fixed throughout the text, and $p$ is possibly much larger than $T$.

We emphasize that in model (2.1.1), only $y_{it}$ is observable. It is intuitively clear that the unknown common factors can only be inferred reliably when there are sufficiently many cases, that is, $p \to \infty$. In a data-rich environment, $p$ can diverge at a rate faster than $T$. The factor model (2.1.1) can be put in a matrix form as

$$y_t = B f_t + u_t. \quad (2.1.2)$$

where $y_t = (y_{1t}, ..., y_{pt})'$, $B = (b_1, ..., b_p)'$ and $u_t = (u_{1t}, ..., u_{pt})'$. We are interested in $\Sigma$, the $p \times p$ covariance matrix of $y_t$, and its inverse, which are assumed to be time-invariant. Under model (2.1.1), $\Sigma$ is given by

$$\Sigma = B \text{cov}(f_t) B' + \Sigma_u, \quad (2.1.3)$$

where $\Sigma_u = (\sigma_{u,ij})_{p \times p}$ is the covariance matrix of $u_t$. The literature on approximate factor models typically assumes that the first $K$ eigenvalues of $B \text{cov}(f_t) B'$ diverge at rate $O(p)$, whereas all the eigenvalues of $\Sigma_u$ are bounded as $p \to \infty$. This assumption holds easily when the factors are pervasive in the sense that a non-negligible fraction of factor loadings should be non-vanishing. The decomposition (2.1.3) is then asymptotically identified as $p \to \infty$. In addition to it, in this text we assume that $\Sigma_u$ is
approximately sparse as in Bickel and Levina (2008) and Rothman et al. (2009): for some $q \in [0, 1)$,

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

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

The conditional sparsity structure of (2.1.2) was explored in Chapter 1 of this thesis, in estimating the covariance matrix, when the factors $\{f_t\}$ are observable. As a reminder, this allowed us to use regression analysis to estimate $\{u_t\}_{t=1}^T$. This chapter deals with the situation in which the factors are unobservable and have to be inferred. Our approach is simple, optimization-free and it uses the data only through the sample covariance matrix. Run the singular value decomposition on the sample covariance matrix $\hat{\Sigma}_{sam}$ of $y_t$, keep the covariance matrix formed by the first $K$ principal components, and apply the thresholding procedure to the remaining covariance matrix. This results in a Principal Orthogonal complement Thresholding (POET) estimator. When the number of common factors $K$ is unknown, it can be estimated from the data. See Section 2 for additional details. We will investigate various properties of POET under the assumption that the data are serially dependent, which includes independent observations as a specific example. The rate of convergence under various norms for both estimated $\Sigma$ and $\Sigma_u$ and their precision (inverse) matrices will be derived. We show that the effect of estimating the unknown factors on the rate of convergence vanishes when $p \log p \gg T$, and in particular, the rate of convergence for $\Sigma_u$ achieves the optimal rate in Cai and Zhou (2012).

This text focuses on the high-dimensional static factor model (2.1.2), which is innately related to the principal component analysis (PCA), as clarified in Section 2. This feature makes it different from the classical factor model with fixed dimensionality (e.g., Lawley and Maxwell 1971). In the last ten years, much theory on the
estimation and inference of the static factor model has been developed, for example, Stock and Watson (1998, 2002), Bai and Ng (2002), Bai (2003), Doz, Giannone and Reichlin (2011), among others. Our contribution is on the estimation of covariance matrices and their inverse in large factor models.

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

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

There is a big difference between our model and those considered in the aforementioned literature. In the current project, the first $K$ eigenvalues of $\Sigma$ are spiked and grow at a rate $O(p)$, whereas the eigenvalues of the matrices studied in the existing literature on covariance estimation are usually assumed to be either bounded or slowly growing. Due to this distinctive feature, the common components and the idiosyncratic components can be identified, and in addition, PCA on the sample covariance matrix can consistently estimate the space spanned by the eigenvectors of $\Sigma$. The existing methods of either thresholding directly or solving a constrained optimization method can fail in the presence of very spiked principal eigenvalues. However, there is a price to pay here: as the first $K$ eigenvalues are “too spiked”, one can hardly obtain a satisfactory rate of convergence for estimating $\Sigma$ in absolute term, but it can be estimated accurately in relative term (see Section 3.3 for details). In addition, $\Sigma^{-1}$ can be estimated accurately.

We would like to further note that the low-rank plus sparse representation of our model is on the population covariance matrix, whereas Candès et al. (2011), Wright et al. (2009), Lin et al. (2009) considered such a representation on the data matrix. As there is no $\Sigma$ to estimate, their goal is limited to producing a low-rank plus sparse matrix decomposition of the data matrix, which corresponds to the identifiability issue of our study, and does not involve estimation and inference. In contrast, our ultimate goal is to estimate the population covariance matrices as well as the precision matrices. For this purpose, we require the idiosyncratic components and common factors to be uncorrelated and the data generating process to be strictly stationary. The covariances considered in this project are constant over time, though slow-time-varying covariance matrices are applicable through localization in time (time-domain smoothing). Our consistency result on $\Sigma_u$ demonstrates that the decomposition (2.1.3) is
identifiable, and hence our results also shed the light of the “surprising phenomenon” of Candès et al. (2011) that one can separate fully a sparse matrix from a low-rank matrix when only the sum of these two components is available.

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


2.2 Regularized Covariance Matrix via PCA

There are three main objectives of this section: (i) understand the relationship between principal component analysis (PCA) and the high-dimensional factor analysis; (ii) estimate both covariance matrices $\Sigma$ and the idiosyncratic $\Sigma_u$ and their precision matrices in the presence of common factors, and (iii) investigate the impact of estimating the unknown factors on the covariance estimation. The propositions in Section 2.2.1 below show that the space spanned by the principal components in the population level $\Sigma$ is close to the space spanned by the columns of the factor loading matrix $B$.

2.2.1 High-dimensional PCA and factor model

Consider a factor model

$$y_{it} = b_i'f_t + u_{it}, \quad i \leq p, t \leq T,$$

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

**Example 2.2.1.** Consider a single-factor model $y_{it} = b_i f_t + u_{it}$ where $b_i \in \mathbb{R}$. Suppose that the factor is pervasive in the sense that it has non-negligible impact on a non-vanishing proportion of outcomes. It is then reasonable to assume $\sum_{i=1}^p b_i^2 > cp$ for some $c > 0$. Therefore, assuming that $\lambda_{\max}(\Sigma_u) = o(p)$, an application of (2.1.3) yields,

$$\lambda_{\max}(\Sigma) \geq \text{var}(f_t) \sum_{i=1}^p b_i^2 - \lambda_{\max}(\Sigma_u) > \frac{c}{2} \text{var}(f_t)p$$
for all large $p$, assuming $\text{var}(f_i) > 0$.

We now elucidate why PCA can be used for the factor analysis in the presence of spiked eigenvalues. Write $B = (b_1, ..., b_p)'$ as the $p \times K$ loading matrix. Note that the linear space spanned by the first $K$ principal components of $B\text{cov}(f_i)B'$ is the same as that spanned by the columns of $B$ when $\text{cov}(f_i)$ is non-degenerate. Thus, we can assume without loss of generality that the columns of $B$ are orthogonal and $\text{cov}(f_i) = I_K$, the identity matrix. This canonical form corresponds to the identifiability condition in decomposition (2.1.3). Let $\tilde{b}_1, \cdots, \tilde{b}_K$ be the columns of $B$, ordered such that $\{\|\tilde{b}_j\|\}_{j=1}^K$ is in a non-increasing order. Then, $\{\tilde{b}_j/\|\tilde{b}_j\|\}_{j=1}^K$ are eigenvectors of the matrix $BB'$ with eigenvalues $\{\|\tilde{b}_j\|^2\}_{j=1}^K$ and the rest zero. We will impose the pervasiveness assumption that all eigenvalues of the $K \times K$ matrix $p^{-1}B'B$ are bounded away from zero, which holds if the factor loadings $\{b_i\}_{i=1}^p$ are independent realizations from a non-degenerate population. Since the non-vanishing eigenvalues of the matrix $BB'$ are the same as those of $B'B$, from the pervasiveness assumption it follows that $\{\|\tilde{b}_j\|^2\}_{j=1}^K$ are all growing at rate $O(p)$.

Let $\{\lambda_j\}_{j=1}^p$ be the eigenvalues of $\Sigma$ in a descending order and $\{\xi_j\}_{j=1}^p$ be their corresponding eigenvectors. Then, an application of Weyl’s eigenvalue theorem (see the appendix) yields that

**Proposition 2.2.1.** Assume that the eigenvalues of $p^{-1}B'B$ are bounded away from zero for all large $p$. For the factor model (2.1.3) with the canonical condition

$$\text{cov}(f_i) = I_K \text{ and } B'B \text{ is diagonal,}$$

we have

$$|\lambda_j - \|\tilde{b}_j\|^2| \leq \|\Sigma_u\|, \quad \text{for } j \leq K, \quad |\lambda_j| \leq \|\Sigma_u\|, \quad \text{for } j > K.$$
In addition, for \( j \leq K \), \( \liminf_{p \to \infty} \| \tilde{b}_j \|_2^2 / p > 0 \).

Using Proposition 2.2.1 and the sin \( \theta \) theorem of Davis and Kahn (1970, see the appendix), we have the following:

**Proposition 2.2.2.** Under the assumptions of Proposition 2.2.1, if \( \{ \| \tilde{b}_j \| \}_{j=1}^K \) are distinct, then

\[
\| \xi_j - \tilde{b}_j / \| \tilde{b}_j \| \| = O(p^{-1} \| \Sigma_u \|), \quad \text{for} \quad j \leq K.
\]

Propositions 2.2.1 and 2.2.2 state that PCA and factor analysis are approximately the same if \( \| \Sigma_u \| = o(p) \). This is assured through a sparsity condition on \( \Sigma_u = (\sigma_{u,ij})_{p \times p} \), which is frequently measured through

\[
m_p = \max_{i \leq p} \sum_{j \leq p} |\sigma_{u,ij}|^q, \quad \text{for some} \quad q \in [0, 1]. \tag{2.2.2}
\]

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

\[
\| \Sigma_u \| \leq \| \Sigma_u \|_1 \leq \max_{i} \sum_{j=1}^{p} |\sigma_{u,ij}|^q (\sigma_{u,ii} \sigma_{u,jj})^{(1-q)/2} = O(m_p),
\]

if the noise variances \( \{ \sigma^2_{u,ii} \} \) are bounded. Therefore, when \( m_p = o(p) \), Proposition 2.2.1 implies that we have distinguished eigenvalues between the principal components \( \{ \lambda_j \}_{j=1}^K \) and the rest of the components \( \{ \lambda_j \}_{j=K+1}^p \) and Proposition 2.2.2 ensures that the first \( K \) principal components are approximately the same as the columns of the factor loadings.

The aforementioned sparsity assumption appears reasonable in empirical applications. Boivin and Ng (2006) conducted an empirical study and showed that imposing zero correlation between weakly correlated idiosyncratic components improves fore-
More recently, Phan (2012) empirically estimated the level of sparsity of the idiosyncratic covariance using the UK market data.

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

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

implies that the first $K$ eigenvalues are growing at a rate $p$. Moreover, when $p$ is large, the principal components $\{\xi_j\}_{j=1}^K$ are close to the normalized vectors $\{\tilde{b}_j\}_{j=1}^K$ when $m_p = o(p)$. This provides the mathematics for using the first $K$ principal components as a proxy of the space spanned by the columns of the factor loading matrix $B$. In addition, due to (2.2.3), the signals of the first $K$ eigenvalues are stronger than those of the spiked covariance model considered by Jung and Marron (2009) and Birnbaum et al. (2012). Therefore, our other conditions for the consistency of principal components at the population level are much weaker than those in the spiked covariance literature. On the other hand, this also shows that, under our setting the PCA is a valid approximation to factor analysis only if $p \to \infty$. The fact that the PCA on the sample covariance is inconsistent when $p$ is bounded was also previously demonstrated in the literature (See e.g., Bai (2003)).

With assumption (2.2.3), the standard literature on approximate factor models has shown that the PCA on the sample covariance matrix $\tilde{\Sigma}_{sam}$ can consistently estimate the space spanned by the factor loadings (e.g., Stock and Watson (1998), Bai (2003)).

Our contribution in Propositions 2.1 and 2.2 is that we connect the high-dimensional

\footnote{We thank a referee for this interesting reference.}
factor model to the principal components, and obtain the consistency of the spectrum in the population level $\Sigma$ instead of the sample level $\hat{\Sigma}_{sam}$. The spectral consistency also enhances the results in Chamberlain and Rothschild (1983). This provides the rationale behind the consistency results in the factor model literature.

2.2.2 POET

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

$$\hat{\Sigma}_{sam} = \sum_{i=1}^{K} \hat{\lambda}_i \hat{\xi}_i \hat{\xi}_i' + \sum_{i=K+1}^{p} \hat{\lambda}_i \hat{\xi}_i \hat{\xi}_i' = \sum_{i=1}^{K} \hat{\lambda}_i \hat{\xi}_i \hat{\xi}_i' + \hat{R}_K,$$

(2.2.4)

where $\hat{R}_K = \sum_{i=K+1}^{p} \hat{\lambda}_i \hat{\xi}_i \hat{\xi}_i'$ is the principal orthogonal complement, and $K$ is the number of diverging eigenvalues of $\Sigma$. Let us first assume $K$ is known.

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

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

(2.2.5)

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

$$\tau_{ij} = \tau (\hat{r}_{ii} \hat{r}_{jj})^{1/2}, \text{ for a given } \tau > 0$$  \hspace{1cm} (2.2.6)$$

where $\hat{r}_{ii}$ is the $i^{th}$ diagonal element of $\hat{R}_K$. This corresponds to applying the thresholding with parameter $\tau$ to the correlation matrix of $\hat{R}_K$.

The estimator of $\Sigma$ is then defined as:

$$\hat{\Sigma}_K = \sum_{i=1}^{K} \hat{\lambda}_i \hat{\xi}_i \hat{\xi}_i' + \hat{R}_K^T.$$  \hspace{1cm} (2.2.7)

We will call this estimator the Principal Orthogonal complement thresholding (POET) estimator. It is obtained by thresholding the remaining components of the sample covariance matrix, after taking out the first $K$ principal components. One of the attractiveness of POET is that it is optimization-free, and hence is computationally appealing.  \hspace{1cm} (2)\hspace{1cm}

With the choice of $\tau_{ij}$ in (2.2.6) and the hard thresholding rule, our estimator encompasses many popular estimators as its specific cases. When $\tau = 0$, the estimator is the sample covariance matrix and when $\tau = 1$, the estimator becomes that based on the strict factor model (Fan, Fan, and Lv, 2008). When $K = 0$, our estimator is the same as the thresholding estimator of Bickel and Levina (2008) and (with a more general thresholding function) Rothman et al. (2009) or the adaptive thresholding estimator of Cai and Liu (2011) with a proper choice of $\tau_{ij}$.

In practice, the number of diverging eigenvalues (or common factors) can be estimated based on the sample covariance matrix. Determining $K$ in a data-driven way is an important topic, and is well understood in the literature. We will describe the POET with a data-driven $K$ in Section 2.2.4.

---

\textsuperscript{2}We have written an R package for POET, which outputs the estimated $\Sigma$, $\Sigma_u$, $K$, the factors and loadings.
There has been a growing literature on estimating a low-rank plus sparse matrix recently, e.g., Luo (2011), Agarwal et al. (2012), Pati et al. (2012). In contrast to these existing methods, besides being optimization-free, our proposed POET handles the covariance estimation when the eigenvalues grow fast with the dimensionality. Moreover, another class of literature on recovering the low-rank plus sparse decomposition, e.g., Candès et al. (2011), Lin et al. (2009) assumes such a decomposition on the data matrix instead of on $\Sigma$. As there is no $\Sigma$ to be estimated, they aim at exact recovery of the low rank and sparse parts and hence their problem is different from ours. Indeed, when applied to the population covariance matrix, their problem becomes whether one can separate the sparse covariance matrix from the low rank covariance matrix based on their summation, namely whether one can determine $\Sigma_u$ from $\Sigma$ in (2.1.3). This corresponds to the identifiability issue in our study.

### 2.2.3 Least squares point of view

The POET (2.2.7) has an equivalent representation using a constrained least squares method. The least squares method seeks for $\hat{\Lambda}_K = (\hat{b}_1^K, ..., \hat{b}_p^K)'$ and $\hat{F}_K = (\hat{f}_1^K, ..., \hat{f}_T^K)$ such that

$$
(\hat{\Lambda}_K, \hat{F}_K) = \arg \min_{b_i \in \mathbb{R}^K, f_t \in \mathbb{R}^K} \sum_{i=1}^{p} \sum_{t=1}^{T} (y_{it} - b_i^t f_t)^2,
$$

subject to the normalization

$$
\frac{1}{T} \sum_{i=1}^{T} f_i f_t' = I_K, \quad \text{and} \quad \frac{1}{p} \sum_{i=1}^{p} b_i b_i' \text{ is diagonal}.
$$

The constraints (2.2.9) correspond to the normalization (2.2.1). Here we assume that the mean of each variable \( \{y_{it}\}_{t=1}^{T} \) has been removed, that is, $E y_{it} = E f_{jt} = 0$ for all $i \leq p, j \leq K$ and $t \leq T$. Putting it in a matrix form, the optimization problem can
be written as

\[
\begin{align*}
\arg \min_{B \mathbf{F}} & \| \mathbf{Y} - \mathbf{B} \mathbf{F}' \|_F^2, \\
T^{-1} \mathbf{F} \mathbf{F}' &= \mathbf{I}_K, \quad \mathbf{B}' \mathbf{B} \text{ is diagonal.}
\end{align*}
\] (2.2.10)

where \( \mathbf{Y} = (\mathbf{y}_1, ..., \mathbf{y}_T) \) and \( \mathbf{F}' = (\mathbf{f}_1, \cdots, \mathbf{f}_T) \). For each given \( \mathbf{F} \), the least-squares estimator of \( \mathbf{B} \) is \( \mathbf{\Lambda} = T^{-1} \mathbf{Y} \mathbf{F}' \), using the constraint (2.2.9) on the factors. Substituting this into (2.2.10), the objective 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 \( \hat{\mathbf{F}}_K / \sqrt{T} \) are the eigenvectors corresponding to the \( K \) largest eigenvalues of the \( T \times T \) matrix \( \mathbf{Y}' \mathbf{Y} \) and \( \hat{\mathbf{\Lambda}}_K = T^{-1} \mathbf{Y} \hat{\mathbf{F}}_K \) (see e.g., Stock and Watson (2002)).

We will show that under some mild regularity conditions, as \( p \) and \( T \to \infty \), \( \hat{\mathbf{b}}_i \mathbf{f}_t \) consistently estimates the true \( \mathbf{b}_i \mathbf{f}_t \) uniformly over \( i \leq p \) and \( t \leq T \). Since \( \Sigma_u \) is assumed to be sparse, we can construct an estimator of \( \Sigma_u \) using the adaptive thresholding method by Cai and Liu (2011) as follows. Let \( \hat{\mathbf{u}}_{it} = \mathbf{y}_{it} - \hat{\mathbf{b}}_i \mathbf{f}_t \), \( \hat{\sigma}_{ij} = \frac{1}{T} \sum_{t=1}^T \hat{\mathbf{u}}_{it} \hat{\mathbf{u}}_{jt} \), and \( \hat{\theta}_{ij} = \frac{1}{T} \sum_{t=1}^T \left( \hat{\mathbf{u}}_{it} \hat{\mathbf{u}}_{jt} - \hat{\sigma}_{ij} \right)^2 \). For some pre-determined decreasing sequence \( \omega_T > 0 \), and large enough \( C > 0 \), define the adaptive threshold parameter as \( \tau_{ij} = C \sqrt{\hat{\theta}_{ij} \omega_T} \). The estimated idiosyncratic covariance estimator is then given by

\[
\hat{\Sigma}^T_{u,K} = (\hat{\sigma}^T_{ij})_{p \times p}, \quad \hat{\sigma}^T_{ij} = \begin{cases} 
\hat{\sigma}_{ij}, & i = j \\
\hat{s}_{ij}(\hat{\sigma}_{ij}), & i \neq j,
\end{cases}
\] (2.2.11)

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

\[
\hat{s}_{ij}(z) = 0 \text{ when } |z| \leq \tau_{ij}, \quad |\hat{s}_{ij}(z) - z| \leq \tau_{ij}.
\]
It is easy to verify that \( s_{ij}(\cdot) \) includes many interesting thresholding functions such as the hard thresholding \( (s_{ij}(z) = zI(|z| \geq \tau_{ij})) \), soft thresholding \( (s_{ij}(z) = \text{sign}(z)(|z| - \tau_{ij})+) \), SCAD, and adaptive lasso (See Rothman et al. (2009)).

Analogous to the decomposition (2.1.3), we obtain the following substitution estimators

\[
\tilde{\Sigma}_K = \hat{\Lambda}_K \hat{\Lambda}'_K + \hat{\Sigma}_{u,K},
\]

and by the Sherman-Morrison-Woodbury formula, noting that \( \frac{1}{T} \sum_{t=1}^{T} \tilde{f}_t \tilde{f}'_t = I_K \),

\[
(\tilde{\Sigma}_K)^{-1} = (\tilde{\Sigma}_{u,K})^{-1} - (\tilde{\Sigma}_{u,K})^{-1} \hat{\Lambda}_K [I_K + \hat{\Lambda}'_K (\tilde{\Sigma}_{u,K})^{-1} \hat{\Lambda}_K]^{-1} \hat{\Lambda}'_K (\tilde{\Sigma}_{u,K})^{-1},
\]

In practice, the true number of factors \( K \) might be unknown to us. However, for any determined \( K_1 \leq p \), we can always construct either \( (\tilde{\Sigma}_{K_1}, \hat{R}_{K_1}^T) \) as in (2.2.7) or \( (\tilde{\Sigma}_{K_1}, \tilde{\Sigma}_{u,K_1}^T) \) as in (2.2.12) to estimate \( (\Sigma, \Sigma_u) \). The following theorem shows that for each given \( K_1 \), the two estimators based on either regularized PCA or least squares substitution are equivalent. Similar results were obtained by Bai (2003) when \( K_1 = K \) and no thresholding was imposed.

**Theorem 2.2.1.** Suppose that the entry-dependent threshold in (2.2.5) is the same as the thresholding parameter used in (2.2.11). Then for any \( K_1 \leq p \), the estimator (2.2.7) is equivalent to the substitution estimator (2.2.12), that is,

\[
\tilde{\Sigma}_{K_1} = \tilde{\Sigma}_{K_1}, \quad \text{and} \quad \tilde{\Sigma}_{u,K_1} = \hat{R}_{K_1}^T.
\]

In this text, we will use a data-driven \( \hat{K} \) to construct the POET (see Section 2.4 below), which has two equivalent representations according to Theorem 2.2.1.
2.2.4 POET with Unknown $K$

In many applications of approximate factor models, the number of true factors $K$ is usually unknown to us. Thus, determining the number of factors in a data-driven way has been an important research topic in the econometric literature. Bai and Ng (2002) proposed a consistent estimator as both $p$ and $T$ diverge. Other recent criteria are proposed by Kapetanios (2010), Onatski (2010), Alessi et al. (2010), etc.

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

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

(2.2.14)

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

**IC1**: $g(T, p) = \frac{p + T}{pT} \log \left( \frac{pT}{p + T} \right)$,

**IC2**: $g(T, p) = \frac{p + T}{pT} \log \min\{p, T\}$.

based on empirical experience because they tend to produce stable results. Recently Alessi et al. (2010) introduced a tuning multiplicative constant to the penalty function to improve the finite sample performance of the selection procedure.

Throughout the text, we let $\hat{K}$ be the solution to (2.2.14) using either IC1 or IC2. The asymptotic results are not affected regardless of the specific choice of $g(T, p)$.
define the POET estimator with unknown $K$ as

$$\hat{\Sigma}_{\hat{K}} = \sum_{i=1}^{\hat{K}} \hat{\lambda}_i \hat{\xi}_i \hat{\xi}_i' + \hat{R}_{\hat{K}}^T.$$  \hfill (2.2.15)

The procedure is as stated in Section 2.2.2 except that $\hat{K}$ is now data-driven.
2.3 Asymptotic Properties

2.3.1 Assumptions

This section presents the assumptions on the model (2.1.2), in which only \( \{y_t\}_{t=1}^T \) are observable. Recall the identifiability condition (2.2.1).

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

**Assumption 2.3.1.** All the eigenvalues of the \( K \times K \) matrix \( p^{-1}B'B \) are bounded away from both zero and infinity as \( p \to \infty \).

**Remark 2.3.1.**

1. It implies from Proposition 2.1 in Section 2 that the first \( K \) eigenvalues of \( \Sigma \) grow at rate \( O(p) \). This unique feature distinguishes our work from most of other low-rank plus sparse covariances considered in the literature, e.g., Luo (2011), Pati et al. (2012), Agarwal et al. (2012), Birnbaum et al. (2012). \(^3\)

2. This assumption is equivalent to Assumption 1.3.5 in Chapter 1, namely:

\[ \|p^{-1}B'B - \Omega\| = o(1) \]

for some \( K \times K \) symmetric positive definite matrix \( \Omega \) such that \( \lambda_{\min}(\Omega) \) and \( \lambda_{\max}(\Omega) \) are bounded away from both zero and infinity.

3. Assumption 3.1 requires the factors to be pervasive, that is, to impact a non-vanishing proportion of individual time series. To illustrate its meaning, consider a one-factor model \((K = 1)\) where \( B = (b_1, \ldots, b_p)' \). Assumption 2.3.1 then reduces to that \( p^{-1} \sum_{i=1}^p b_i^2 \) is bounded away from both zero and infinity. Assuming \( \max_{i \leq p} |b_i| \) is bounded, then \( \sum_{i=1}^p b_i^2 = O(p) \) is automatically satisfied.

\(^3\)To our best knowledge, the only other papers that estimate large covariances with diverging eigenvalues (growing at the rate of dimensionality \( O(p) \)) are Fan et al. (2008, 2011) and Bai and Shi (2011). While Fan et al. (2008, 2011) assumed the factors are observable, Bai and Shi (2011) considered the strict factor model in which \( \Sigma_u \) is diagonal.
On the other hand, the assumption that $p^{-1} \sum_{i=1}^{p} b_i^2$ is bounded away from zero is not stringent since $\sum_{i=1}^{p} b_i^2 = o(p)$ implies many trivial factor loadings. In that case, the factor cannot effectively interpret the variations of $\{y_{it}\}$ across $i$. It is important to distinguish the model we consider in this thesis from the “sparse factor model” in the literature, e.g., Carvalho et al. (2009), Pati et al. (2012), which assumes that the loading matrix $B$ is sparse. The intuition of a sparse loading matrix is that each factor is related to only a relatively small number of stocks, assets, genes, etc. With $B$ being sparse, all the eigenvalues of $B'B$ and hence those of $\Sigma$ are bounded.

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

**Assumption 2.3.2.** (i) $\{u_t, f_t\}_{t \geq 1}$ is strictly stationary. In addition, $E u_{it} = E u_{it} f_{jt} = 0$ for all $i \leq p, j \leq K$ and $t \leq T$.

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

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

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

Condition (i) requires strict stationarity as well as the non-correlation between $\{u_t\}$ and $\{f_t\}$. These conditions are slightly stronger than those in the literature, e.g., Bai (2003), but are still standard and simplify our technicalities. Condition (ii) requires that $\Sigma_u$ be well-conditioned. The condition $\|\Sigma_u\|_1 \leq c_2$ instead of a weaker
condition $\lambda_{\text{max}}(\Sigma_u) \leq c_2$ is imposed here in order to consistently estimate $K$. But it is still standard in the approximate factor model literature as in Bai and Ng (2002), Bai (2003), etc. When $K$ is known, such a condition can be removed. Our working paper\footnote{See Fan, Liao and Mincheva (2011), working paper, arxiv.org/pdf/1201.0175.pdf} shows that the results continue to hold for a growing (known) $K$ under the weaker condition $\lambda_{\text{max}}(\Sigma_u) \leq c_2$. Condition (iii) requires exponential-type tails, which allows us to apply the large deviation theory to $\frac{1}{T} \sum_{t=1}^{T} u_{it} u_{jt} - \sigma_{u,ij}$ and $\frac{1}{T} \sum_{t=1}^{T} f_{jt} u_{it}$.

We impose the strong mixing condition. Let $\mathcal{F}_{-\infty}^0$ and $\mathcal{F}_{T}^\infty$ denote the $\sigma$-algebras generated by $\{(f_t, u_t) : t \leq 0\}$ and $\{(f_t, u_t) : t \geq T\}$ respectively. In addition, define the mixing coefficient

$$\alpha(T) = \sup_{A \in \mathcal{F}_{-\infty}^0, B \in \mathcal{F}_{T}^\infty} |P(A)P(B) - P(AB)|. \quad (2.3.1)$$

**Assumption 2.3.3.** Strong mixing: There exists $r_3 > 0$ such that $3r_1^{-1} + 1.5r_2^{-1} + r_3^{-1} > 1$, and $C > 0$ satisfying: for all $T \in \mathbb{Z}^+$,

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

In addition, we impose the following regularity conditions.

**Assumption 2.3.4.** There exists $M > 0$ such that for all $i \leq p$, $t \leq T$ and $s \leq T$,

(i) $\|b_i\|_{\text{max}} < M$,

(ii) $E[p^{-1/2}(u'_s u_t - EU'_s u_t)]^4 < M$,

(iii) $E[p^{-1/2} \sum_{i=1}^{p} b_i u_{it}]^4 < M$.

These conditions are needed to consistently estimate the transformed common factors as well as the factor loadings. Similar conditions were also assumed in Bai (2003), and Bai and Ng (2006). The number of factors is assumed to be fixed. Our
conditions in Assumption 2.3.4 are weaker than those in Bai (2003) as we focus on different aspects of the study.

The condition $\|\Sigma_u\|_1 \leq c_2$ instead of a weaker condition $\lambda_{\text{max}}(\Sigma_u)$ is imposed here in order to consistently estimate $K$. When $K$ is known, such a condition can be removed. In fact, the results continue to hold for growing $K$ under these weaker conditions.

### 2.3.2 Convergence of the idiosyncratic covariance

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

Suppose $\Sigma_u$ is sparse in the sense of Bickel and Levina (2008), that is, for some $q \in [0, 1),$

$$m_p = \max_{i \leq p} \sum_{j=1}^p |\sigma_{u,ij}|^q$$

is either bounded or grows slowly with $T$ and $p$. In particular, when $q = 0$, many components of $\Sigma_u$ are exactly zero, which corresponds to the “strictly sparse” case.

We estimate $\Sigma_u$ by thresholding the principal orthogonal complements after the first $\hat{K}$ principal components of the sample covariance are taken out: $\hat{\Sigma}_{u,\hat{K}}^T = \hat{R}_{\hat{K}}^T$. By Theorem 2.2.1, it also has an equivalent expression given by (2.2.11), with $\hat{u}_{it} = y_{it} - (\hat{b}_i^{\hat{K}})^T \hat{f}_t$. Here $\hat{K}$ is the data-driven number of factors when the true $K$ is unknown. Throughout the text, we apply the adaptive threshold

$$\tau_{ij} = C \sqrt{\hat{\theta}_{ij}} \omega_T, \quad \omega_T = \frac{1}{\sqrt{p}} + \sqrt{\frac{\log p}{T}}$$  \hspace{1cm} (2.3.2)
where \( C > 0 \) is a sufficiently large constant, though the results hold for other types of thresholding. As in Bickel and Levina (2008) and Cai and Liu (2011), the threshold chosen in the current project is in fact obtained from the optimal uniform rate of convergence of \( \max_{i \leq p, j \leq p} |\tilde{\sigma}_{ij} - \sigma_{u,ij}| \). When direct observation of \( u_{it} \) is not available, the effect of estimating the unknown factors also contributes to this uniform estimation error, which is why \( p^{-1/2} \) appears in the threshold.

The following theorem gives the rate of convergence of the estimated idiosyncratic covariance. Let \( \gamma^{-1} = 3r_1^{-1} + 1.5r_2^{-1} + r_3^{-1} + 1 \). In the convergence rate below, recall that \( m_p \) and \( q \) are defined in the measure of sparsity (2.2.2).

**Theorem 2.3.1.** Suppose \( \log p = o(T^{\gamma/6}) \), \( T = o(p^2) \), and Assumptions 2.3.1-2.3.4 hold. Then for a sufficiently large constant \( C > 0 \) in the threshold (2.3.2), the POET estimator \( \hat{\Sigma}_{u,K}^T \) satisfies

\[
\| \hat{\Sigma}_{u,K}^T - \Sigma_u \| = O_p \left( \omega_1^{-q} m_p \right).
\]

If further \( \omega_1^{-q} m_p = o(1) \), then the eigenvalues of \( \hat{\Sigma}_{u,K}^T \) are all bounded away from zero with probability approaching one, and

\[
\| (\hat{\Sigma}_{u,K}^T)^{-1} - \Sigma_u^{-1} \| = O_p \left( \omega_1^{-q} m_p \right).
\]

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

**Remark 2.3.2.** 1. Consistent estimation of \( \Sigma_u \) indicates that \( \Sigma_u \) is identifiable in (2.1.3), namely, the sparse \( \Sigma_u \) can be separated perfectly from the low-rank matrix there. The result here gives another proof (when assuming \( \omega_1^{-q} m_p = \)
of the “surprising phenomenon” in Candès et al (2011) under different technical conditions.

2. Fan, Liao and Mincheva (2011) recently showed that when $\{f_t\}_{t=1}^T$ are observable and $q = 0$, the rate of convergence of the adaptive thresholding estimator is given by $
abla(\hat{\Sigma}_u - \Sigma_u) = O_p\left(m_p \sqrt{\log p}\right) = \nabla((\hat{\Sigma}_u)^{-1} - \Sigma_u^{-1})$. Hence when the common factors are unobservable, the rate of convergence has an additional term $m_p/\sqrt{p}$, coming from the impact of estimating the unknown factors. This impact vanishes when $p \log p \gg T$, in which case the minimax rate as in Cai and Zhou (2010) is achieved. As $p$ increases, more information about the common factors is collected, which results in more accurate estimation of the common factors $\{f_t\}_{t=1}^T$.

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

2.3.3 Convergence of the POET estimator

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

Example 2.3.1. Consider an ideal case where we know that $K = 1$, $b_i = 1$ for each $i = 1, \ldots, p$, $\Sigma_u = I_p$, and $\{f_{1t}\}_{t=1}^T$ are observable. Then this one-factor model has $\Sigma = \text{var}(f_{1t})1_p1_p^T + I_p$, where $1_p$ denotes the $p$-dimensional column vector of ones. When estimating $\Sigma$, we only need to estimate $\text{var}(f_{1t})$ by its sample covariance and
obtain a substitution estimator. Simple calculations yield to

\[
\|\hat{\Sigma} - \Sigma\| = p\frac{1}{T} \sum_{t=1}^{T} (f_{tt} - \bar{f}_1)^2 - \text{var}(f_{tt}),
\]

Therefore, \(\sqrt{T}\|\hat{\Sigma} - \Sigma\|/p \rightarrow d\) \(|N(0, s_f)|\) for some \(s_f > 0\), and \(\|\hat{\Sigma} - \Sigma\| = o_p(1)\) only if \(p = o(\sqrt{T})\). □

More generally, when a covariance matrix has fast diverging eigenvalues, it is in general hard to be consistently estimated under either Frobenius norm or spectral norm.

**Example 2.3.2.** Consider an ideal case where we know the spectrum except for the first eigenvector of \(\Sigma\). Let \(\{\lambda_j, \xi_j\}_{j=1}^p\) be the eigenvalues and vectors, and assume that the largest eigenvalue \(\lambda_1 \geq c p\) for some \(c > 0\). Let \(\hat{\xi}_1\) be the estimated first eigenvector and define the covariance estimator \(\hat{\Sigma} = \lambda_1 \hat{\xi}_1 \hat{\xi}_1' + \sum_{j=2}^p \lambda_j \xi_j \xi_j'\). Assume that \(\hat{\xi}_1\) is a good estimator in the sense that \(\|\hat{\xi}_1 - \xi_1\|^2 = O_p(T^{-1})\). However,

\[
\|\hat{\Sigma} - \Sigma\| = \lambda_1 \|\hat{\xi}_1 \hat{\xi}_1' - \xi_1 \xi_1'\| = \lambda_1 O_p(\|\hat{\xi} - \xi\|) = O_p(\lambda_1 T^{-1/2}),
\]

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

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

\[
\Sigma^{-1/2} \hat{\Sigma} \Sigma^{-1/2} - I_p,
\]

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

\[
p^{-1/2}\|\Sigma^{-1/2} \hat{\Sigma} \Sigma^{-1/2} - I_p\|_F = \left(p^{-1} \text{tr}[((\Sigma^{-1/2} \hat{\Sigma} \Sigma^{-1/2} - I_p)^2)]\right)^{1/2}.
\]
In the last equality, there are $p$ terms being added in the trace operation and the factor $p^{-1}$ plays the role of normalization. The loss (2.3.3) is closely related to the entropy loss, introduced by James and Stein (1961). Also note that

$$p^{-1/2}||\Sigma^{-1/2}\tilde{\Sigma}\Sigma^{-1/2} - I_p||_F = ||\tilde{\Sigma} - \Sigma||_\Sigma$$

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

To avoid accumulating the estimation errors, we normalize it by $p^{-1}$, which gives an averaged estimation error. In particular, $||\Sigma||_\Sigma = 1$. It is easy to apply the properties of the Frobenius norm to check that the triangular inequality holds for $||.||_\Sigma$. Due to the invertibility of $\Sigma$, $||A||_\Sigma = 0$ if and only if $A = 0$. Hence it is indeed a norm. In addition, for any two $p \times p$ matrices $A_1$ and $A_2$:

$$||A_1 - A_2||_\Sigma = p^{-1/2}||\Sigma^{-1/2}(A_1 - A_2)\Sigma^{-1/2}||_F.$$

$$\leq ||\Sigma^{-1/2}(A_1 - A_2)\Sigma^{-1/2}||_F.$$

$$\leq ||A_1 - A_2|| \cdot \lambda_{\text{max}}(\Sigma^{-1}).$$

Fan et al. (2008) showed that in a large factor model, the sample covariance is such that $||\hat{\Sigma}_{\text{sam}} - \Sigma||_\Sigma = O_p(\sqrt{p/T})$, which does not converge if $p > T$. On the other hand, Theorem 2.3.2 below shows that $||\hat{\Sigma}_K - \Sigma||_\Sigma$ can still be convergent as long as $p = o(T^2)$. Technically, the impact of high-dimensionality on the convergence rate of $\hat{\Sigma}_K - \Sigma$ is via the number of rows in $B$. We show in the appendix that $B$ appears in $||\hat{\Sigma}_K - \Sigma||_\Sigma$ through $B'\Sigma^{-1}B$ whose eigenvalues are bounded. Therefore it successfully cancels out the curse of high-dimensionality introduced by $B$.

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

**Theorem 2.3.2.** Under the assumptions of Theorem 2.3.1, the POET estimator defined in (2.2.15) satisfies

\[
\| \hat{\Sigma} - \Sigma \| = O_p \left( \sqrt{\log \frac{p}{T}} + m_p \omega_T^{1-q} \right), \quad \| \hat{\Sigma} - \Sigma \|_{\text{max}} = O_p(\omega_T).
\]

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

\[
\| \hat{\Sigma}^{-1} - \Sigma^{-1} \| = O_p \left( m_p \omega_T^{1-q} \right).
\]

**Remark 2.3.3.**

1. When estimating \( \Sigma^{-1} \), \( p \) is allowed to grow exponentially fast in \( T \), and the estimator has the same rate of convergence as that of the estimator \( \hat{\Sigma}^T_{u\hat{K}} \) in Theorem 2.3.1. When \( p \) becomes much larger than \( T \), the precision matrix can be estimated at the same rate as if the factors were observable.

2. As in Remark 2.3.2, when \( K > 0 \) is known and grows with \( p \) and \( T \), the working paper Fan et al. (2011) proves the following results (when \( q = 0 \))

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

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

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

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

3. The relative error \( \| \Sigma^{-1/2} \hat{\Sigma} \Sigma^{-1/2} - I_p \| \) in operator norm can be shown to have the same order as the maximum relative error of estimated eigenvalues. It does

\[\text{The assumptions in the working paper Fan et al. (2011) are slightly weaker than those presented here, in that it required } \lambda_{\max}(\Sigma_u) \text{ instead of } \| \Sigma_u \|_1 \text{ to be bounded.}\]
not converge to zero nor diverge. It is much smaller than $\|\hat{\Sigma}_K - \Sigma\|$, which is of order $p/\sqrt{T}$ (see Example 2.3.2).

2.3.4 Convergence of unknown factors and factor loadings

Many applications of the factor model require estimating the unknown factors. In general, factor loadings in $B$ and the common factors $f_t$ are not separably identifiable, as for any matrix $H$ such that $H'H = I_K$, $Bf_t = BH'Hf_t$. Hence $(B, f_t)$ cannot be identified from $(BH', Hf_t)$. However, this ambiguity is eliminated by the identifiability condition (2.2.1), subject to a permutation. Note that the linear space spanned by the rows of $B$ is the same as that by those of $BH'$. In practice, it often does not matter which one is used.

Let $V$ denote the $\hat{K} \times \hat{K}$ diagonal matrix of the first $\hat{K}$ largest eigenvalues of the sample covariance matrix in decreasing order. Recall that $F' = (f_1, ..., f_T)$ and define a $\hat{K} \times \hat{K}$ matrix $H = \frac{1}{T} V^{-1}\hat{F}'FB'B$. Then for $t \leq T$, $Hf_t = T^{-1} V^{-1}\hat{F}'(Bf_1, ..., Bf_T)'Bf_t$. Note that $Hf_t$ depends only on the data $V^{-1}\hat{F}'$ and an identifiable part of parameters $\{Bf_t\}_{t=1}^T$. Therefore, there is no identifiability issue in $Hf_t$ regardless of the imposed identifiability condition.

Bai (2003) obtained the rate of convergence for both $\hat{b}_i$ and $\hat{f}_t$ for any fixed $(i, t)$. However, the uniform rate of convergence is more relevant for many applications (see Example 5.1). The following theorem extends those results in Bai (2003) in a uniformity sense. In particular, with a more refined technique, we have improved the uniform convergence rate for $\hat{f}_t$.

**Theorem 2.3.3.** Under the assumptions of Theorem 2.3.1,

$$\max_{i \leq p} \|\hat{b}_i - Hb_i\| = O_p(\omega_T), \quad \max_{t \leq T} \|\hat{f}_t - Hf_t\| = O_p\left(\frac{1}{T^{1/2}} + \frac{T^{1/4}}{\sqrt{p}}\right).$$
We show in the appendix that \( \mathbf{H}' \mathbf{H} = \mathbf{I}_K + o_p(1) \). Given the selection consistency that \( P(\hat{K} = K) \to 1 \), \( \mathbf{H} \) is invertible and therefore \( \mathbf{H}' = \mathbf{H}^{-1} \) with probability approaching one.

As a consequence of Theorem 2.3.3, we obtain the following: (recall that the constant \( r_2 \) is defined in Assumption 2.3.2.)

**Corollary 2.3.1.** Under the assumptions of Theorem 2.3.1,

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

The rates of convergence obtained above also explain the condition \( T = o(p^2) \) in Theorems 2.3.1 and 2.3.2. It is needed in order to estimate the common factors \( \{f_t\}_{t=1}^T \) uniformly in \( t \leq T \). When we do not observe \( \{f_t\}_{t=1}^T \), in addition to the factor loadings, there are \( KT \) factors to estimate. Intuitively, the condition \( T = o(p^2) \) requires the number of parameters introduced by the unknown factors be “not too many”, so that we can consistently estimate them uniformly. Technically, as demonstrated by Bickel and Levina (2008), Cai and Liu (2011) and many other authors, achieving uniform accuracy is essential for large covariance estimations.

2.4 Choice of Threshold

### 2.4.1 Finite-sample positive definiteness

Theorems 2.3.1 and 2.3.2 guarantee the positive definiteness asymptotically as long as \( \omega_T^{1-q} m_p = o(1) \) and the threshold is set to \( \tau_{ij} = C \sqrt{\theta_{ij} \omega_T} \) for some large constant \( C > 0 \).

Recall that the threshold value \( \tau_{ij} = C \sqrt{\theta_{ij} \omega_T} \), where \( C \) is determined by the users. To make POET operational in practice, one has to choose \( C \) to maintain the positive definiteness of the estimated covariances for any given finite sample. We write
\[
\hat{\Sigma}_{u, \hat{K}}^T(C) = \hat{\Sigma}_{u, \hat{K}}^T, \text{ where the covariance estimator depends on } C \text{ via the threshold.}
\]

We choose \( C \) in the range where \( \lambda_{\min}(\hat{\Sigma}_{u, \hat{K}}^T) > 0 \). Define

\[
C_{\min} = \inf\{C > 0 : \lambda_{\min}(\hat{\Sigma}_{u, \hat{K}}^T(M)) > 0, \forall M > C\} \tag{2.4.1}
\]

When \( C \) is sufficiently large, the estimator becomes diagonal, while its minimum eigenvalue must retain strictly positive. Thus, \( C_{\min} \) is well defined and for all \( C > C_{\min}, \hat{\Sigma}_{u, \hat{K}}^T(C) \) is positive definite under finite sample. We can obtain \( C_{\min} \) by solving \( \lambda_{\min}(\hat{\Sigma}_{u, \hat{K}}^T(C)) = 0, C \neq 0 \). We can also approximate \( C_{\min} \) by plotting \( \lambda_{\min}(\hat{\Sigma}_{u, \hat{K}}^T(C)) \) as a function of \( C \), as illustrated in Figure 2.1. In practice, we can choose \( C \) in the range \( (C_{\min} + \epsilon, M) \) for a small \( \epsilon \) and large enough \( M \). Choosing the threshold in a range to guarantee the finite-sample positive definiteness has also been previously suggested by Fryzlewicz (2010).

Figure 2.1: Minimum eigenvalue of \( \hat{\Sigma}_{u, \hat{K}}^T(C) \) as a function of \( C \) for three choices of thresholding rules. The plot is based on the simulated data set in Section 6.2.
2.4.2 Multifold Cross-Validation

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

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

$$C^* = \arg \min_{C_{\min} + \epsilon \leq C \leq M} \frac{1}{H} \sum_{j=1}^H \| \hat{\Sigma}_u^T(C) - \hat{\Sigma}_u^j \|^2_F. \quad (2.4.2)$$

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

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

### 2.5 Applications of POET

We give four examples to which the results in Theorems 2.3.1–2.3.3 can be applied. Detailed pursuits of these are beyond the scope of this thesis.

**Example 2.5.1** (Large-scale hypothesis testing). Controlling the false discovery rate in large-scale hypothesis testing based on correlated test statistics is an important and challenging problem in statistics (Leek and Storey, 2008; Efron, 2010; Fan, et al., 2012). Suppose that the test statistic for each of the hypothesis

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

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

If the covariance \( \Sigma \) admits the approximate factor structure (2.1.3), then the test statistics can be stochastically decomposed as

\[ Z = \mu + Bf + u, \quad \text{where } \Sigma_u \text{ is sparse.} \tag{2.5.1} \]

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

\[ V(x) = \sum_{i=1}^{p} \{ \Phi(a_i(z_{x/2} + \eta_i)) + \Phi(a_i(z_{x/2} - \eta_i)) \} + o_P(p), \tag{2.5.2} \]
when \( m_p = o(p) \) and the number of true significant hypothesis \( \{ i : \mu_i \neq 0 \} \) is \( o(p) \),
where \( z_x \) is the upper \( x \)-quantile of the standard normal distribution, \( \eta_i = (Bf)_i \) and 
\( a_i = \text{var}(u_i)^{-1} \).

Now suppose that we have \( n \) repeated measurements from the model (2.5.1). Then, by Corollary 2.3.1, \( \{ \eta_i \} \) can be uniformly consistently estimated, and hence \( p^{-1}V(x) \) and \( \text{FDP}(x) \) can be consistently estimated. Efron (2010) obtained these repeated test statistics based on the bootstrap sample from the original raw data. Our theory (Theorem 2.3.3) gives a formal justification to the framework of Efron (2007, 2010). Detailed developments using POET are described in a forthcoming project by Fan, Gu and Han, in which Theorem 3.3 plays a pivotal role.

**Example 2.5.2** (Risk management). The maximum elementwise estimation error \( \| \hat{\Sigma}_K - \Sigma \|_{\text{max}} \) appears in risk assessment as in Fan, Zhang and Yu (2012). For a fixed portfolio allocation vector \( w \), the true portfolio variance and the estimated one are given by \( w'\Sigma w \) and \( w'\hat{\Sigma}_K w \) respectively. The estimation error is bounded by

\[
|w'\hat{\Sigma}_K w - w'\Sigma w| \leq \|\hat{\Sigma}_K - \Sigma\|_{\text{max}}\|w\|^2_1,
\]

where \( \|w\|_1 \), the \( L_1 \)-norm of \( w \), 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 \( \|w\|_1 \leq c \) for some \( c > 0 \). In particular, \( c = 1 \) corresponds to a portfolio with no-short positions (all weights are nonnegative). Theorem 2.3.2 quantifies the maximum approximation error.

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

\[
|w'\hat{\Sigma}_K w/w'\Sigma w - 1| \leq \|\Sigma^{-1/2}\hat{\Sigma}_K \Sigma^{-1/2} - I_p\|
\]

for any allocation vector \( w \). Theorem 2.3.2 quantifies this relative error.
Example 2.5.3 (Panel regression with a factor structure in the errors). Consider the following panel regression model

\[ Y_{it} = x_{it}'\beta + \varepsilon_{it}, \quad \varepsilon_{it} = b_i'f_t + u_{it}, \quad i \leq p, t \leq T, \]

where \( x_{it} \) is a vector of observable regressors with fixed dimension. The regression error \( \varepsilon_{it} \) has a factor structure and is assumed to be independent of \( x_{it} \), but \( b_i, f_t \) and \( u_{it} \) are all unobservable. We are interested in the common regression coefficients \( \beta \).

The above panel regression model has been considered by many researchers, such as Ahn, Lee and Schmidt (2001), Pesaran (2006), and has broad applications in social sciences. For example, in the income studies, \( Y_{it} \) represents the income of individual \( i \) at age \( t \), \( x_{it} \) is a vector of observable characteristics that are associated with income. Here \( b_i \) represents a vector of unmeasured skills, such as innate ability, motivation, and hardworking; \( f_t \) 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 \( \beta \), a more efficient estimation can be obtained by GLS (generalized least squares). The GLS method depends, however, on an estimator of \( \Sigma^{-1}_\varepsilon \), the inverse of the covariance matrix of \( \varepsilon_t = (\varepsilon_{1t}, ..., \varepsilon_{pt})' \). By assuming the covariance matrix of \((u_{1t}, ..., u_{pt})\) to be sparse, we can successfully solve this problem by applying Theorem 2.3.2. Although \( \varepsilon_{it} \) is unobservable, it can be replaced by the regression residuals \( \hat{\varepsilon}_{it} \), obtained via first regressing \( Y_{it} \) on \( x_{it} \). We then apply the POET estimator to \( T^{-1} \sum_{t=1}^{T} \hat{\varepsilon}_t\hat{\varepsilon}_t' \). By Theorem 2.3.2, the inverse of the resulting estimator is a consistent estimator of \( \Sigma^{-1}_\varepsilon \) under the spectral norm. A slight difference lies in the fact that when we apply POET, \( T^{-1} \sum_{t=1}^{T} \varepsilon_t\varepsilon_t' \) is replaced with \( T^{-1} \sum_{t=1}^{T} \hat{\varepsilon}_t\hat{\varepsilon}_t' \), which introduces an additional term \( O_p(\sqrt{\frac{\log p}{T}}) \) in the estimation error.
Example 2.5.4 (Validating an asset pricing theory). A celebrated financial economic theory is the capital asset pricing model (CAPM, Sharpe 1964) that makes William Sharpe win the Nobel prize in Economics in 1990, whose extension is the multi-factor model (Ross, 1976, Chamberlain and Rothschild, 1983). It states that in a frictionless market, the excessive return of any financial asset equals the excessive returns of the risk factors times its factor loadings plus noises. In the multi-period model, the excess return $y_{it}$ of firm $i$ at time $t$ follows model (2.1.1), in which $f_t$ is the excess returns of the risk factors at time $t$. To test the null hypothesis (2.1.2), one embeds the model into the multivariate linear model

$$y_t = \alpha + Bf_t + u_t, \quad t = 1, \ldots, T$$

(2.5.3)

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

To provide some insights, let $\hat{\alpha}$ be the least-squares estimator of (2.5.3). Then, when $u_t \sim N(0, \Sigma_u)$, $\hat{\alpha} \sim N(\alpha, \Sigma_u/c_T)$ for a constant $c_T$ which depends on the observed factors. When $\Sigma_u$ is known, the Wald test statistic is $W = c_T \hat{\alpha}' \Sigma_u^{-1} \hat{\alpha}$. When it is unknown and $p$ is large, it is natural to use the F-type of test statistic $\hat{W} = c_T \hat{\alpha}' (\hat{\Sigma}_{u,K}^{T})^{-1} \hat{\alpha}$. The difference between these two statistics is bounded by

$$|\hat{W} - W| \leq c_T \| (\hat{\Sigma}_{u,K}^{T})^{-1} - \Sigma_u^{-1} \| \| \hat{\alpha} \|^2.$$
Since under the null hypothesis $\hat{\alpha} \sim N(0, \Sigma_u/c_T)$, we have $c_T\| \Sigma_u^{-1/2}\hat{\alpha}\|^2 = O(p)$. Thus, it follows from boundness of $\|\Sigma_u\|$ that $|\hat{W} - W| = O(p)\|\tilde{\Sigma}^{-1}_{u, \tilde{K}} - \Sigma_u^{-1}\|$. Theorem 3.1 provides the rate of convergence for the above difference. Detailed development is out of the scope of the current thesis, and we will leave it as a separate research project.
Chapter 3

Numerical Studies and Applications to Finance

3.1 Monte Carlo Experiments with Observable Factors

In this section, we use simulation to demonstrate the rates of convergence of the estimators $\hat{\Sigma}^T$ and $(\hat{\Sigma}^T)^{-1}$ that we have obtained so far. 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.

3.1.1 Calibration

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

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

Table 3.1: Mean and covariance matrix used to generate $\mathbf{b}$

<table>
<thead>
<tr>
<th>$\mu_B$</th>
<th>$\Sigma_B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0641</td>
<td>0.0475</td>
</tr>
<tr>
<td>0.1233</td>
<td>0.0218</td>
</tr>
<tr>
<td>-0.0119</td>
<td>0.0488</td>
</tr>
</tbody>
</table>

2. For each fixed $p$, create the sparse matrix $\Sigma_u = \mathbf{D} + \mathbf{ss}' - \text{diag}(s_1^2, ..., s_p^2)$ in the following way. Let $\hat{\mathbf{u}}_t = \tilde{y}_t - \tilde{\mathbf{B}}\tilde{f}_t$. For $i = 1, ..., 30$, let $\hat{\sigma}_i$ denote the standard deviation of the residuals of the $i$th 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 \( \hat{\sigma}_i \)'s, namely \( \bar{\sigma} = 0.6055 \) and \( \sigma_{SD} = 0.2621 \).

Let \( D = \text{diag}\{\sigma_1^2, ..., \sigma_p^2\} \), 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 \( \sigma_i \) if it is between \( \min(\hat{\sigma}_i) = 0.3533 \) and \( \max(\hat{\sigma}_i) = 1.5222 \).

Create \( s = (s_1, ..., s_p)' \) to be a sparse vector. We set each \( s_i \sim N(0, 1) \) with probability \( \frac{0.2}{\sqrt{2\log p}} \), and \( s_i = 0 \) otherwise. This leads to an average of \( \frac{0.2\sqrt{\beta}}{\log p} \) nonzero elements per each row of the error covariance matrix.

Create a loop that generates \( \Sigma_u \) multiple times until it is positive definite.

3. Assume the factors follow the vector autoregressive model (VAR(1)) model \( f_t = \mu + \Phi f_{t-1} + \varepsilon_t \) for some \( 3 \times 3 \) matrix \( \Phi \), where \( \varepsilon_t \)'s are i.i.d. \( N_3(0, \Sigma_\varepsilon) \). We estimate \( \Phi, \mu \) and \( \Sigma_\varepsilon \) from the data, and obtain \( \text{cov}(f_t) \).

<table>
<thead>
<tr>
<th>( \mu )</th>
<th>\text{cov}(f_t)</th>
<th>\Phi</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1074</td>
<td>2.2540 0.2735 0.9197</td>
<td>-0.1149 0.0024 0.0776</td>
</tr>
<tr>
<td>0.0357</td>
<td>0.2735 0.3767 0.0430</td>
<td>0.0016 -0.0162 0.0387</td>
</tr>
<tr>
<td>0.0033</td>
<td>0.9197 0.0430 0.6822</td>
<td>-0.0399 0.0218 0.0351</td>
</tr>
</tbody>
</table>

### 3.1.2 Simulation

For each fixed \( p \), we generate \( \{b_1, ..., b_p\} \) independently from \( N_3(\mu_B, \Sigma_B) \), and generate \( \{f_t\}_{t=1}^T \) and \( \{u_t\}_{t=1}^T \) independently from and from respectively. 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.

Repeat the following steps \( N = 200 \) times for each fixed \( p \):

1. Generate \( \{b_t\}_{t=1}^p \) independently from \( N_3(\mu_B, \Sigma_B) \), and set \( B = (b_1, ..., b_p)' \).
2. Generate \( \{u_t\}_{t=1}^T \) independently from \( N_p(0, \Sigma_u) \).

3. Generate \( \{f_t\}_{t=1}^T \) independently from the VAR(1) model \( f_t = \mu + \Phi f_{t-1} + \varepsilon_t \).

4. Calculate \( y_t = Bf_t + u_t \) for \( t = 1, ..., T \).

5. Set \( \omega_T = 0.10K \sqrt{\log p/T} \) to obtain the thresholding estimator (1.2.5) \( \hat{\Sigma}_u^T \) and the sample covariance matrices \( \hat{\text{cov}}(f_t), \hat{\Sigma}_y = \frac{1}{T-1} \sum_{t=1}^T (y_t - \bar{y})(y_t - \bar{y})^T \).

We graph the convergence of \( \hat{\Sigma}_u^T \) and \( \hat{\Sigma}_y \) to \( \Sigma \), the covariance matrix of \( y \), under the entropy-loss norm \( \| \cdot \|_\Sigma \) and the L-infinity norm \( \| \cdot \|_\infty \). We also graph the convergence of the inverses \( (\hat{\Sigma}_u^T)^{-1} \) and \( \hat{\Sigma}_y^{-1} \) to \( \Sigma^{-1} \) 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, \( \hat{\Sigma} \) is nearly singular, which leads to abnormally large values of the operator norm. Lastly, we record the standard deviations of these norms.

### 3.1.3 Results

In figures 1-3, the dashed curves correspond to \( \hat{\Sigma}_u^T \) and the solid curves- to the sample covariance matrix \( \hat{\Sigma} \). Figure 1 and 2 presents the averages and standard deviations of the estimation error of both of these matrices with respect to the \( \Sigma \)-norm and infinity norm, respectively. Figure 3 presents the averages and estimation errors of the inverses with respect to the operator norm. Based on the simulation results, we can make the following observations:

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

2. Under the \( \| \cdot \|_\Sigma \), our estimate of the covariance matrix of \( y, \hat{\Sigma}_u^T \) performs much better than the sample covariance matrix \( \hat{\Sigma}_y \). Note that, in the proof of Theo-
Figure 3.1: Averages and standard deviations of $\|\hat{\Sigma}^T - \Sigma\|_\Sigma$ (dashed curve) and $\|\hat{\Sigma}_y - \Sigma\|_\Sigma$ (solid curve) over $N = 200$ iterations, as a function of the dimensionality $p$. 
Figure 3.2: Averages and standard deviations of $\|\tilde{\Sigma}^T - \Sigma\|_\infty$ (dashed curve) and $\|\tilde{\Sigma}_y - \Sigma\|_\infty$ (solid curve) over $N = 200$ iterations, as a function of the dimensionality $p$. 
Figure 3.3: Averages and standard deviations of $\|\left(\Sigma^T\right)^{-1} - \Sigma^{-1}\|$ (dashed curve) and $\|\hat{\Sigma}_y^{-1} - \Sigma^{-1}\|$ (solid curve) over $N = 200$ iterations, as a function of the dimensionality $p$.

rem 2 in Fan, Fan, Lv(2008), it was shown that:

$$
\|\hat{\Sigma}_y - \Sigma\|_\Sigma^2 = O_p\left(\frac{K^3}{T_p}\right) + O_p\left(\frac{p}{T}\right) + O_p\left(\frac{K^{3/2}}{T}\right). \quad (3.1.1)
$$
For a small fixed value of $K$, such as $K = 3$, the dominating term in (1.3) is $O \left( \frac{p}{T} \right)$. From Theorem 4.1, and given that $m_T = o(p^{1/4})$, the dominating term in the convergence of $\| \hat{\Sigma}_T - \Sigma \|^2_{\Sigma}$ is $O_p \left( \frac{p}{T^2} + \frac{m_T^2 \log p}{T} \right)$. So, we would expect our 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, $(\hat{\Sigma}_T)^{-1}$ also performs significantly better than the inverse of the sample covariance matrix.

5. Finally, when $p > 500$, the thresholding estimators $\hat{\Sigma}_u^T$ and $\hat{\Sigma}^T$ are still nonsingular.

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

### 3.2 Monte Carlo Experiments with Unobservable Factors

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

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

The factor loadings are drawn from a trivariate normal distribution $b \sim N_3(\mu_B, \Sigma_B)$, the idiosyncratic errors from $u_t \sim N_p(0, \Sigma_u)$, and the factor returns $f_t$ follow a VAR(1) model. To make the simulation more realistic, model parameters are calibrated from the financial returns, as detailed in the following section.

### 3.2.1 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 ($\tilde{y}_t$) are computed for the period from January 1st, 2009 to December 31st, 2010. Here, we present a short outline of the calibration procedure.

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

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

3. We run the stationary vector autoregressive model $f_t = \mu + \Phi f_{t-1} + \varepsilon_t$, a VAR(1) model, to the data $\tilde{\mathbf{F}}$ to obtain the multivariate least squares estimator for $\mu$. 

63
Table 3.3: Mean and covariance matrix used to generate $b$

<table>
<thead>
<tr>
<th>$\mu_B$</th>
<th>$\Sigma_B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0047</td>
<td>0.0767 -0.00004 0.0087</td>
</tr>
<tr>
<td>0.0007</td>
<td>-0.00004 0.0841 0.0013</td>
</tr>
<tr>
<td>-1.8078</td>
<td>0.0087 0.0013 0.1649</td>
</tr>
</tbody>
</table>

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

Table 3.4: Parameters of $f_t$ generating process

<table>
<thead>
<tr>
<th>$\mu$</th>
<th>$\text{cov}(f_t)$</th>
<th>$\Phi$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.0050</td>
<td>1.0037 0.0011 -0.0009</td>
<td>-0.0712 0.0468 0.1413</td>
</tr>
<tr>
<td>0.0335</td>
<td>0.0011 0.9999 0.0042</td>
<td>-0.0764 -0.0008 0.0646</td>
</tr>
<tr>
<td>-0.0756</td>
<td>-0.0009 0.0042 0.9973</td>
<td>0.0195 -0.0071 -0.0544</td>
</tr>
</tbody>
</table>

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

$$\Sigma_u = D\Sigma_0 D.$$ 

Here, $\Sigma_0$ is the error correlation matrix, and $D$ is the diagonal matrix of the standard deviations of the errors. We set $D = \text{diag}(\sigma_1, ..., \sigma_p)$, where each $\sigma_i$ is generated independently from a Gamma distribution $G(\alpha, \beta)$, and $\alpha$ and $\beta$ 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 $\Sigma_0$ 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 $\Sigma_0$ sparse, where the threshold
is found as the smallest constant that provides the positive definiteness of $\Sigma_0$.
More precisely, start with threshold value 1, which gives $\Sigma_0 = I_p$ and then
decrease the threshold values in a grid until positive definiteness is violated.

### 3.2.2 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 $\{b_i\}_{i=1}^p \sim N_3(\mu_B, \Sigma_B)$, and set $B = (b_1, ..., b_p)'$.
2. Generate independently $\{u_t\}_{t=1}^T \sim N_p(0, \Sigma_u)$.
3. Generate $\{f_t\}_{t=1}^T$ as a vector autoregressive sequence of the form $f_t = \mu + \Phi f_{t-1} + \epsilon_t$.
4. Calculate $\{y_t\}_{t=1}^T$ from $y_t = Bf_t + u_t$.
5. Set hard-thresholding with threshold $0.5\sqrt{\hat{\theta}_{ij}(\sqrt{\log p}/T + 1/\sqrt{p})}$. Estimate $K$ using
   Bai and Ng (2002)'s IC1. Calculate covariance estimators using the POET
   method. Calculate the sample covariance matrix $\hat{\Sigma}_{sam}$.

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

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

From the simulation results, reported in Figures 3.4-3.7, we observe that POET under the unobservable factor model performs just as well as the estimator in Fan et al. (2011) if the factors are known, when $p$ is large enough. The cost of not knowing the factors is approximately of order $O_p(1/\sqrt{p})$. It can be seen in Figures 3.4 and 3.5 that this cost vanishes for $p \geq 200$. To give a better insight of the impact of estimating the unknown factors for small $p$, a separate set of simulations is conducted for $p \leq 100$. As we can see from Figures 3.4 (bottom panel) and 3.5 (middle and bottom panels), the impact decreases quickly. In addition, when estimating $\Sigma^{-1}$, it is hard to distinguish the estimators with known and unknown factors, whose performances are quite stable compared to the sample covariance matrix. Also, the maximum absolute elementwise error (Figure 3.6) of our estimator performs very similarly to that of the sample covariance matrix, which coincides with our asymptotic result. Figure 3.7 shows that the performances of the three methods are indistinguishable in the spectral norm, as expected.

Next, let us focus on the importance of selecting an appropriate value for the threshold rate $\omega_T$. Choosing $\omega_T$ to be too small creates noninvertibility issues with $\hat{\Sigma}_u^T$ and therefore, with $\hat{\Sigma}^T$. On the other hand, the larger the threshold, the sparser the matrix $\hat{\Sigma}_u^T$, so choosing the threshold large enough will bring us to the diagonal case of Fan, Fan, Lv (2008).

3.2.4 Robustness to the estimation of $K$

The POET estimator depends on the estimated number of factors. Our theory uses a consistent estimator $\hat{K}$. To assess the robustness of our procedure to $\hat{K}$ in finite
Figure 3.4: Averages (left panel) and standard deviations (right panel) of the relative error $p^{-1/2} \| \Sigma^{-1/2} \hat{\Sigma} \Sigma^{-1/2} - I_p \|_F$ with known factors ($\hat{\Sigma} = \hat{\Sigma}_{obs}$ solid red curve), POET ($\hat{\Sigma} = \hat{\Sigma}_{R}$ solid blue curve), and sample covariance ($\hat{\Sigma} = \hat{\Sigma}_{sam}$ dashed curve) over 200 simulations, as a function of the dimensionality $p$. Top panel: $p$ ranges in 20 to 600 with increment 20; bottom panel: $p$ ranges in 1 to 100 with increment 1.
Figure 3.5: Averages (left panel) and standard deviations (right panel) of $\|\hat{\Sigma}^{-1} - \Sigma^{-1}\|$ with known factors ($\hat{\Sigma} = \hat{\Sigma}_{\text{obs}}$ solid red curve), POET ($\hat{\Sigma} = \hat{\Sigma}_{\hat{\Sigma}}$ solid blue curve), and sample covariance ($\hat{\Sigma} = \hat{\Sigma}_{\text{sam}}$ dashed curve) over 200 simulations, as a function of the dimensionality $p$. Top panel: $p$ ranges in 20 to 600 with increment 20; middle panel: $p$ ranges in 1 to 100 with increment 1; Bottom panel: the same as the top panel with dashed curve excluded.
Figure 3.6: Averages (left panel) and standard deviations (right panel) of $\|\hat{\Sigma} - \Sigma\|_{\text{max}}$ with known factors ($\hat{\Sigma} = \hat{\Sigma}_{\text{obs}}$ solid red curve), POET ($\hat{\Sigma} = \hat{\Sigma}_K$ solid blue curve), and sample covariance ($\hat{\Sigma} = \hat{\Sigma}_{\text{sam}}$ dashed curve) over 200 simulations, as a function of the dimensionality $p$. They are nearly indifferenciable.

Figure 3.7: Averages of $\|\hat{\Sigma} - \Sigma\|$ (left panel) and $\|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - I_p\|$ with known factors ($\hat{\Sigma} = \hat{\Sigma}_{\text{obs}}$ solid red curve), POET ($\hat{\Sigma} = \hat{\Sigma}_K$ solid blue curve), and sample covariance ($\hat{\Sigma} = \hat{\Sigma}_{\text{sam}}$ dashed curve) over 200 simulations, as a function of the dimensionality $p$. The three curves are hardly distinguishable on the left panel.
sample, we calculate $\hat{\Sigma}^T_{u,K}$ for $K = 1, 2, \ldots, 10$. Again, the threshold is fixed to be $0.5\sqrt{\theta_{ij}(\sqrt{\frac{\log p}{T}} + \frac{1}{\sqrt{p}})}$.

**Design 1**

The simulation setup is the same as before where the true $K_0 = 3$. We calculate $\|\hat{\Sigma}^T_{u,K} - \Sigma_u\|$, $\|\hat{\Sigma}^T_{u,K} - \Sigma^{-1}_{u}\|$, $\|\hat{\Sigma}^{-1}_{K} - \Sigma^{-1}\|$ and $\|\hat{\Sigma} - \Sigma\|_\Sigma$ for $K = 1, 2, \ldots, 10$. Figure 3.9 plots these norms as $p$ increases but with a fixed $T = 300$. The results demonstrate a trend that is quite robust when $K \geq 3$; especially, the estimation accuracy of the spectral norms for large $p$ are close to each other. When $K = 1$ or 2, the estimators perform badly due to modeling bias. Therefore, POET is robust to over-estimated $K$, but not to under-estimation.

**Design 2**

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

$$
\sigma_{u,ij} = \begin{cases} 
0.5|i-j|, & |i-j| \leq 9 \\
0, & |i-j| > 9
\end{cases}, \quad (u_1, \ldots, u_T) \sim^{i.i.d.} N_p(0, \Sigma_u).
$$

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

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

Table 3.6 summarizes the average estimation error of covariance matrices across $K$ in the spectral norm. Each simulation is replicated 50 times and $T = 200$.

Table 3.6 illustrates some interesting patterns. First of all, the best estimation accuracy is achieved when $K = K_0$. Second, the estimation is robust for $K \geq K_0$. As $K$ increases from $K_0$, the estimation error becomes larger, but is increasing slowly in
Figure 3.8: Robustness of $K$ as $p$ increases for various choices of $K$ (Design 1). Top left: $\|\hat{\Sigma}^T_{u,K} - \Sigma_u\|_F / \sqrt{p}$; top right: $\|\hat{\Sigma}^T_{u,K} - \Sigma_u^{-1}\|_F / \sqrt{p}$; bottom left: $\|\hat{\Sigma}_K - \Sigma\|_F$; bottom right: $\|\hat{\Sigma}_K^{-1} - \Sigma^{-1}\|_F / \sqrt{p}$. $T$ is fixed to be 300.
Figure 3.9: Robustness of $K$ as $p$ increases for various choices of $K$ (Design 1, $T = 300$). Top left: $\|\hat{\Sigma}_{u,K}^T - \Sigma_u\|$; top right: $\|\hat{\Sigma}_{u,K}^{-1} - \Sigma_u^{-1}\|$; bottom left: $\|\hat{\Sigma}_K - \Sigma\|$; bottom right: $\|\hat{\Sigma}_K^{-1} - \Sigma^{-1}\|$.
Table 3.5: Robustness of $K$. Design 2, estimation errors in spectral norm

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

In general, which indicates the robustness when a slightly larger $K$ has been used. Third, when the number of factors is under-estimated, corresponding to $K = 1, 2$, all the estimators perform badly, which demonstrates the danger of missing any common factors. Therefore, over-estimating the number of factors, while still maintaining a satisfactory estimation accuracy of the covariance matrices, is much better than under-estimating. The resulting bias caused by under-estimation is more severe than the additional variance introduced by over-estimation. Finally, estimating $\Sigma$, the covariance of $y_t$, does not achieve a good accuracy even when $K = K_0$ in the absolute term $\|\hat{\Sigma} - \Sigma\|$, but the relative error $\|\Sigma^{-1/2}\Sigma_K\Sigma^{-1/2} - I_p\|$ is much smaller. This is consistent with our discussions in Section 3.3.
<table>
<thead>
<tr>
<th>K</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>8</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>p = 100</td>
<td>(\hat{\Sigma}_{u,K}^T)</td>
<td>2.86</td>
<td>1.76</td>
<td>0.50</td>
<td>0.54</td>
<td>0.62</td>
<td>0.65</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>((\hat{\Sigma}_{u,K}^T))^{-1}</td>
<td>1.43</td>
<td>1.23</td>
<td>0.69</td>
<td>0.71</td>
<td>0.68</td>
<td>0.72</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>(\hat{\Sigma}_{K}^{-1})</td>
<td>1.39</td>
<td>1.16</td>
<td>0.67</td>
<td>0.72</td>
<td>0.65</td>
<td>0.67</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>(\hat{\Sigma}_{K})</td>
<td>12.60</td>
<td>9.91</td>
<td>3.84</td>
<td>4.01</td>
<td>4.10</td>
<td>4.33</td>
<td>4.58</td>
</tr>
<tr>
<td></td>
<td>(\Sigma^{-1/2}\hat{\Sigma}_K\Sigma^{-1/2})</td>
<td>4.90</td>
<td>2.90</td>
<td>0.58</td>
<td>0.64</td>
<td>0.68</td>
<td>0.71</td>
<td>0.75</td>
</tr>
<tr>
<td>p = 200</td>
<td>(\hat{\Sigma}_{u,K}^T)</td>
<td>2.66</td>
<td>1.75</td>
<td>0.54</td>
<td>0.56</td>
<td>0.58</td>
<td>0.61</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>((\hat{\Sigma}_{u,K}^T))^{-1}</td>
<td>1.47</td>
<td>1.29</td>
<td>0.72</td>
<td>0.71</td>
<td>0.71</td>
<td>0.74</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>(\hat{\Sigma}_{K}^{-1})</td>
<td>1.45</td>
<td>1.27</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.72</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>(\hat{\Sigma}_{K})</td>
<td>19.42</td>
<td>14.83</td>
<td>5.69</td>
<td>6.06</td>
<td>6.12</td>
<td>5.67</td>
<td>5.60</td>
</tr>
<tr>
<td></td>
<td>(\Sigma^{-1/2}\hat{\Sigma}_K\Sigma^{-1/2})</td>
<td>5.13</td>
<td>3.13</td>
<td>0.66</td>
<td>0.72</td>
<td>0.77</td>
<td>0.81</td>
<td>0.88</td>
</tr>
<tr>
<td>p = 300</td>
<td>(\hat{\Sigma}_{u,K}^T)</td>
<td>2.83</td>
<td>1.86</td>
<td>0.55</td>
<td>0.56</td>
<td>0.58</td>
<td>0.60</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>((\hat{\Sigma}_{u,K}^T))^{-1}</td>
<td>1.55</td>
<td>1.39</td>
<td>0.74</td>
<td>0.75</td>
<td>0.78</td>
<td>0.84</td>
<td>1.05</td>
</tr>
<tr>
<td></td>
<td>(\hat{\Sigma}_{K}^{-1})</td>
<td>1.53</td>
<td>1.35</td>
<td>0.73</td>
<td>0.74</td>
<td>0.77</td>
<td>0.82</td>
<td>1.01</td>
</tr>
<tr>
<td></td>
<td>(\hat{\Sigma}_{K})</td>
<td>24.95</td>
<td>17.13</td>
<td>6.91</td>
<td>7.22</td>
<td>7.46</td>
<td>6.80</td>
<td>7.44</td>
</tr>
<tr>
<td></td>
<td>(\Sigma^{-1/2}\hat{\Sigma}_K\Sigma^{-1/2})</td>
<td>5.48</td>
<td>3.08</td>
<td>0.70</td>
<td>0.77</td>
<td>0.83</td>
<td>0.88</td>
<td>0.97</td>
</tr>
</tbody>
</table>

Figure 3.10 plots the estimation errors as functions of \(K\), which shows a significant drop at \(K = 3\). When \(K \geq 3\), as \(K\) becomes larger, the estimation error slightly increases.

### 3.2.5 Comparisons with Other Methods

**Comparison with related methods**

We compare POET with related methods that address low-rank plus sparse covariance estimation, specifically, LOREC proposed by Luo (2012), the strict factor model (SFM) by Fan, Fan and Lv (2008), the Dual Method (Dual) by Lin et al. (2009), and finally, the singular value thresholding (SVT) by Cai, Candès and Shen (2008). In particular, SFM is a special case of POET which employs a large threshold that
Figure 3.10: Robustness of $K$ for various $p$. Design 2

Top left: $\| \hat{\Sigma}_{u,K} - \Sigma_u \|_F / \sqrt{p}$, top right: $\| (\hat{\Sigma}_{u,K})^{-1} - \Sigma_u^{-1} \|_F / \sqrt{p}$; bottom left: $\| \hat{\Sigma}_K^{-1} - \Sigma^{-1} \|_F / \sqrt{p}$, bottom right: $\| \hat{\Sigma}_K - \Sigma \|_F / \sqrt{p}$. $T$ is fixed to be 300.
forces $\hat{\Sigma}_u$ to be diagonal even when the true $\Sigma_u$ might not be. Note that Dual, SVT and many others dealing with low-rank plus sparse, such as Candès et al. (2011) and Wright et al. (2009), assume a known $\Sigma$ and focus on recovering the decomposition. Hence they do not estimate $\Sigma$ or its inverse, but decompose the sample covariance into two components. The resulting sparse component may not be positive definite, which can lead to large estimation errors for $\hat{\Sigma}_u^{-1}$ and $\hat{\Sigma}^{-1}$.

Data are generated from the same setup as Design 2 in Section 6.4. Table 3.9 reports the averaged estimation error of the four comparing methods, calculated based on 50 replications for each simulation. Dual and SVT assume the data matrix has a low-rank plus sparse representation, which is not the case for the sample covariance matrix (though the population $\Sigma$ has such a representation). The tuning parameters for POET, LOREC, Dual and SVT are chosen to achieve the best performance for each method.\footnote{We used the R package for LOREC developed by Luo (2012) and the Matlab codes for Dual and SVT provided on Yi Ma’s website “Low-rank matrix recovery and completion via convex optimization” at University of Illinois. The tuning parameters for each method have been chosen to minimize the sum of relative errors $\|\Sigma^{-1/2} \hat{\Sigma} \Sigma^{-1/2} - I_p\|_F + \|\Sigma_u^{-1/2} \hat{\Sigma}_u \Sigma_u^{-1/2} - I_p\|_F$. We have also written an R package for POET.} We see that POET outperforms all the other methods. It is not surprising to see that LOREC has large estimation errors since it is not designed for estimating covariances with diverging eigenvalues.

Comparison with direct thresholding

This section compares POET with direct thresholding on the sample covariance matrix without taking out common factors (Rothman et al. 2009, Cai and Liu 2011. We denote this method by THR). We also run simulations to demonstrate the finite sample performance when $\Sigma$ itself is sparse and has bounded eigenvalues, corresponding to the case $K = 0$. Three models are considered and both POET and THR use the soft thresholding. We fix $T = 200$. Reported results are the average of 100 replications.
Table 3.7: Method Comparison under spectral norm for $T = 100$. RelE represents the relative error $\| \Sigma^{-1/2} \hat{\Sigma} \Sigma^{-1/2} - I_p \|$

<table>
<thead>
<tr>
<th></th>
<th>$\Sigma_u$</th>
<th>$\Sigma_u^{-1}$</th>
<th>RelE</th>
<th>$\Sigma^{-1}$</th>
<th>$\Sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p = 100$</td>
<td>POET</td>
<td>1.624</td>
<td>1.336</td>
<td>2.080</td>
<td>1.309</td>
</tr>
<tr>
<td></td>
<td>LOREC</td>
<td>2.274</td>
<td>1.880</td>
<td>2.564</td>
<td>1.511</td>
</tr>
<tr>
<td></td>
<td>SFM</td>
<td>2.084</td>
<td>2.039</td>
<td>2.707</td>
<td>2.022</td>
</tr>
<tr>
<td></td>
<td>Dual</td>
<td>2.306</td>
<td>5.654</td>
<td>2.707</td>
<td>4.674</td>
</tr>
<tr>
<td></td>
<td>SVT</td>
<td>2.59</td>
<td>13.64</td>
<td>2.806</td>
<td>103.1</td>
</tr>
<tr>
<td>$p = 200$</td>
<td>POET</td>
<td>1.641</td>
<td>1.358</td>
<td>3.295</td>
<td>1.346</td>
</tr>
<tr>
<td></td>
<td>LOREC</td>
<td>2.179</td>
<td>1.767</td>
<td>3.874</td>
<td>1.543</td>
</tr>
<tr>
<td></td>
<td>SFM</td>
<td>2.098</td>
<td>2.071</td>
<td>3.758</td>
<td>2.065</td>
</tr>
<tr>
<td></td>
<td>Dual</td>
<td>2.41</td>
<td>6.554</td>
<td>4.541</td>
<td>5.813</td>
</tr>
<tr>
<td></td>
<td>SVT</td>
<td>2.930</td>
<td>362.5</td>
<td>4.680</td>
<td>47.21</td>
</tr>
<tr>
<td>$p = 300$</td>
<td>POET</td>
<td>1.662</td>
<td>1.394</td>
<td>4.337</td>
<td>1.395</td>
</tr>
<tr>
<td></td>
<td>LOREC</td>
<td>2.364</td>
<td>1.635</td>
<td>4.909</td>
<td>1.742</td>
</tr>
<tr>
<td></td>
<td>SFM</td>
<td>2.091</td>
<td>2.064</td>
<td>4.874</td>
<td>2.061</td>
</tr>
<tr>
<td></td>
<td>Dual</td>
<td>2.475</td>
<td>2.602</td>
<td>6.190</td>
<td>2.234</td>
</tr>
<tr>
<td></td>
<td>SVT</td>
<td>2.681</td>
<td>$&gt;10^3$</td>
<td>6.247</td>
<td>$&gt;10^3$</td>
</tr>
</tbody>
</table>

**Model 1: one-factor.** The factors and loadings are independently generated from $N(0, 1)$. The error covariance is the same banded matrix as Design 2 in Section 6.4. Here $\Sigma$ has one diverging eigenvalue.

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

**Model 3: cross-sectional AR(1).** Set $K = 0$, but $\Sigma = \Sigma_u = (0.85)^{|i-j|} p \times p$. Now $\Sigma$ is no longer sparse (or banded), but is not too dense either since $\Sigma_{ij}$ decreases to zero exponentially fast as $|i - j| \to \infty$. This is the correlation matrix if $\{y_{it}\}_{i=1}^p$ follows a cross-sectional AR(1) process: $y_{it} = 0.85 y_{i-1,t} + \varepsilon_{it}$.

For each model, POET uses an estimated $\hat{K}$ based on IC1 of Bai and Ng (2002), while THR thresholds the sample covariance directly. We find that in Model 1, POET performs significantly better than THR as the latter misses the common factor. For Model 2, IC1 estimates $\hat{K} = 0$ precisely in each replication, and hence POET is
identical to THR. For Model 3, POET still outperforms. The results are summarized in Table 3.8.

<table>
<thead>
<tr>
<th></th>
<th>(|\Sigma - \Sigma|)</th>
<th>(|\Sigma^{-1} - \Sigma^{-1}|)</th>
<th>(\hat{K})</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>POET</td>
<td>THR</td>
<td>POET</td>
</tr>
<tr>
<td>(p = 200)</td>
<td>(\times) 200</td>
<td>(\times) 240.18</td>
<td>1.31</td>
</tr>
<tr>
<td>Model 1</td>
<td>26.20</td>
<td>240.18</td>
<td>1.31</td>
</tr>
<tr>
<td>Model 2</td>
<td>2.04</td>
<td>2.04</td>
<td>2.07</td>
</tr>
<tr>
<td>Model 3</td>
<td>7.73</td>
<td>11.24</td>
<td>8.48</td>
</tr>
<tr>
<td>(p = 300)</td>
<td>(\times) 300</td>
<td>(\times) 314.43</td>
<td>2.18</td>
</tr>
<tr>
<td>Model 1</td>
<td>32.60</td>
<td>314.43</td>
<td>2.18</td>
</tr>
<tr>
<td>Model 2</td>
<td>2.03</td>
<td>2.03</td>
<td>2.08</td>
</tr>
<tr>
<td>Model 3</td>
<td>9.41</td>
<td>11.29</td>
<td>8.81</td>
</tr>
</tbody>
</table>

The reported numbers are the averages based on 100 replications.

3.2.6 Simulated portfolio allocation

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

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

\[
\hat{\mathbf{w}} = \text{argmin}_{\mathbf{w}^{\prime} = 1} \mathbf{w}^{\prime}\hat{\Sigma}\mathbf{w},
\]

the estimated (minimum variance) portfolio. Then the actual risk of the estimated portfolio is defined as \(R(\hat{\mathbf{w}}) = \hat{\mathbf{w}}^{\prime}\Sigma\hat{\mathbf{w}}\), and the estimated risk (also called empirical risk) is equal to \(\hat{R}(\hat{\mathbf{w}}) = \hat{\mathbf{w}}^{\prime}\hat{\Sigma}\hat{\mathbf{w}}\). In practice, the actual risk is unknown, and only the empirical risk can be calculated. The actual risk is usually higher than the empirical
<table>
<thead>
<tr>
<th></th>
<th>$\Sigma_u$</th>
<th>$\Sigma_u^{-1}$</th>
<th>Rel. Error</th>
<th>$\Sigma^{-1}$</th>
<th>$\Sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p = 100$</td>
<td>POET</td>
<td>0.522</td>
<td>0.537</td>
<td>0.552</td>
<td>4.104</td>
</tr>
<tr>
<td></td>
<td>LOREC</td>
<td>0.825</td>
<td>0.775</td>
<td>0.750</td>
<td>4.339</td>
</tr>
<tr>
<td></td>
<td>SFM</td>
<td>0.825</td>
<td>1.115</td>
<td>1.103</td>
<td>4.420</td>
</tr>
<tr>
<td></td>
<td>Dual</td>
<td>0.979</td>
<td>1.261</td>
<td>1.003</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SVT</td>
<td>1.115</td>
<td>4.038</td>
<td>1.025</td>
<td></td>
</tr>
<tr>
<td>$p = 200$</td>
<td>POET</td>
<td>0.521</td>
<td>0.580</td>
<td>0.572</td>
<td>5.808</td>
</tr>
<tr>
<td></td>
<td>LOREC</td>
<td>0.764</td>
<td>0.671</td>
<td>0.916</td>
<td>5.922</td>
</tr>
<tr>
<td></td>
<td>SFM</td>
<td>0.827</td>
<td>1.125</td>
<td>1.154</td>
<td>5.806</td>
</tr>
<tr>
<td></td>
<td>Dual</td>
<td>0.953</td>
<td>1.124</td>
<td>1.415</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SVT</td>
<td>1.239</td>
<td>$&gt; 10^3$</td>
<td>1.408</td>
<td></td>
</tr>
<tr>
<td>$p = 300$</td>
<td>POET</td>
<td>0.524</td>
<td>0.594</td>
<td>0.638</td>
<td>6.064</td>
</tr>
<tr>
<td></td>
<td>LOREC</td>
<td>0.889</td>
<td>0.973</td>
<td>1.110</td>
<td>6.944</td>
</tr>
<tr>
<td></td>
<td>SFM</td>
<td>0.828</td>
<td>1.127</td>
<td>1.178</td>
<td>6.736</td>
</tr>
<tr>
<td></td>
<td>Dual</td>
<td>0.973</td>
<td>1.201</td>
<td>1.728</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SVT</td>
<td>1.276</td>
<td>$&gt; 10^3$</td>
<td>1.701</td>
<td></td>
</tr>
</tbody>
</table>

Risk, but ideally a good estimator of $\hat{\Sigma}_y$ should make them converge to each other as the dimensionality $p$ increases.

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

It is interesting to examine the accuracy and the performance of the actual risk of our portfolio $\hat{\Sigma}$ in comparison to the oracle risk $R^* = \min_{\mathbf{w}} \mathbf{1}_1=1 \mathbf{w}' \Sigma \mathbf{w}$, which is the theoretical risk of the portfolio we would have created if we knew the true covariance matrix $\Sigma$. We thus compare the regret $R(\hat{\Sigma}) - R^*$, which is always nonnegative, for three estimators of $\hat{\Sigma}$. They are summarized by using the box plots over the
200 simulations. The results are reported in Figure 3.11. In practice, we are also concerned about the difference between the actual and empirical risk of the chosen portfolio $\hat{w}$. Hence, in Figure 3.12, we also compare the average estimation error $|R(\hat{w}) - \hat{R}(\hat{w})|$ and the average relative estimation error $|\hat{R}(\hat{w})/R(\hat{w}) - 1|$ over 200 simulations. When $\hat{w}$ is obtained based on the strict factor model, both differences - between actual and oracle risk, and between actual and empirical risk, are persistently greater than the corresponding differences for the approximate factor estimator. Also, in terms of the relative estimation error, the factor model based method is negligible, where as the sample covariance does not process such a property.

Figure 3.11: Box plots of regrets $R(\hat{w}) - R^*$ for $p = 40$, $p = 80$, $p = 120$ and $p = 140$. In each panel, the box plots from left to right correspond to $\hat{w}$ obtained using $\hat{\Sigma}$ based on approximate factor model, strict factor model, and sample covariance, respectively.
Figure 3.12: Estimation errors for risk assessments as a function of the portfolio size $p$. Left panel plots the average absolute error $|R(\hat{w}) - \hat{R}(\hat{w})|$ and right panel depicts the average relative error $|\hat{R}(\hat{w})/R(\hat{w}) - 1|$. Here, $\hat{w}$ and $\hat{R}$ are obtained based on three estimators of $\hat{\Sigma}$. 
When the sample covariance is used, its performance is better for small values of \( p \), such as \( p = 40 \), but both its mean and variability increase rapidly with \( p \). On the other hand, the variability of both factor models decreases with dimensionality. Thus, approximate factor model gives a portfolio whose actual risk is closest to both the oracle risk and the empirical risk among the three methods of estimating \( \Sigma \).

This result is based on a simulation in which \( \Sigma_u \) is in an average of around 10% off-diagonal sparsity for \( p < 100 \), 1% off-diagonal sparsity for \( p \) between 100 and 140. When simulating \( \Sigma_u \), we found larger dimensionality requires larger threshold to achieve invertibility. In reality, we can expect the true error covariance matrix to be less sparse than that, which would make our estimator perform better than the strict factor model more significantly.

### 3.3 Real Data Example

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

#### 3.3.1 Sparsity of Idiosyncratic Errors

The data were obtained from the CRSP (The Center for Research in Security Prices) database, and consists of \( p = 50 \) stocks and their annualized daily returns for the period January 1\(^{st}\), 2010-December 31\(^{st}\), 2010 \((T = 252)\). The stocks are chosen from 5 different industry sectors, (more specifically, Consumer Goods-Textile & Apparel Clothing, Financial-Credit Services, Healthcare-Hospitals, Services-Restaurants, Utilities-Water utilities), with 10 stocks from each sector. We made this selection to demonstrate a block diagonal trend in the sparsity. More specifically, we show that the non-zero elements are clustered mainly within companies in the same industry.
We also notice that these are the same groups that show predominantly positive correlation.

The largest eigenvalues of the sample covariance equal 0.0102, 0.0045 and 0.0039, while the rest are bounded by 0.0020. Hence $K = 0, 1, 2, 3$ are the possible values of the number of factors. Figure 3.13 shows the heatmap of the thresholded error correlation matrix (for simplicity, we applied hard thresholding). The threshold has been chosen using the cross validation as described in Section 4. We compare the level of sparsity (percentage of non-zero off-diagonal elements) for the 5 diagonal blocks of size $10 \times 10$, versus the sparsity of the rest of the matrix. For $K = 2$, our method results in 25.8% non-zero off-diagonal elements in the 5 diagonal blocks, as opposed to 7.3% non-zero elements in the rest of the covariance matrix. Note that, out of the non-zero elements in the central 5 blocks, 100% are positive, as opposed to a distribution of 60.3% positive and 39.7% negative amongst the non-zero elements in off-diagonal blocks. There is a strong positive correlation between the returns of companies in the same industry after the common factors are taken out, and the thresholding has preserved them. The results for $K = 1, 2$ and 3 show the same characteristics. These provide stark evidence that the strict factor model is not appropriate.

### 3.3.2 Portfolio Allocation

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

The number of factors is determined using the penalty function proposed by Bai and Ng (2002), as defined in (2.2.14). For calibration, we use the last 100 consecutive
Figure 3.13: Heatmap of thresholded error correlation matrix for number of factors $K = 0$, $K = 1$, $K = 2$ and $K = 3$.

Business days of the above data, and both IC1 and IC2 give $\hat{K} = 3$. On the 1st of each month, we estimate $\hat{\Sigma}_{\text{diag}}$ (SFM) and $\hat{\Sigma}_{\hat{K}}$ (POET with soft thresholding) using the historical data of excess daily returns for the proceeding 12 months ($T = 252$). The value of the threshold is determined using the cross-validation procedure. We minimize the empirical risk of both portfolios to obtain the two respective optimal portfolio allocations $\hat{\mathbf{w}} = \hat{\mathbf{w}}_1$ and $\hat{\mathbf{w}}_2$ (based on $\hat{\Sigma} = \hat{\Sigma}_{\text{diag}}$ and $\hat{\Sigma}_{\hat{K}}$): $\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \mathbf{w}' \mathbf{1} \Sigma \mathbf{w}$.

At the end of the month (21 trading days), their actual risks are compared, calculated by

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

We can see from Figure 3.14 that the minimum-risk portfolio created by the POET estimator performs significantly better, achieving lower variance 76% of the time.
Amongst those months, the risk is decreased by 48.63%. On the other hand, during the months that POET produces a higher-risk portfolio, the risk is increased by only 17.66%.

Next, we demonstrate the impact of the choice of number of factors and threshold on the performance of POET. If cross-validation seems computationally expensive, we can choose a common soft-threshold throughout the whole investment process. The average constant in the cross-validation was 0.53, close to our suggested constant 0.5 used for simulation. We also present the results based on various choices of constant $C = 0.5, 0.75, 1$ and $1.25$, with soft threshold $C\sqrt{\hat{\theta}_{ij}^T\omega_T}$. The results are summarized in Table 3.10. The performance of POET seems consistent across different choices of these parameters.

Figure 3.14: Risk of portfolios created with POET and SFM (strict factor model)
Table 3.10: Comparisons of the risks of portfolios using POET and SFM: The first number is proportion of the time POET outperforms and the second number is percentage of average risk improvements. $C$ represents the constant in the threshold.

<table>
<thead>
<tr>
<th>$C$</th>
<th>$\hat{K} = 1$</th>
<th>$\hat{K} = 2$</th>
<th>$\hat{K} = 3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>0.58/29.6%</td>
<td>0.68/38%</td>
<td>0.71/33%</td>
</tr>
<tr>
<td>0.5</td>
<td>0.66/31.7%</td>
<td>0.70/38.2%</td>
<td>0.75/33.5%</td>
</tr>
<tr>
<td>0.75</td>
<td>0.68/29.3%</td>
<td>0.70/29.6%</td>
<td>0.71/25.1%</td>
</tr>
<tr>
<td>1</td>
<td>0.66/20.7%</td>
<td>0.62/19.4%</td>
<td>0.69/18%</td>
</tr>
</tbody>
</table>
Chapter 4

Conclusion

In this thesis, we studied the rate of convergence of high dimensional covariance matrix of approximate factor models under various norms. By assuming conditional sparsity of the error covariance matrix, we allow for the presence of the cross-sectional correlation even after taking out common factors. Realizing unconditional sparsity assumption is inappropriate in many applications, 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. We also expand the results based on whether the common factors are observable or latent and establish the price one has to pay for not knowing the 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 generalized shrinkage function of Antoniadis and Fan (2001). This includes the adaptive, hard and soft thresholding techniques as specific examples. We then constructed the covariance matrix of $y_t$ 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 $(\hat{\Sigma}^T)^{-1}$ and $(\hat{\Sigma}_u^T)^{-1}$ is of order $O_p(m_p(\sqrt{\log p}/T + \sqrt{1/p}))$, where the second term comes from the impact of estimating the unobservable factors. This demonstrates when estimating the inverse covariance matrix, $p$ is allowed to be much larger than $T$.

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

We provide extensive simulation and empirical studies, for both the known and unknown factors case. We create a setup similar to the Fama-French model, and compare the performance of our estimator to others, such as the strict factor model and the direct thresholding of the sample covariance matrix.

The proposed POET also has wide applicability in statistical genomics. For example, Carvalho et al. (2008) applied a Bayesian sparse factor model to study the breast cancer hormonal pathways. Their real-data results have identified about two common factors that have highly loaded genes (about half of 250 genes). As a result, these factors should be treated as “pervasive” (see the explanation in Example 2.1), which will result in one or two very spiked eigenvalues of the gene expressions’ covariance matrix. The POET can be applied to estimate such a covariance matrix and its network model.
Chapter 5

Technical Proofs

5.1 Proofs for Section 1.2

5.1.1 Lemmas

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

Lemma 5.1.1. Suppose that \( A \) and \( B \) are symmetric semi-positive definite matrices, and \( \lambda_{\text{min}}(B) > c_T \) for a sequence \( c_T > 0 \). If \( \|A - B\| = o_p(c_T) \), then \( \lambda_{\text{min}}(A) > c_T/2 \), and

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

Proof. Suppose both \( A \) and \( B \) are \( m \times m \). For any \( v \in \mathbb{R}^m \) such that \( \|v\| = 1 \),

\[
|v'(A - B)v| \leq \|v\|^2 \|A - B\| = o_p(c_T).
\]

Hence for all large \( T \), \( v'Av \geq v'Bv - 0.5c_T \geq \lambda_{\text{min}}(B) - 0.5c_T > 0.5c_T \). Hence, \( \lambda_{\text{min}}(A) \geq 0.5c_T \). In addition,

\[
\|A^{-1} - B^{-1}\| = \|A^{-1}(B - A)B^{-1}\| \\
\leq \lambda_{\text{min}}(A)^{-1}\|A - B\|\lambda_{\text{min}}(B)^{-1} \\
= O_p(c_T^{-2})\|A - B\|.
\]
Lemma 5.1.2. 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 for some $r_3$ and $b_3 > 0$, and any $s > 0$,

$$P(|Z_1Z_2| > s) \leq \exp(1 - (s/b_3)^{r_3}). \quad (5.1.1)$$

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 $r = r_1 r_2 / (r_1 + r_2)$,

$$P(|Z_1Z_2| > s) \leq P(M|Z_1| > s) + P(|Z_2| > M) \leq \exp(1 - (s/b_1 M)^{r_1}) + \exp(1 - (M/b_2)^{r_2}) \leq 2 \exp(1 - (s/b)^r).$$

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 $s > b_3$,

$$P(|Z_1Z_2| > s) \leq 2 \exp(1 - (s/b)^r) \leq \exp(1 - (s/b_3)^{r_3}).$$

When $s \leq b_3$,

$$P(|Z_1Z_2| > s) \leq 1 \leq \exp(1 - (s/b_3)^{r_3}).$$

Q.E.D.
Lemma 5.1.3. Under Assumption 2.3.2, $a_T = o(1)$ 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) $$\max_{i,j \leq p} |\hat{\sigma}_{ij} - \sigma_{ij}| = O_p(\max\{\sqrt{\frac{\log p}{T}}, a_T\}).$$

Proof. (i) By Assumption 2.3.2(iii) and Lemma 5.1.2, $u_{it}u_{jt}$ 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 2.3.2(iii)),

$$P(\left| \sum_{t=1}^{T} u_{it}u_{jt} - \sigma_{ij} \right| \geq s) \leq T \exp\left( -\frac{(Ts)^{\gamma}}{C_1} \right) + \exp\left( -\frac{T^2s^2}{C_2(1+TC_3)} \right) + \exp\left( -\frac{(Ts)^2}{C_4T} \exp\left( \frac{(Ts)^{(1-\gamma)}}{C_5(\log Ts)^{\gamma}} \right) \right).$$

Using Bonferroni’s method, we have

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

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

$$p^2T \exp\left( -\frac{(Ts)^{\gamma}}{C_1} \right) + p^2 \exp\left( -\frac{(Ts)^2}{C_4T} \exp\left( \frac{(Ts)^{(1-r)}}{C_5(\log Ts)^{r}} \right) \right) = o(1).$$

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

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

which implies the desired result.
(ii) We have, by part (i) and the triangular inequality,

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

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

\[
A \leq \max_{i,j \leq p} \left| \frac{1}{T} \sum_{t=1}^{T} (\hat{u}_{it} - u_{it})(\hat{u}_{jt} - u_{jt}) \right| + 2 \max_{i,j \leq p} \left| \frac{1}{T} \sum_{t=1}^{T} u_{it}(\hat{u}_{jt} - u_{jt}) \right|
\]

\[
\leq \max_{i \leq p} \frac{1}{T} \sum_{t=1}^{T} (\hat{u}_{it} - u_{it})^2 + 2 \sqrt{\max_{i \leq p} \frac{1}{T} \sum_{t=1}^{T} u_{it}^2} \sqrt{\max_{i \leq p} \frac{1}{T} \sum_{t=1}^{T} (\hat{u}_{it} - u_{it})^2}
\]

\[
= O_p(a_T^2) + 2 \sqrt{o_p(1) + \max_{i \leq p} \sigma_{ii}^2} \sqrt{a_T^2}
\]

\[
= O_p(a_T).
\]

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

**Lemma 5.1.4.** 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

\[
\hat{\theta}_{i,j} = \frac{1}{T} \sum_{t=1}^{T} (\hat{u}_{it} \hat{u}_{jt} - \frac{1}{T} \sum_{l} \hat{u}_{il} \hat{u}_{jl})^2
\]

\[
\leq \frac{2}{T} \sum_{t=1}^{T} (\hat{u}_{it} \hat{u}_{jt} - \sigma_{ij})^2 + 2 \max_{i,j} (\sigma_{ij} - \frac{1}{T} \sum_{l} \hat{u}_{il} \hat{u}_{jl})^2
\]

\[
\leq \frac{2}{T} \sum_{t=1}^{T} (\hat{u}_{it} \hat{u}_{jt} - \sigma_{ij})^2 + o_p(1),
\]

92
where the term $o_p(1)$ does not depend on $i, j$, by Lemma 5.6.2. Still by adding and subtracting terms,

$$\frac{1}{2} \sum_t (\hat{u}_{it}\hat{u}_{jt} - \sigma_{ij})^2 \leq \sum_t (\hat{u}_{it} - u_{it})^2 \hat{u}_{jt}^2 + \sum_t (\hat{u}_{jt} - u_{jt})^2 u_{it}^2$$

$$+ \sum_t (u_{it}u_{jt} - \sigma_{ij})^2 \leq 2 \max_{it} |\hat{u}_{it} - u_{it}|^2 (\max_j \sum_t (\hat{u}_{jt} - u_{jt})^2)$$

$$+ \max_j \sum_t u_{jt}^2) + \sum_t (u_{it}u_{jt} - \sigma_{ij})^2 \leq o_p(1)(o_p(T\sigma_j^2) + \max_j \sum_t u_{jt}^2)) + \sum_t (u_{it}u_{jt} - \sigma_{ij})^2.$$

Since $(u_t)_{t \geq 1}$ are i.i.d. normal random vectors, the same arguments as those in the proof of Lemma 2 in Cai and Liu (2011) imply that

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

and var$(u_{it}u_{jt})$ is bounded away from both zero and infinity. Therefore, $\frac{4}{T} \sum_t (u_{it}u_{jt} - \sigma_{ij})^2$ is bounded away from zero and infinity with probability approaching one. In addition, by Lemma 5.6.2(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, $\max_{i,j} \hat{\theta}_{ij}$ is bounded away from infinity with probability approaching one.

(ii) By adding and subtracting terms, we obtain

$$\sum_t (u_{it}u_{jt} - \sigma_{ij})^2 \leq 4 \sum_t (u_{it}u_{jt} - \hat{u}_{it}\hat{u}_{jt})^2$$

$$+ 4 \sum_t (\hat{u}_{it}\hat{u}_{jt} - \frac{1}{T} \sum_t \hat{u}_{it}\hat{u}_{jt})^2 + 4 \sum_t (\sigma_{ij} - \frac{1}{T} \sum_t \hat{u}_{it}\hat{u}_{jt})^2$$
\[
\leq 8 \sum_t u_t^2(u_{jt} - \hat{u}_{jt})^2 + 8 \sum_t \hat{u}_{jt}^2(u_{it} - \hat{u}_{it})^2 + 4T\hat{\theta}_{ij} + o_p(T)
\]
\[
\leq 16 \max_t |\hat{u}_{it} - u_{it}|^2(\max_j (\hat{u}_{jt} - u_{jt})^2 + \max_j u_{jt}^2) + 4T\hat{\theta}_{ij} + o_p(T),
\]

where \(o_p(T)\) does not depend on \(i, j\) due to Lemma 5.6.2. As is demonstrated in part (i),

\[
16 \max_t |\hat{u}_{it} - u_{it}|^2(\max_j (\hat{u}_{jt} - u_{jt})^2 + \max_j u_{jt}^2) = o_p(T),
\]
\[
\frac{1}{T} \sum_t (u_{it}u_{jt} - \sigma_{ij})^2 \geq C \text{ uniformly in } i, j \text{ for some } C > 0 \text{ w.p.a.1.}
\]

This establishes the result. Q.E.D.

5.1.2 Proof of Theorem 1.2.1

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

\[
\|\tilde{\Sigma}_u - \Sigma_u\| \leq \|\Sigma_u - \Sigma_T\| + \|\tilde{\Sigma}_u - \Sigma_T\|,
\]

where

\[
\Sigma^T_u = (\sigma^T_{ij}), \quad \sigma^T_{ij} = \sigma_{ij}I(\|\sigma_{ij}\| > \sqrt{\hat{\theta}_{ij}}\omega_T).
\]

and \(\omega_T = C \max(\sqrt{\log p/T}, a_T)\) for some \(C > 0\). We bound \(\|\Sigma^T_u - \Sigma_u\|\) and \(\|\tilde{\Sigma}^T_u - \Sigma^T_u\|\) separately.

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

\[
\|\Sigma^T_u - \Sigma_u\| \leq \max_{i \leq p} \sum_{j=1}^p |\sigma_{ij}|I(|\sigma_{ij}| \leq \omega_T\hat{\theta}_{ij}^{1/2})
\]
\[
\leq \max_i \sum_{j: \sigma_{ij} \neq 0} \omega_T\hat{\theta}_{ij}^{1/2} = O_p(\omega_T m_T),
\]

94
where the last equality is due to \( \hat{\theta}_{ij} \) is bounded above uniformly in \( i,j \) with probability approaching one.

On the other hand,

\[
\| \hat{\Sigma}_u^T - \Sigma_u^T \| \leq \max_{i \leq p} \sum_{j=1}^p |\sigma_{ij}| I(|\hat{\sigma}_{ij}| \leq \omega_T \hat{\theta}_{ij}^{1/2}, |\sigma_{ij}| > \omega_T \hat{\theta}_{ij}^{1/2}) \\
+ \max_{i \leq p} \sum_{j=1}^p |\sigma_{ij} - \hat{\sigma}_{ij}| I(|\hat{\sigma}_{ij}| > \omega_T \hat{\theta}_{ij}^{1/2}, |\sigma_{ij}| > \omega_T \hat{\theta}_{ij}^{1/2}) \\
+ \max_{i \leq p} \sum_{j=1}^p |\hat{\sigma}_{ij}| I(|\sigma_{ij}| \leq \omega_T \hat{\theta}_{ij}^{1/2}, |\hat{\sigma}_{ij}| > \omega_T \hat{\theta}_{ij}^{1/2}).
\]

Since by Lemma 5.1.4, \( \hat{\theta}_{ij}^{1/2} \) 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,

\[
\max_{i \leq p} \sum_{j=1}^p I(|\hat{\sigma}_{ij} - \sigma_{ij}| \geq (1 - r)\omega_T \hat{\theta}_{ij}) = O_p(1). \tag{5.1.2}
\]

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

\[
\max_{i \leq p} \sum_{j=1}^p |\hat{\sigma}_{ij} - \sigma_{ij}| I(|\hat{\sigma}_{ij}| \geq \omega_T \hat{\theta}_{ij}, |\sigma_{ij}| \leq r\omega_T \hat{\theta}_{ij}) \\
\leq O_p(\omega_T) \max_{i \leq p} \sum_{j=1}^p I(|\hat{\sigma}_{ij}| \geq \omega_T \hat{\theta}_{ij}, |\sigma_{ij}| \leq r\omega_T \hat{\theta}_{ij}) \\
= O_p(\omega_T). \tag{5.1.3}
\]

To show (5.1.2), let \( C_1 > 0 \) be such that \( P(\min_{ij} \hat{\theta}_{ij} \leq C_1) = o(1) \), whose existence is guaranteed by Lemma 5.1.4. Since \( \max_{ij} |\hat{\sigma}_{ij} - \sigma_{ij}| = O_p(\omega_T) \), for any \( \epsilon, M > 0 \), and
sufficiently large $C > 0$,

\[
P(\max_{i \leq p} \sum_{j=1}^{p} I(|\hat{\sigma}_{ij} - \sigma_{ij}| \geq (1 - r)\omega_T\hat{\theta}_{ij}) > M)
\leq P(\max_{ij} |\hat{\sigma}_{ij} - \sigma_{ij}| \geq (1 - r)\omega_T\hat{\theta}_{ij})
\leq P\left(\frac{\max_{ij} |\hat{\sigma}_{ij} - \sigma_{ij}|}{\max\{\sqrt{(\log p)/T},a_T\}} \geq (1 - r)CC_1\right) + o(1) < \epsilon,
\]

which yields the result.

(ii) Since both $\hat{\Sigma}_T$ and $\Sigma_u$ are symmetric and $\lambda_{\min}(\Sigma_u) > C$ for some $C > 0$, the result follows immediately from Lemma 5.6.1.

Q.E.D.
5.2 Proofs for Section 1.3

5.2.1 Proof of Theorem 1.3.1

Lemma 5.2.1. (i) \( \max_{i \leq p} \| \hat{b}_i - b_i \| = O_p \left( \sqrt{\frac{K \log p}{T}} \right) \).

(ii) \( \max_{i,j \leq K} \left| \frac{1}{T} \sum_{t=1}^{T} f_{it} f_{jt} - E f_{it} f_{jt} \right| = O_p \left( \frac{\log K}{T} \right) \).

Proof. As \( \hat{b}_i - b_i = (XX')^{-1}Xu_i \), we have

\[
\| \hat{b}_i - b_i \|^2 = u_i'X'(XX')^{-2}Xu_i.
\]

Since \( \{f_t\}_{t \leq T} \) is ergodic and stationary, \( \| \frac{1}{T}XX' - Ef_t f_t' \| = o_p(1) \). Then \( \lambda_{\min}(\text{cov}(f_t)) > c > 0 \) implies that the smallest eigenvalue of \( \frac{1}{T}XX' \) is bounded away from zero with probability approaching one. Then

\[
\| \hat{b}_i - b_i \|^2 \leq C \| \frac{1}{T}Xu_i \|^2 = C \sum_{k=1}^{K} \left( \frac{1}{T} \sum_{t=1}^{T} f_{kt} u_{it} \right)^2 \leq O_p(K) \max_{k \leq K, i \leq p} \left( \frac{1}{T} \sum_{t=1}^{T} f_{kt} u_{it} \right)^2.
\]

Let \( Z_{ki} = \frac{1}{T} \sum_{t=1}^{T} f_{kt} u_{it} \). We bound \( |Z_{ki}| \) using Bernstein type inequality. By Lemma 5.1.2, and Assumptions 2.3.2(iii) and 2.3.4(ii), \( f_{kt} u_{it} \) satisfies the exponential tail condition (1.2.6) for some \( r_3 = \frac{r_2}{2r_2 + 2} \in (0, 1) \), as well as the strong mixing condition. Hence by the Bernstein inequality for weakly dependent data in Merlevède (2009, Theorem 1), there exist \( C_i > 0, i = 1, \ldots, 5 \), for any \( s > 0 \)

\[
P(|Z_{ki}| > s) \leq T \exp \left( -\frac{(Ts)^{r_3}}{C_1} \right) + \exp \left( -\frac{T^2 s^2}{C_2(1 + TC_3)} \right) + \exp \left( -\frac{C_4 T}{(Ts)^{r_3(1-r_3)}} \right) \exp \left( \frac{(Ts)^{r_3(1-r_3)}}{C_5(\log TS)^r_3} \right).
\]
The Bonferroni’s method then implies that, if \((\log p)^{2/r_3-1} = o(T)\),

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

It then yields the desired result.

(ii) Lemma 5.1.2 implies that for any \(i\) and \(j \leq K\), \(f_{it}f_{jt}\) satisfies the exponential tail condition \((1.2.6)\) 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 5.2.2.** When \((f_t)_{t \geq 1}\) are observable,

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

(ii) \(\max_{i,t} |u_{it} - \hat{u}_{it}| = o_p(1)\).

**Proof.** (i) \(\max_{i \leq p} \frac{1}{T} \sum_{t=1}^{T} |u_{it} - \hat{u}_{it}|^2 \leq \max_{i \leq p} \frac{1}{T} \sum_{t} \|f_t\|^2 \|\hat{b}_i - b_i\|^2\). Note that

\[
\frac{1}{T} \sum_{t} \|f_t\|^2 \leq K \max_{k \leq K} \frac{1}{T} \sum_{t=1}^{T} f_{kt}^2 - E f_{kt}^2 + K \max_{k \leq K} E f_{kt}^2 = O_p(K).
\]

The result then follows immediately from Lemma 5.6.7.

(ii) By Lemma 2.3.4, \(E\|K^{-1/2}f_t\|^4 < M\). Hence Lemma D.2 in Kitamura et al (2004) yields \(\max_{t \leq T} \|f_t\| = o(T^{1/4}/\sqrt{K})\). We then have

\[
\max_{t \leq T, i \leq p} |u_{it} - \hat{u}_{it}| = \max_{t \leq T, i \leq p} |(\hat{b}_i - b_i)' f_t| = O_p(K \sqrt{\log p} T^{-1/4}) = o_p(1).
\]

**Proof of Theorem 1.3.1** Theorem 1.3.1 follows immediately from Theorem 1.2.1 and Lemma 5.6.8. Q.E.D.
5.2.2 Proof of Theorem 1.3.2 Part (i)

We follow similar lines of proof as in Fan, Fan and Lv (2008). Define

\[ D_T = \hat{\text{cov}}(f_t) - \text{cov}(f_t), \quad H = X'(XX')^{-1}X, \]

\[ C_T = \hat{B} - B, \quad E = (u_1, ..., u_T). \]

With probability approaching one,

\[
\| \hat{\Sigma}^T - \Sigma \|^2 \leq \text{Const} \left[ \| BD_T B' \|^2 + \| B\hat{\text{cov}}(f)C_T' \|^2 \right.
\]

\[
\left. + \| C_T\hat{\text{cov}}(f)C_T' \|^2 + \| \hat{\Sigma}^T_u - \Sigma_u \|^2 \right]. \tag{5.2.1}
\]

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

**Lemma 5.2.3.**

(i) \( \| D_T \|^2_F = O_p\left( \frac{K^2 \log K}{T} \right) \).

(ii) \( \| C_T \|^2_F = O_p(pK/T) \).

**Proof.** (i) It follows immediately from Lemma 5.6.7(ii) since

\[ 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} \right)^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} \right)^2 \).

(ii) \( C_T = EX'(XX')^{-1} \). By the facts that \( E(u_{it}^2) \) is bounded uniformly in \( i, t \), and \( (u_t)_{t=1}^T \) are i.i.d., we have

\[ E\| C_T \|^2_F = E\text{tr}((XX')^{-1}XE(E'X)X'(XX')^{-1}) = O_p(KP/T). \]

**Lemma 5.2.4.** \( \| BD_T B' \|^2_F + \| B\hat{\text{cov}}(f)C_T' \|^2_F = O_p(K/T + K^2 \log K/(Tp)) \).

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

\[ \| B'\Sigma^{-1}B \| \leq 2\| \text{cov}(f_i)^{-1} \| = O(1). \]
Hence $\|BD_TB'\|_F^2 \leq p^{-1}\|D_TB'\Sigma^{-1}B\|_F^2 = O_p(p^{-1})\|D_T\|_F^2 = K^2 \log K/(Tp)$.

On the other hand,

$$\|B\tilde{\text{cov}}(f)C_T'\|_F^2 \leq 8T^{-2}\|BXX'C_T'\|_F^2 + 8T^{-4}\|BX11'X'C_T'\|_F^2.$$ 

Respectively, $\|BXX'C_T'\|_F^2 \leq p^{-1}\|XX'C_n\Sigma^{-1}\|_F\|C_nXX'\Sigma^{-1}B\|_F = O_p(TK)$.

Likewise, $\|BX11'X'C_T'\|_F^2 = O_p(KT^3)$. This yields the result.

**Lemma 5.2.5.** $\|C_T\tilde{\text{cov}}(f)C_T'\|_F^2 = O_p\left(\frac{pK^2}{T}\right)$.

**Proof.** Straightforward calculation yields:

$$p\|C_T\tilde{\text{cov}}(f)C_T'\|_F^2 = \text{tr}(C_T\tilde{\text{cov}}(f)C_T'\Sigma^{-1}C_T\tilde{\text{cov}}(f)C_T'\Sigma^{-1})$$

$$\leq \|EX'(XX')^{-1}\tilde{\text{cov}}(f)(XX')^{-1}XE\Sigma^{-1}\|_F^2$$

$$= O_p(1)\|EX'(XX')^{-1}\|_F^2.$$

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

**Proof of Theorem 1.3.2 Part (i)**

(a) By Corollary 1.4.1, we have

$$\|\hat{\Sigma}_u^T - \Sigma_u\|_\Sigma = p^{-1/2}\|\Sigma^{-1/2}(\hat{\Sigma}_u^T - \Sigma_u)\Sigma^{-1/2}\|_F \leq \|\Sigma^{-1/2}(\hat{\Sigma}_u^T - \Sigma_u)\Sigma^{-1/2}\|_F$$

$$\leq \|\hat{\Sigma}_u^T - \Sigma_u\| \cdot \lambda_{\text{max}}(\Sigma^{-1}) = O_p(Km_T\sqrt{\log p/T}).$$

(5.2.2)

Therefore, (5.2.1) and Lemma 5.6.3-5.6.10 yield

$$\|\hat{\Sigma}^T - \Sigma\|_\Sigma^2 = O_p\left(\frac{pK^2}{T^2} + \frac{K^2m_T^2\log p}{T}\right).$$
(b) For the infinity norm, it is straightforward to find that

\[
\|\hat{\Sigma}^T - \Sigma\|_\infty \leq K^2 \|B\|_\infty \|\hat{\text{cov}}(f) - \text{cov}(f)\|_\infty (\|\hat{B} - B\|_\infty + \|B\|_\infty) \\
+ K^2 \|\hat{B} - B\|_\infty (\|\hat{\text{cov}}(f) - \text{cov}(f)\|_\infty + \|\text{cov}(f)\|_\infty)(\|\hat{B} - B\|_\infty + \|B\|_\infty) \\
+ K^2 \|B\|_\infty \|\text{cov}(f)\|_\infty \|\hat{B} - B\|_\infty \\
+ \|\hat{\Sigma}_u^T - \Sigma_u\|_\infty.
\]

In addition, we have

\[
\left\| \frac{1}{T}EX' \right\|_\infty = \max_{i \leq K, j \leq p} \left\| \frac{1}{T} \sum_{t=1}^T f_{it} u_{jt} \right\| = O_p(\sqrt{\log p T})
\]

and

\[
\|\hat{\text{cov}}(f) - \text{cov}(f)\|_\infty = O_p(\sqrt{\log K T}).
\]

In addition, \(\left\| \left(\frac{1}{T}XX'\right)^{-1} \right\|_\infty \leq \lambda_{\max}\left(\left(\frac{1}{T}XX'\right)^{-1}\right) \leq \lambda_{\min}^{-1}(\hat{\text{cov}}(f)) = O_p(1). \) Hence

\[
\|\hat{B} - B\|_\infty \leq K \left\| \frac{1}{T}EX' \right\|_\infty \left\| \left(\frac{1}{T}XX'\right)^{-1} \right\|_\infty = O_p(K \sqrt{\log p T}).
\]

Inserting \(\|B\|_\infty = O(1), \|\hat{\text{cov}}(f)\|_\infty, \|B - \hat{B}\|_\infty,\) and \(\|\hat{\text{cov}}(f) - \text{cov}(f)\|_\infty\) into (5.2.3) yields

\[
\|\hat{\Sigma}^T - \Sigma\|_\infty \leq O_p(K^3 \sqrt{\log p T}) + \|\hat{\Sigma}_u^T - \Sigma_u\|_\infty.
\]

Moreover, the \((i, j)\)th entry of \(\hat{\Sigma}_u^T - \Sigma_u\) is given by

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

When \(|\hat{\sigma}_{ij}| < \omega T \sqrt{\hat{\theta}_{ij}}, |\sigma_{ij}| \leq |\sigma_{ij} - \hat{\sigma}_{ij}| + |\hat{\sigma}_{ij}| = O_p(\omega T), \) by Lemmas 5.6.2 and 5.1.4. Hence \(\|\hat{\Sigma}_u^T - \Sigma_u\|_\infty = O_p(\omega T), \) which yields the result. Q.E.D.
5.2.3 Proof of Theorem 1.3.2 Part (ii)

Lemma 5.2.6. (i) $\|\hat{\text{cov}}(f)^{-1} + \hat{B}'(\hat{\Sigma}_u^T)^{-1}\hat{B}]^{-1}\| = O_p(p^{-1}),$

$\|\text{cov}(f)^{-1} + B'S_u^{-1}B]^{-1}\| = O(p^{-1}).$

(ii) $\lambda_{\text{min}}(B'S_u^{-1}B) \geq cp$ for some $c > 0.$

Proof. By Assumption, $\lambda_{\text{min}}(\text{cov}(f)^{-1} + B'S_u^{-1}B) \geq \lambda_{\text{min}}(B'S_u^{-1}B) > cp$ for a constant $c > 0.$ In addition, $\|\hat{B} - B\| = O_p(\sqrt{pK/T}), \|\text{cov}(f)^{-1} - \hat{\text{cov}}(f)^{-1}\|_F = O_p(K/\sqrt{T}),$ and that $\|(\hat{\Sigma}_u^T)^{-1} - \Sigma_u^{-1}\| = o_p(1).$ The same argument of Fan, Fan and Lv (2008) (eq. 14) implies that $\|B\|_F = O(\sqrt{p}).$ Therefore,

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

The results in (i) then follow from Lemma 5.6.1.

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

$$\lambda_{\text{min}}(B'S_u^{-1}B) = v'B'S_u^{-1}Bv \geq \lambda_{\text{min}}(\Sigma_u^{-1})v'B'Bv \geq c\lambda_{\text{min}}(B'B)$$

given that $\|\Sigma_u\|$ is bounded. By Assumption 2.3.1 and Lemma 5.6.1, the smallest eigenvalue of $B'B \geq c_1p$ for some $c_1 > 0,$ which completes the proof.

Q.E.D.
Using the Sherman-Morrison-Woodbury formula, we have

\[
\| (\hat{\Sigma}_T^T)^{-1} - \Sigma^{-1} \| = \| (\hat{\Sigma}_u^T)^{-1} - \Sigma_u^{-1} \| \tag{5.2.5}
\]

\[
+ \| ((\hat{\Sigma}_T^T)^{-1} - \Sigma_u^{-1}) \tilde{B} [\tilde{\text{cov}}(f)^{-1} + \tilde{B}' (\hat{\Sigma}_u^T)^{-1} \tilde{B}]^{-1} \tilde{B}' (\hat{\Sigma}_u^T)^{-1} \| \\
+ \| ((\hat{\Sigma}_T^T)^{-1} - \Sigma_u^{-1}) \tilde{B} [\tilde{\text{cov}}(f)^{-1} + \tilde{B}' (\hat{\Sigma}_u^T)^{-1} \tilde{B}]^{-1} \tilde{B}' \Sigma_u^{-1} \|
\]

\[
+ \| \Sigma_u^{-1} (\tilde{B} - B) [\tilde{\text{cov}}(f)^{-1} + \tilde{B}' (\hat{\Sigma}_u^T)^{-1} \tilde{B}]^{-1} \Sigma_u^{-1} \|
\]

\[
+ \| \Sigma_u^{-1} (\tilde{B} - B) [\tilde{\text{cov}}(f)^{-1} + \tilde{B}' (\hat{\Sigma}_u^T)^{-1} \tilde{B}]^{-1} B' \Sigma_u^{-1} \|
\]

\[
+ \| \Sigma_u^{-1} B ([\tilde{\text{cov}}(f)^{-1} + \tilde{B}' (\hat{\Sigma}_u^T)^{-1} \tilde{B}]^{-1} - [\text{cov}(f)^{-1} + B' \Sigma_u^{-1} B]^{-1}) B' \Sigma_u^{-1} \|
\]

\[= L_1 + L_2 + L_3 + L_4 + L_5 + L_6. \]

Theorem 1.3.1 implies \( L_1 = O_p(Km_T\sqrt{\frac{\log p}{T}}). \)

Let \( G = [\tilde{\text{cov}}(f)^{-1} + \tilde{B}' (\hat{\Sigma}_u^T)^{-1} \tilde{B}]^{-1} \), then

\[
L_2 \leq \| ((\hat{\Sigma}_u^T)^{-1} - \Sigma_u^{-1}) (\hat{\Sigma}_u^T)^{-1/2} \| \cdot \| (\hat{\Sigma}_u^T)^{-1/2} \tilde{B} G \tilde{B}' (\hat{\Sigma}_u^T)^{-1/2} \| \cdot \| (\hat{\Sigma}_u^T)^{-1/2} \|. \tag{5.2.6}
\]

Note that \( \| \hat{\Sigma}_u^T \| \leq \| \hat{\Sigma}_T^T - \Sigma_u \| + \lambda_{\text{max}}(\Sigma_u) < C \) for a constant \( C > 0 \). By Lemma 5.6.1 and Theorem 1.3.1, \( \lambda_{\text{max}}(\hat{\Sigma}_u^T)^{-1/2} \leq C. \)

In addition, the middle term in (5.2.6) can be treated in the same way as in the proof of (28) in Fan, Fan and Lv (2008):

\[
(\hat{\Sigma}_u^T)^{-1/2} \tilde{B} G \tilde{B}' (\hat{\Sigma}_u^T)^{-1/2} = I - (\hat{\Sigma}_u^T)^{1/2} (\hat{\Sigma}_u^T)^{-1} (\hat{\Sigma}_u^T)^{1/2} \leq I.
\]

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

\[
L_3 \leq \| ((\hat{\Sigma}_u^T)^{-1} - \Sigma_u^{-1}) (\hat{\Sigma}_u^T)^{-1/2} \| \cdot \| (\hat{\Sigma}_u^T)^{-1/2} \tilde{B} G \tilde{B}' (\hat{\Sigma}_u^T)^{-1/2} \| \cdot \| \Sigma_u^{-1} (\hat{\Sigma}_u^T)^{1/2} \|
\]

\[= O_p(L_1). \]
Lemma 5.6.11 shows that $\|G\| = O_p(p^{-1})$. Hence

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

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

$$\|G - G_1\| \leq \|G'(G^{-1} - G_1^{-1})G_1\|
\leq \|G\|\|G_1\|\|\text{cov}(f)^{-1} + B' \Sigma_u^{-1}B - (\hat{\text{cov}}(f)^{-1} + \hat{B}' (\hat{\Sigma}_u^T)^{-1}\hat{B})\|
\leq O_p(p^{-2})\|\text{cov}(f)^{-1} - \hat{\text{cov}}(f)^{-1}\| + O_p(p^{-2})\|B' \Sigma_u^{-1}B - \hat{B}' (\hat{\Sigma}_u^T)^{-1}\hat{B}\|
= O_p(p^{-1}K m_T \sqrt{\frac{\log p}{T}}).$$

Therefore $L_6 \leq \|\Sigma_u^{-1}B\|^2 \|G - G_1\| = O_p(K m_T \sqrt{\frac{\log p}{T}})$. The proof is completed by combining $L_1 \sim L_6$. Q.E.D.
5.3 Proofs for Section 4

The OLS is given by

$$\hat{b}_i = (X'_i X_i)^{-1} X'_i y_i, i \leq p.$$ 

The same arguments in the proof of Lemma 5.6.7 can yield

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

which then implies the rate of

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

The result then follows from a straightforward application of Theorem 1.2.1.

5.4 Estimating a sparse covariance with contaminated data

We estimate $\Sigma_u$ by applying the adaptive thresholding given by (2.2.11). However, the task here is slightly different from the standard problem of estimating a sparse covariance matrix in the literature, as no direct observations for $\{u_t\}_{t=1}^T$ are available. In many cases the original data are contaminated, including any type of estimate of the data when direct observations are not available. This typically happens when $\{u_t\}_{t=1}^T$ represent the error terms in regression models or when data is subject to measurement of errors. Instead, we may observe $\{\hat{u}_t\}_{t=1}^T$. For instance, in the approximate factor models, $\hat{u}_t = y_t - \hat{b}_i f_t$. 

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

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

$$
\hat{\Sigma}_{u}^T = (s_{ij}(\hat{\sigma}_{ij}))_{p \times p}, \quad (5.4.1)
$$

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

When $\{\hat{u}_t\}_{t=1}^T$ is close enough to $\{u_t\}_{t=1}^T$, we can show that $\hat{\Sigma}_{u}^T$ is also consistent. The following theorem extends the standard thresholding results in Bickel and Levina (2008) and Cai and Liu (2011) to the case when no direct observations are available, or the original data are contaminated. For the tail and mixing parameters $r_1$ and $r_3$ defined in Assumptions 2.3.2 and 2.3.3, let $\alpha = 3r_1^{-1} + r_3^{-1} + 1$.

**Theorem 5.4.1.** Suppose $(\log p)^6 \alpha = o(T)$, and Assumptions 2.3.2 and 2.3.3 hold. In addition, suppose there is a sequence $a_T = o(1)$ so that $\max_{i \leq p} \frac{1}{T} \sum_{t=1}^{T} |u_{it} - \hat{u}_{it}|^2 = O_p(a_T^2)$, and $\max_{i \leq p, t \leq T} |u_{it} - \hat{u}_{it}| = o_p(1)$; Then there is a constant $C > 0$ in the adaptive thresholding estimator (5.4.1) with

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

such that

$$
\|\hat{\Sigma}_{u}^T - \Sigma_u\| = O_p(\omega_T^{1-\alpha} m_p).
$$

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

$$
\|(\hat{\Sigma}_{u}^T)^{-1} - \Sigma_u^{-1}\| = O_p(\omega_T^{1-\alpha} m_p).
$$

**Proof.** By Assumptions 2.3.2 and 2.3.3, the conditions of Lemmas A.3 and A.4 of Fan, Liao and Mincheva (2011, *Ann. Statist*, 39, 3320-3356) are satisfied. Hence for
any $\epsilon > 0$, there are positive constants $M, \theta_1$ and $\theta_2$ such that each of the events

\[
A_1 = \left\{ \max_{i \leq p, j \leq p} |\tilde{\sigma}_{ij} - \sigma_{u,ij}| < M\omega_T \right\}
\]

\[
A_2 = \{ \theta_1 > \sqrt{|\tilde{\theta}_{ij}|} > \theta_2, \text{ all } i \leq p, j \leq p \}.
\]

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

\[
\|\tilde{\Sigma}_u^T - \Sigma_u\| \leq \max_{i \leq p} \sum_{j=1}^p |s_{ij}(\tilde{\sigma}_{ij}) - \sigma_{u,ij}|
\]

\[
= \max_{i \leq p} \sum_{j=1}^p |s_{ij}(\tilde{\sigma}_{ij})I_{(|\tilde{\sigma}_{ij}| > C\omega_T \sqrt{\theta_{ij}})} - \sigma_{u,ij}I_{(|\tilde{\sigma}_{ij}| > C\omega_T \sqrt{\theta_{ij}})} - \sigma_{u,ij}I_{(|\tilde{\sigma}_{ij}| \leq C\omega_T \sqrt{\theta_{ij}})}| + \sum_{j=1}^p |\sigma_{u,ij}|I_{(|\tilde{\sigma}_{ij}| \leq C\omega_T \sqrt{\theta_{ij}})}| 
\]

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

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

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

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

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

Let $M_1 = (C\theta_1 + M)(M^{-q} + (C\theta_1 + M)^{-q})$. Then with probability at least $1 - 2\epsilon$, $\|\tilde{\Sigma}_u^T - \Sigma_u\| \leq m_p \omega_T^{1-q} M_1$. Since $\epsilon$ is arbitrary, we have $\|\tilde{\Sigma}_u^T - \Sigma_u\| = O_p(\omega_T^{1-q} m_p)$.

If in addition, $\omega_T m_p = o(1)$, then the minimum eigenvalue of $\tilde{\Sigma}_u^T$ is bounded away
from zero with probability approaching one since \( \lambda_{\min}(\Sigma_u) > c_1 \). This then implies
\[
\| (\hat{\Sigma}_u)^{-1} - \Sigma_u^{-1} \| = O_p(\omega_1^{1-q}m_p).
\]

5.5 Proofs for Section 2.2

We first cite two useful theorems, which are needed to prove propositions 2.1 and 2.2. In Lemma 5.5.1 below, let \( \{\lambda_i\}_{i=1}^p \) be the eigenvalues of \( \Sigma \) in descending order and \( \{\xi_i\}_{i=1}^p \) be their associated eigenvectors. Correspondingly, let \( \{\hat{\lambda}_i\}_{i=1}^p \) be the eigenvalues of \( \hat{\Sigma} \) in descending order and \( \{\hat{\xi}_i\}_{i=1}^p \) be their associated eigenvectors.

Lemma 5.5.1.  1. (Weyl’s Theorem) \( |\hat{\lambda}_i - \lambda_i| \leq \| \hat{\Sigma} - \Sigma \| \).

2. (\( \sin \theta \) Theorem, Davis and Kahan, 1970)
\[
\| \hat{\xi}_i - \xi_i \| \leq \frac{\sqrt{2}\| \hat{\Sigma} - \Sigma \|}{\min(|\hat{\lambda}_{i-1} - \lambda_i|, |\lambda_i - \hat{\lambda}_{i+1}|)}.
\]

Proof of Proposition 2.1

Proof. Since \( \{\lambda_j\}_{j=1}^p \) are the eigenvalue of \( \Sigma \) and \( \{\| \hat{\mathbf{b}}_j \|^2\}_{j=1}^K \) are the first \( K \) eigenvalues of \( \mathbf{B} \mathbf{B}' \) (the remaining \( p - K \) eigenvalues are zero), then by the Weyl’s theorem, for each \( j \leq K \),
\[
|\lambda_j - \| \hat{\mathbf{b}}_j \|^2| \leq \| \Sigma - \mathbf{B} \mathbf{B}' \| = \| \Sigma_u \|.
\]

For \( j > K \), \( |\lambda_j| = |\lambda_j - 0| \leq \| \Sigma_u \| \). On the other hand, the first \( K \) eigenvalues of \( \mathbf{B} \mathbf{B} \) are also the eigenvalues of \( \mathbf{B}' \mathbf{B} \). By the assumption, the eigenvalues of \( p^{-1} \mathbf{B}' \mathbf{B} \) are bounded away from zero. Thus when \( j \leq K \), \( \| \hat{\mathbf{b}}_j \|^2/p \) are bounded away from zero for all large \( p \).

Proof of Proposition 2.2
Proof. Applying the sin $\theta$ theorem yields

$$\|\xi_j - \tilde{b}_j\| / \|\tilde{b}_j\| \leq \frac{\sqrt{2} \|\Sigma_u\|}{\min(\|\lambda_{j-1} - \|\tilde{b}_j\|^2, \|\tilde{b}_j\|^2 - \lambda_{j+1})}$$

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

Proof of Theorem 2.2.1

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

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

$$= \frac{1}{T} (YY' - Y \hat{F} \hat{\Lambda}' - \hat{\Lambda} \hat{F}' Y' + \hat{\Lambda} \hat{F}' \hat{F} \hat{\Lambda}')$$

$$= \frac{1}{T} (YY' - Y \hat{F} \hat{\Lambda}' - \hat{\Lambda} \hat{F}' Y' + T \hat{\Lambda} \hat{\Lambda}')$$

$$= \frac{1}{T} YY' - \hat{\Lambda} \hat{\Lambda}'. $$

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

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

we have $\hat{R} = \hat{\Sigma}_u$. Consequently, applying thresholding on $\hat{\Sigma}_u$ is equivalent to applying thresholding on $\hat{R}$, which gives the desired result.
We now show \( \hat{\Lambda}' = \sum_{i=1}^{K} \hat{\lambda}_i \hat{\xi}_i \hat{\xi}'_i \) indeed holds. Consider again the least squares problem (2.2.8) but with the following alternative normalization constraints: \( \frac{1}{p} \sum_{i=1}^{p} b_i b'_i = I_K \), and \( \frac{1}{T} \sum_{t=1}^{T} f_t f'_t \) is diagonal. Let \( (\tilde{\Lambda}, \tilde{F}) \) be the solution to the new optimization problem. Switching the roles of \( B \) and \( F \), then the solution of (2.2.10) is \( \tilde{\Lambda} = (\hat{\xi}_1, \cdots, \hat{\xi}_K) \) and \( \tilde{F} = p^{-1} Y' \tilde{\Lambda} \). In addition, \( T^{-1} \tilde{F}' \tilde{F} = \text{diag}(\hat{\lambda}_1, \cdots, \hat{\lambda}_K) \).

From \( \tilde{\Lambda} \tilde{F}' = \tilde{\Lambda} F' \), it follows that \( \tilde{\Lambda} \tilde{\Lambda}' = \frac{1}{T} \tilde{F}' \tilde{F} \tilde{\Lambda}' = \frac{1}{T} \tilde{F}' \tilde{F} \tilde{\Lambda}' = \sum_{i=1}^{K} \hat{\lambda}_i \hat{\xi}_i \hat{\xi}'_i. \) \( \square \)

5.6 Proofs for Section 2.3

We will proceed by subsequently showing Theorems 2.3.3, 2.3.1 and 2.3.2.

5.6.1 Preliminary lemmas

The following results are to be used subsequently. The proofs of Lemmas 5.6.1, 5.6.2 and 5.6.3 are found in Fan, Liao and Mincheva (2011).

Lemma 5.6.1. Suppose \( A, B \) are symmetric semi-positive definite matrices, and \( \lambda_{\min}(B) > c_T \) for a sequence \( c_T > 0 \). If \( \|A - B\| = o_p(c_T) \), then \( \lambda_{\min}(A) > c_T / 2 \), and

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

Lemma 5.6.2. 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 \( \forall s > 0, \)

\[
P(|Z_i| > s) \leq \exp\left(-\frac{s}{b_i}^r\right), \quad i = 1, 2.
\]

Then for some \( r_3 \) and \( b_3 > 0 \), and any \( s > 0, \)

\[
P(|Z_1 Z_2| > s) \leq \exp(1 - \frac{s}{b_3}^r).
\] (5.6.1)
Lemma 5.6.3. Under the assumptions of Theorem 2.3.1,

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

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

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

Lemma 5.6.4. Let \( \hat{\lambda}_K \) denote the \( K \)th largest eigenvalue of \( \hat{\Sigma}_{\text{sam}} = \frac{1}{T} \sum_{t=1}^{T} y_t y_t' \),

then \( \hat{\lambda}_K > C_1 p \) with probability approaching one for some \( C_1 > 0 \).

Proof. First of all, by Proposition 2.2.1, under Assumption 2.3.1, the \( K \)th largest eigenvalue \( \lambda_K \) of \( \Sigma \) satisfies: for some \( c > 0 \),

\[
\lambda_K \geq \| \hat{b}_K \|^2 - |\lambda_K - \| \hat{b}_K \|^2 | \geq cp\| \Sigma_u \| \geq cp/2
\]

for sufficiently large \( p \). Using Weyl’s theorem, we need only to prove that \( \| \hat{\Sigma}_{\text{sam}} - \Sigma \| = o_p(p) \). Without loss of generality, we prove the result under the identifiability condition (2.2.1). Using model (2.1.2), \( \hat{\Sigma}_{\text{sam}} = T^{-1} \sum_{t=1}^{T} (Bf_t + u_t)(Bf_t + u_t)' \). Using this and (2.1.3), \( \hat{\Sigma}_{\text{sam}} - \Sigma \) can be decomposed as the sum of the four terms:

\[
D_1 = (T^{-1} B \sum_{t=1}^{T} f_t f_t' - I_K)B', \quad D_2 = T^{-1} \sum_{t=1}^{T} (u_t u_t' - \Sigma_u),
\]

\[
D_3 = BT^{-1} \sum_{t=1}^{T} f_t u_t', \quad D_4 = D_3'
\]

We now deal them term by term. We will repeatedly use the fact that for a \( p \times p \) matrix \( A \),

\[
\| A \| \leq p \| A \|_{\text{max}}.
\]

First of all, by Lemma 5.6.3, \( \| T^{-1} \sum_{t=1}^{T} f_t f_t' - I_K \| \leq K \| T^{-1} \sum_{t=1}^{T} f_t f_t' - I_K \|_{\text{max}} = O_p(\sqrt{1/T}) \), which is \( o_p(p) \) if \( K \log p = o(T) \). Consequently, by Assumption 2.3.1, we have

\[
\| D_1 \| \leq O_p(K \sqrt{(\log K)/T}) \| BB' \| = O_p(p \sqrt{1/T}).
\]
We now deal with $\mathbf{D}_2$. It follows from Lemma 5.6.3 that

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

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

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

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

Lemma 5.6.5. Under Assumption 2.3.3, $\max_{s=1}^{T} \sum_{t=1}^{T} |E \mathbf{u}_t' \mathbf{u}_t|/p = O(1)$.

Proof. Since $\{\mathbf{u}_t\}_{t=1}^{T}$ is weakly stationary, $\max_{s=1}^{T} \sum_{t=1}^{T} |E \mathbf{u}_t' \mathbf{u}_t|/p \leq 2 \sum_{t=1}^{\infty} |E \mathbf{u}_t' \mathbf{u}_t|/p$.

In addition, $E|u_{it}|^4 < M$ for some constant $M$ and any $i, t$ since $u_{it}$ has exponential tail. Hence by Davydov’s inequality (Corollary 16.2.4 in Athreya and Lahiri 2006), there is a constant $C > 0$, for all $i \leq p, t \leq T$, $|Eu_{1t}u_{it}| \leq C \sqrt{\alpha(t)}$, where $\alpha(t)$ is the $\alpha$-mixing coefficient. By Assumption 2.3.3, $\sum_{t=1}^{\infty} \sqrt{\alpha(t)} < \infty$. Thus uniformly in $T$,

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

5.6.2 Proof of Theorem 2.3.3

Our derivation below relies on a result obtained by Bai and Ng (2002), which showed that the estimated number of factors is consistent, in the sense that $\hat{K}$ equals the true $K$ with probability approaching one. Note that under our Assumptions 2.3.1-2.3.4, all the assumptions in Bai and Ng (2002) are satisfied. Thus immediately we have the following Lemma.
Lemma 5.6.7. For all \( f \) where \( \zeta \)

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

Proof. See Bai and Ng (2002).

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

\[
\hat{f}_t - Hf_t = (V/p)^{-1} \left( \sum_{s=1}^{T} \hat{f}_s E(u'_s u_t)/p + \sum_{s=1}^{T} \hat{f}_s \zeta_{st} + \sum_{s=1}^{T} \hat{f}_s \eta_{st} + \sum_{s=1}^{T} \hat{f}_s \xi_{st} \right)
\]

(5.6.2)

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

We first prove some preliminary results in the following Lemmas. Denote by \( \hat{f}_t = (\hat{f}_{1t}, ..., \hat{f}_{Tt})' \).

Lemma 5.6.7. For all \( i \leq \hat{K} \),

(i) \( \frac{1}{T} \sum_{t=1}^{T} (\frac{1}{T} \sum_{s=1}^{T} \hat{f}_{is} E(u'_s u_t)/p)^2 = O_p(T^{-1}) \),

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

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

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

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

\[
\frac{1}{T} \sum_{t=1}^{T} (\frac{1}{T} \sum_{s=1}^{T} \hat{f}_{is} E(u'_s u_t)/p)^2 \leq \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{T} (E(u'_s u_t)/p)^2
\]

\[
\leq \max_{t \leq T} \frac{1}{T} \sum_{s=1}^{T} (E(u'_s u_t)/p)^2 \leq \max_{s,t} |E(u'_s u_t)/p| \max_{t \leq T} \frac{1}{T} \sum_{s=1}^{T} |E(u'_s u_t)/p|
\]

By Lemma 5.6.5, \( \max_{t \leq T} \sum_{s=1}^{T} |E(u'_s u_t)/p| = O(1) \), which then yields the result.

(ii) By the Cauchy-Schwarz inequality,

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

113
\[
\frac{1}{T^3} \sum_{s=1}^{T} \sum_{i,s} \hat{f}_{is}^2 \left( \sum_{s,t} \left( \sum_{t=1}^{T} \zeta_{st} \zeta_{st}^2 \right) \right)^{1/2} = \frac{1}{T^2} \left( \sum_{s=1}^{T} \sum_{t=1}^{T} \zeta_{st} \zeta_{st}^2 \right)^{1/2}.
\]

Note that \( E(\sum_{s=1}^{T} \sum_{t=1}^{T} (\sum_{t=1}^{T} \zeta_{st} \zeta_{st}^2)) = T^2 E(\sum_{s=1}^{T} \zeta_{st} \zeta_{st}^2) \leq T^4 \max_{s,t} E|\zeta_{st}|^4 \). By Assumption 2.3.4, \( \max_{s,t} E \zeta_{st}^4 = O(p^{-2}) \), which implies that \( \sum_{s,t} (\sum_{t=1}^{T} \zeta_{st} \zeta_{st}^2) = O_p(T^4/p^2) \), and yields the result.

(iii) By definition, \( \eta_{st} = f' \sum_{i=1}^{p} b_i u_{it}/p \). We first bound \( \| \sum_{i=1}^{p} b_i u_{it} \| \). Assumption 2.3.4 implies \( E \left( \sum_{s=1}^{T} \| \sum_{i=1}^{p} b_i u_{it} \|^2 \right) = E \| \sum_{i=1}^{p} b_i u_{it} \|^2 = O(p) \). Therefore, by the Cauchy-Schwarz inequality,

\[
\frac{1}{T} \sum_{t=1}^{T} \left( \frac{1}{T} \sum_{s=1}^{T} \hat{f}_{is} \eta_{st} \right)^2 \leq \frac{1}{T} \sum_{s=1}^{T} \hat{f}_{is} f' \sum_{s=1}^{T} \sum_{j=1}^{p} b_j u_{jt} \| \frac{1}{p} \| \sum_{j=1}^{p} b_j u_{jt} \| \leq \frac{1}{T} \sum_{s=1}^{T} \| f' \| \cdot \left( \frac{1}{T} \sum_{s=1}^{T} \sum_{j=1}^{p} b_j u_{jt} \| \hat{f}_{is} \| \right)^2 = O_p \left( \frac{1}{p} \right).
\]

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

\[
\frac{1}{T} \sum_{t=1}^{T} \left( \frac{1}{T} \sum_{s=1}^{T} \hat{f}_{is} \xi_{st} \right)^2 = \frac{1}{T} \sum_{t=1}^{T} \left( \frac{1}{T} \sum_{s=1}^{T} f' \sum_{j=1}^{p} b_j u_{jt} \| \hat{f}_{is} \| \right)^2 \leq \frac{1}{T} \sum_{t=1}^{T} \| f' \| \cdot \left( \frac{1}{T} \sum_{s=1}^{T} \sum_{j=1}^{p} b_j u_{jt} \| \hat{f}_{is} \| \right)^2 \leq \frac{1}{T} \sum_{t=1}^{T} \| f' \| \cdot \left( \frac{1}{T} \sum_{s=1}^{T} \sum_{j=1}^{p} b_j u_{jt} \| \hat{f}_{is} \| \right)^2 = O_p \left( \frac{1}{p} \right),
\]

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

\[\square\]

**Lemma 5.6.8.** (i) \( \max_{t \leq T} \| \frac{1}{T} \sum_{s=1}^{T} \hat{f}_{is} E(u'_s u_t) \| = O_p(\sqrt{1/T}) \),

(ii) \( \max_{t \leq T} \| \frac{1}{T} \sum_{s=1}^{T} \hat{f}_{is} \| = O_p(T^{1/4}/\sqrt{p}) \),

(iii) \( \max_{t \leq T} \| \frac{1}{T} \sum_{s=1}^{T} \hat{f}_{is} \eta_{st} \| = O_p(T^{1/4}/\sqrt{p}) \),

(iv) \( \max_{t \leq T} \| \frac{1}{T} \sum_{s=1}^{T} \hat{f}_{is} \xi_{st} \| = O_p(T^{1/4}/\sqrt{p}) \).

**Proof.** (i) By the Cauchy-Schwarz inequality and the fact that \( \frac{1}{T} \sum_{t=1}^{T} \| \hat{f}_{it} \|^2 = O_p(1) \),

\[
\max_{t \leq T} \| \frac{1}{T} \sum_{s=1}^{T} \hat{f}_{is} E(u'_s u_t) \| \leq \max_{t \leq T} \left( \frac{1}{T} \sum_{s=1}^{T} \| \hat{f}_{is} \|^2 \frac{1}{T} \sum_{s=1}^{T} (E(u'_s u_t/p)^2) \right)^{1/2}.
\]

114
\[ \leq O_p(1) \max_{t \leq T} \left( \frac{1}{T} \sum_{s=1}^{T} (E{\bf u}'_s{\bf u}_t/p)^2 \right)^{1/2} \leq O_p(1) \max_{s,t} \sqrt{E{\bf u}'_s{\bf u}_t/p} \max_{t \leq T} \left( \frac{1}{T} \sum_{s=1}^{T} |E{\bf u}'_s{\bf u}_t/p| \right)^{1/2}. \]

The result then follows from Assumption 2.3.3.

(ii) By the Cauchy-Schwarz inequality,

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

It follows from Assumption 2.3.4 that \( E\left( \frac{1}{T} \sum_{s=1}^{T} \zeta_{st}^2 \right)^2 \leq \max_{s,t \leq T} E \zeta_{st}^4 = O\left( \frac{1}{p^2} \right) \).

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 2.3.4, \( E\| \frac{1}{T} \sum_{i=1}^{p} b_i u_{it} \|^4 \leq K^2 M \). Chebyshev’s inequality and Bonferroni’s method yield \( \max_{t \leq T} \| \sum_{i=1}^{p} b_i u_{it} \| = O_p(T^{1/4} \sqrt{p}) \) with probability one, which then implies: \( \max_{t \leq T} \| \frac{1}{T} \sum_{s=1}^{T} \hat{f}_s \eta_{st} \| \leq \| \frac{1}{T} \sum_{s=1}^{T} \hat{f}_s \hat{f}_s' \| \max_{t} \| \frac{1}{T} \sum_{i=1}^{p} b_i u_{it} \| = o_p(T^{1/4}/p^{1/2}) \).

(iv) By the Cauchy-Schwarz inequality and Assumption 2.3.4, we have demonstrated that \( \| \frac{1}{T} \sum_{s=1}^{T} \sum_{i=1}^{p} b_i u_{is} \hat{f}_s \| = O_p(p^{-1/2}) \). In addition, since \( E\| K^{-2} \hat{f}_s \|^4 < M \), \( \max_{t \leq T} \| \hat{f}_t \| = O_p(T^{1/4}) \). It follows that \( \max_{t \leq T} \| \frac{1}{T} \sum_{s=1}^{T} \hat{f}_s \xi_{st} \| \leq \max_{t \leq T} \| \hat{f}_t \| \cdot \| \frac{1}{T} \sum_{s=1}^{T} \sum_{i=1}^{p} b_i u_{is} \| = O_p(T^{1/4}/p^{1/2}) \).

\textbf{Lemma 5.6.9.} (i) \( \max_{t \leq K} \frac{1}{T} \sum_{i=1}^{T} (\hat{f}_i - Hf_i)_t^2 = O_p(1/T + 1/p) \).

(ii) \( \frac{1}{T} \sum_{t=1}^{T} \| \hat{f}_t - Hf_t \|^2 = O_p(1/T + 1/p) \).

(iii) \( \max_{t \leq T} \| \hat{f}_t - Hf_t \| = O_p(\sqrt{1/T} + T^{1/4}/\sqrt{p}) \).

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

(i) When \( \hat{K} = K \), by Lemma 5.6.4, all the eigenvalues of \( V/p \) are bounded away from zero. Using the inequality \((a + b + c + d)^2 \leq 4(a^2 + b^2 + c^2 + d^2) \) and the identity
By Lemma 5.6.3, the first term in (5.6.3) is
\[
\max_{i \leq K} \frac{1}{T} \sum_{t=1}^{T} (\hat{f}_t - \mathbf{H}f_t)^2 \leq C \max_{i \leq K} \frac{1}{T} \sum_{t=1}^{T} \left( \frac{1}{T} \sum_{s=1}^{T} \hat{f}_s E(u_s' u_t) / p \right) + C \max_{i \leq K} \frac{1}{T} \sum_{t=1}^{T} \left( \frac{1}{T} \sum_{s=1}^{T} \hat{f}_s \zeta_{st} \right)^2
\[
+ C \max_{i \leq K} \frac{1}{T} \sum_{t=1}^{T} \left( \frac{1}{T} \sum_{s=1}^{T} \hat{f}_s \eta_{st} \right)^2 + C \max_{i \leq K} \frac{1}{T} \sum_{t=1}^{T} \left( \frac{1}{T} \sum_{s=1}^{T} \hat{f}_s \zeta_{st} \right)^2.
\]
Each of the four terms on the right hand side above are bounded in Lemma 5.6.7, which then yields the desired result.

(ii) It follows from part (i) and that \(\frac{1}{T} \sum_{t=1}^{T} \|\hat{f}_t - \mathbf{H}f_t\|^2 \leq K \max_{i \leq K} \frac{1}{T} \sum_{t=1}^{T} (\hat{f}_t - \mathbf{H}f_t)^2\).

Part (iii) is implied by (5.6.2) and Lemma 5.6.8.

\[\square\]

**Lemma 5.6.10.** (i) \(\mathbf{H}^t \mathbf{H} = \mathbf{I}_K + O_p(1/\sqrt{T} + 1/\sqrt{p})\).

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

**Proof.** We first condition on \(\mathbf{I}_K = K\). (i) Lemma 5.6.4 implies \(\|\mathbf{V}^{-1}\| = O_p(\mathbf{p}^{-1})\). Also \(\|\mathbf{F}\| = \lambda_{\max}^{1/2}(\mathbf{F} ' \mathbf{F}) = \lambda_{\max}^{1/2}(\sum_{t=1}^{T} f_t' f_t) = O_p(\sqrt{T})\). In addition, \(\|\mathbf{F}\| = \sqrt{T}\). It then follows from the definition of \(\mathbf{H}\) that \(\|\mathbf{H}\| = O_p(1)\). Define \(\tilde{\text{cov}}(\mathbf{H}f_t) = \frac{1}{T} \sum_{t=1}^{T} \mathbf{H}f_t (\mathbf{H}f_t)'\). Applying the triangular inequality gives:

\[
\|\mathbf{HH}' - \mathbf{I}_K\|_F \leq \|\mathbf{HH}' - \tilde{\text{cov}}(\mathbf{H}f_t)\|_F + \|\tilde{\text{cov}}(\mathbf{H}f_t) - \mathbf{I}_K\|_F
\]  

(5.6.3)

By Lemma 5.6.3, the first term in (5.6.3) is \(\|\mathbf{HH}' - \tilde{\text{cov}}(\mathbf{H}f_t)\|_F \leq \|\mathbf{H}\|^2 \|\mathbf{I}_K - \tilde{\text{cov}}(\mathbf{f}_t)\|_F = O_p \left( \frac{1}{\sqrt{T}} \right) \). The second term of (5.6.3) can be bounded, by the Cauchy-Schwarz inequality and Lemma 5.6.9, as follows:

\[
\left\| \frac{1}{T} \sum_{t=1}^{T} \mathbf{H}f_t (\mathbf{H}f_t)' - \frac{1}{T} \sum_{t=1}^{T} \hat{f}_t \hat{f}_t' \right\|_F \leq \left\| \frac{1}{T} \sum_{t=1}^{T} (\mathbf{H}f_t - \hat{f}_t) (\mathbf{H}f_t)' \right\|_F + \left\| \frac{1}{T} \sum_{t=1}^{T} \hat{f}_t (\hat{f}_t - \langle \mathbf{H}f_t \rangle)' \right\|_F
\]

\[
\leq \left( \frac{1}{T} \sum_{t=1}^{T} \|\mathbf{H}f_t - \hat{f}_t\|^2 \left( \frac{1}{T} \sum_{t=1}^{T} \|\mathbf{H}f_t\|^2 \right) \right)^{1/2} + \left( \frac{1}{T} \sum_{t=1}^{T} \|\mathbf{H}f_t - \hat{f}_t\|^2 \left( \frac{1}{T} \sum_{t=1}^{T} \|\hat{f}_t\|^2 \right) \right)^{1/2}
\]

116
\[= O_p \left( \frac{1}{\sqrt{T}} + \frac{1}{\sqrt{p}} \right). \]

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

Proof of Theorem 2.3.3

Proof. The second part of this theorem was proved in Lemma 5.6.9. We now derive the convergence rate of \( \max_i \|\hat{b}_i - Hb_i\| \).

Using the facts that \( \hat{b}_i = \frac{1}{T} \sum_{t=1}^{T} y_{it} \hat{f}_t \), and that \( \frac{1}{T} \sum_{t=1}^{T} \hat{f}_t \hat{f}_t' = I_K \), we have

\[
\hat{b}_i - Hb_i = \frac{1}{T} \sum_{t=1}^{T} Hf_t u_{it} + \frac{1}{T} \sum_{t=1}^{T} y_{it} (\hat{f}_t - Hf_t) + H\left( \frac{1}{T} \sum_{t=1}^{T} f_t f_t' - I_K \right)b_i. \tag{5.6.4}
\]

We bound the three terms on the right hand side respectively. It follows from Lemmas 5.6.3 and 5.6.10 that \( \max_i \|\frac{1}{T} \sum_{t=1}^{T} Hf_t u_{it}\| \leq \|H\| \max_i \sqrt{\sum_{k=1}^{K} \left( \frac{1}{T} \sum_{t=1}^{T} f_{kt} u_{it} \right)^2} = O_p \left( \sqrt{\frac{\log p}{T}} \right) \). For the second term, \( E y_{it}^2 = O(1) \). Therefore, \( \max_i T^{-1} \sum_{t=1}^{T} y_{it}^2 = O_p(1) \). The Cauchy-Schwarz inequality and Lemma 5.6.9 imply

\[
\max_i \left\| \frac{1}{T} \sum_{t=1}^{T} y_{it} (\hat{f}_t - Hf_t) \right\| \leq \max_i \left( \frac{1}{T} \sum_{t=1}^{T} y_{it}^2 \frac{1}{T} \sum_{t=1}^{T} \|\hat{f}_t - Hf_t\|^2 \right)^{1/2} = O_p \left( \frac{1}{\sqrt{T}} + \frac{1}{\sqrt{p}} \right).
\]

Finally, \( \|\frac{1}{T} \sum_{t=1}^{T} f_t f_t' - I_K\| = O_p(T^{-1/2}) \) and \( \max_i \|b_i\| = O(1) \) imply that the third term is \( O_p(T^{-1/2}) \).

Proof of Corollary 2.3.1
Under Assumption 2.3.3, it can be shown by Bonferroni’s method that
\[ \max_{t \leq T} \|\mathbf{f}_t\| = O_p((\log T)^{1/2}). \] By Theorem 2.3.3, uniformly in \(i\) and \(t\),
\[
\|\hat{\mathbf{b}}_i \mathbf{f}_t - b'_i \mathbf{f}_t\| \leq \|\hat{\mathbf{b}}_i - \mathbf{H} \mathbf{b}_i\| \|\mathbf{f}_t - \mathbf{H} \mathbf{f}_t\| + \|\mathbf{H} \mathbf{b}_i\| \|\hat{\mathbf{f}}_t - \mathbf{H} \mathbf{f}_t\|
+ \|\hat{\mathbf{b}}_i - \mathbf{H} \mathbf{b}_i\| \|\mathbf{f}_t\| \|\mathbf{H}' \mathbf{H} - \mathbf{I}_K\|
= O_p\left((\log T)^{1/2} \sqrt{\frac{\log p}{T}} + \frac{T^{1/4}}{\sqrt{p}}\right).
\]

5.6.3 Proof of Theorem 2.3.1

Lemma 5.6.11. \(\max_{i \leq p} \frac{1}{T} \sum_{t=1}^T |u_{it} - \hat{u}_{it}|^2 = O_p(\omega^2_T)\), and \(\max_{i,t} |u_{it} - \hat{u}_{it}| = o_p(1)\).

Proof. We have, \(u_{it} - \hat{u}_{it} = b'_i (\mathbf{f}_t - \mathbf{H} \mathbf{f}_t) + (\hat{\mathbf{b}}'_i - b'_i \mathbf{H}') \mathbf{f}_t + b'_i (\mathbf{H}' \mathbf{H} - \mathbf{I}_K) \mathbf{f}_t\). Therefore, using the inequality \((a + b + c)^2 \leq 4a^2 + 4b^2 + 4c^2\), we have:
\[
\max_{i \leq p} \frac{1}{T} \sum_{t=1}^T (u_{it} - \hat{u}_{it})^2 \leq 4 \max_i \|b'_i \mathbf{H}'\|^2 \frac{1}{T} \sum_{t=1}^T \|\mathbf{f}_t - \mathbf{H} \mathbf{f}_t\|^2
+ 4 \max_i \|\hat{\mathbf{b}}'_i - b'_i \mathbf{H}'\|^2 \frac{1}{T} \sum_{t=1}^T \|\mathbf{f}_t\|^2 + 4 \max_i \|\mathbf{b}_i\|^2 \frac{1}{T} \sum_{t=1}^T \|\mathbf{f}_t\|^2 \|\mathbf{H}' \mathbf{H} - \mathbf{I}_K\|^2_F,
\]

The first part of the lemma then follows from Theorem 2.3.3 and Lemma 5.6.9. The second part follows from Corollary 2.3.1.

Proof of Theorem 2.3.1: The theorem follows immediately from Theorem 5.4.1 and Lemma 5.6.11.

5.6.4 Proof of Theorem 2.3.2

Define
\[
\mathbf{C}_T = \hat{\mathbf{A}} - \mathbf{B} \mathbf{H}'.
\]
Lemma 5.6.12. (i) \( \|C_T\|_F^2 = O_p(\omega_T^2 p) \), and \( \|C_T'C_T\|_2^2 = O_p(\omega_T^2 p) \).

(ii) \( \|\hat{\Sigma}_{u,K}^T - \Sigma_u\|_2^2 = O_p(\omega_T^2 - 2q m_p^2) \).

(iii) \( \|BH'C_T\|_2^2 = O_p(\omega_T^2) \).

(iv) \( \|B(H'H - I_K)B\|_2^2 = O(p^{-2} + (pT)^{-1}) \).

Proof. (i) We have \( \|C_T\|_F^2 \leq \max_{i \leq p} \|\hat{b}_i - Hb_i\|^2 p = O_p(\omega_T^2 p) \). Moreover, since all the eigenvalues of \( \Sigma \) are bounded away from zero, for any matrix \( A \), \( \|A\|_2^2 = O_p(p^{-1}) \|A\|_2^2 \). Hence \( \|C'_T C_T\|_2^2 = O_p(p^{-1}) \|C_T\|_F^2 = O(p^2) \).

(ii) By Theorem 2.3.1, \( \|\hat{\Sigma}_{u,K}^T - \Sigma_u\|_2^2 = O_p(\omega_T^2 - 2q m_p^2) \).

(iii) The same argument of the proof of Theorem 2 in Fan, Fan and Lv (2008) implies that \( \|B'\Sigma^{-1}B\| = O(1) \). Thus, \( \|BH'C_T\|_2^2 = p^{-1} \text{tr}(H'C_T \Sigma^{-1} C_T HB'\Sigma^{-1} B) \) is upper bounded by \( p^{-1} \|H\|^2 \|B'\Sigma^{-1}B\| \|\Sigma^{-1}\| \|C_T\|_F^2 = O(p^{-1}) \|C_T\|_F^2 = O(\omega_T^2) \).

(iv) Again, by \( \|B'\Sigma^{-1}B\| = O(1) \), and Lemma 5.6.10,

\[
\|B(H'H - I_K)B\|_2^2 = p^{-1} \text{tr}((H'H - I_K)B'\Sigma^{-1}B(H'H - I_K)B'\Sigma^{-1}B) \\
\leq p^{-1} \|H'H - I_K\|_F^2 \|B'\Sigma^{-1}B\|^2 = O(p^{-2} + (pT)^{-1}) \text{[5.6.5]}
\]

Proof of Theorem 2.3.2 (i)

Proof. By Lemma 5.6.12, \( \|B(H'H - I_K)B\|_2^2 + \|BH'C_T\|_2^2 + \|C_T'C_T\|_2^2 = O_p(\omega_T^2 + p \log^2 p / p) \). Hence for a generic constant \( C > 0 \),

\[
\|\hat{\Sigma}_K - \Sigma\|_2^2 \leq C \|\hat{\Lambda} \hat{\Lambda}' - BB\|_2^2 + C \|\hat{\Sigma}_{u,K}^T - \Sigma_u\|_2^2 \\
\leq C \|B(H'H - I_K)B\|_2^2 + \|BH'C_T\|_2^2 + \|C_T'C_T\|_2^2 + \|\hat{\Sigma}_{u,K}^T - \Sigma_u\|_2^2 \\
= O_p(\omega_T^2 - 2q m_p^2 + p \log^2 p / p).
\]
Lemma 5.6.13. $\|\hat{\Lambda}'(\hat{\Sigma}_u^T)^{-1}\hat{\Lambda} - (BH)'^{-1}\Sigma_u^{-1}BH'\| = O_p(p\omega_T^{1-q}m_p)$.

Proof. $\|C_T\|_F^2 = O_p(\omega_T^2)$. Hence

$$\|\hat{\Lambda}'(\hat{\Sigma}_u^T)^{-1}\hat{\Lambda} - (BH)'^{-1}\Sigma_u^{-1}BH'\| \leq \|C_T(\hat{\Sigma}_u^T)^{-1}C_T\| + 2\|C_T(\hat{\Sigma}_u^T)^{-1}BH'\| + \|BH'((\hat{\Sigma}_u^T)^{-1} - \Sigma_u^{-1})BH'\| = O_p(p\omega_T^{1-q}m_p)\text{[5.6.6]}$$

\[\square\]

Lemma 5.6.14. If $\omega_T^{1-q}m_p = o(1)$, then with probability approaching one, for some $c > 0$,

(i) $\lambda_{\min}(I_K + (BH)'^{-1}\Sigma_u^{-1}BH') \geq cp$.

(ii) $\lambda_{\min}(I_K + \hat{\Lambda}'(\hat{\Sigma}_u^T)^{-1}\hat{\Lambda}) \geq cp$.

(iii) $\lambda_{\min}(I_K + B'\Sigma_u^{-1}B) \geq cp$.

(iv) $\lambda_{\min}((HH')^{-1} + B'\Sigma_u^{-1}B) \geq cp$.

Proof. (i) By Lemma 5.6.10, with probability approaching one, $\lambda_{\min}(HH')$ is bounded away from zero. Hence,

$$\lambda_{\min}(I_K + (BH)'^{-1}\Sigma_u^{-1}BH') \geq \lambda_{\min}((BH)'^{-1}\Sigma_u^{-1}BH')$$

$$\geq \lambda_{\min}(\Sigma_u^{-1})\lambda_{\min}(HB'BH') \geq \lambda_{\min}(\Sigma_u^{-1})\lambda_{\min}(B'B)\lambda_{\min}(HH') \geq cp.$$  

(ii) The result follows from part (i) and Lemma 5.6.13. Part (iii) and (iv) follow from a similar argument of part (i) and Lemma 5.6.10.

\[\square\]

Proof of Theorem 2.3.2:

Proof. We derive the rate for $\|\hat{\Sigma}_K^{-1} - \Sigma^{-1}\|$. Define

$$\tilde{\Sigma} = BH'HB' + \Sigma_u.$$
Note that $\hat{\Sigma}_R = \hat{\Lambda}\hat{\Lambda}' + \hat{\Sigma}^T_{u,R}$ and $\Sigma = B\Sigma + \Sigma_u$. The triangular inequality gives

$$
\|\hat{\Sigma}_R^{-1} - \Sigma^{-1}\| \leq \|\hat{\Sigma}_R^{-1} - \hat{\Sigma}_R^{-1}\| + \|\hat{\Sigma}_R^{-1} - \Sigma^{-1}\|.
$$

Using the Sherman-Morrison-Woodbury formula, we have $\|\hat{\Sigma}_R^{-1} - \Sigma^{-1}\| \leq \sum_{i=1}^6 L_i$, where

\begin{align*}
L_1 & = \|((\hat{\Sigma}^T_{u,R})^{-1} - \Sigma_u^{-1})
L_2 & = \|((\hat{\Sigma}^T_{u,R})^{-1} - \Sigma_u^{-1})\hat{\Lambda}[I_K + \hat{\Lambda}'(\hat{\Sigma}^T_{u,R})^{-1}\hat{\Lambda}]^{-1}\hat{\Lambda}'(\hat{\Sigma}^T_{u,R})^{-1}\| \\
L_3 & = \|((\hat{\Sigma}^T_{u,R})^{-1} - \Sigma_u^{-1})\hat{\Lambda}[I_K + \hat{\Lambda}'(\hat{\Sigma}^T_{u,R})^{-1}\hat{\Lambda}]^{-1}\hat{\Lambda}'\Sigma_u^{-1}\| \\
L_4 & = \|\Sigma_u^{-1}(\hat{\Lambda} - BH')[I_K + \hat{\Lambda}'(\hat{\Sigma}^T_{u,R})^{-1}\hat{\Lambda}]^{-1}\hat{\Lambda}'\Sigma_u^{-1}\| \\
L_5 & = \|\Sigma_u^{-1}(\hat{\Lambda} - BH')[I_K + \hat{\Lambda}'(\hat{\Sigma}^T_{u,R})^{-1}\hat{\Lambda}]^{-1}HB\Sigma_u^{-1}\| \\
L_6 & = \|\Sigma_u^{-1}BH'([I_K + \hat{\Lambda}'(\hat{\Sigma}^T_{u,R})^{-1}\hat{\Lambda}]^{-1} - [I_K + HB\Sigma_u^{-1}BH']^{-1})HB\Sigma_u^{-1}(\hat{\Sigma}^T_{u,R})^{-1}\|.
\end{align*}

We bound each of the six terms respectively. First of all, $L_1$ is bounded by Theorem 2.3.1. Let $G = [I_K + \hat{\Lambda}'(\hat{\Sigma}^T_{u,R})^{-1}\hat{\Lambda}]^{-1}$, then

$$
L_2 \leq \|((\hat{\Sigma}^T_{u,R})^{-1} - \Sigma_u^{-1}) \cdot \|\hat{\Lambda}G\hat{\Lambda}'\| \cdot \|((\hat{\Sigma}^T_{u,R})^{-1})\|.
$$

Note that Theorem 2.3.1 implies $\|((\hat{\Sigma}^T_{u,R})^{-1})\| = O_p(1)$. Lemma 5.6.14 then implies $\|G\| = O_p(p^{-1})$. This shows that $L_2 = O_p(L_1)$. Similarly $L_3 = O_p(L_1)$. In addition, since $\|C_T\|^2_F = \|\hat{\Lambda} - BH'\|^2_F = O_p(\omega_T^2 p)$, $L_4 \leq \|\Sigma_u^{-1}(\hat{\Lambda} - BH')\|\|G\|\|\hat{\Lambda}'\Sigma_u^{-1}\| = O_p(\omega_T)$. Similarly $L_5 = O_p(L_4)$. Finally, let $G_1 = [I_K + (BH')\Sigma_u^{-1}BH']^{-1}$. By Lemma 5.6.14, $\|G_1\| = O_p(p^{-1})$. Then by Lemma 5.6.13,

\begin{align*}
\|G - G_1\| & = \|G(G^{-1} - G_1^{-1})G_1\| \leq O_p(p^{-2})\|(BH')\Sigma_u^{-1}BH' - \hat{\Lambda}'(\hat{\Sigma}^T_{u,R})^{-1}\hat{\Lambda}\| \\
& = O_p(p^{-1} \omega_T^{1-q} m_p).
\end{align*}
Consequently, \( L_6 \leq \| \Sigma_u^{-1} B H' \|^2 \| G - G_1 \| = O_p \left( \omega_1^{1-\eta} m_p \right) \). Adding up \( L_1-L_6 \) gives

\[
\| \hat{\Sigma}_K^{-1} - \Sigma^{-1} \| = O_p \left( \omega_1^{1-\eta} m_p \right).
\]

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

\[
\| \hat{\Sigma}_K^{-1} - \Sigma^{-1} \| \leq \| \Sigma_u^{-1} u_B H' \| \| \Sigma_u^{-1} B \|^{-1} - [I_K + B' \Sigma_u^{-1} B]^{-1} \| B' \Sigma_u^{-1} \|
\]

\[
\leq O(p) \| [(H'H)^{-1} + B' \Sigma_u^{-1} B]^{-1} - [I_K + B' \Sigma_u^{-1} B]^{-1} \| = O_p(p^{-1}) \| (H'H)^{-1} - I_K \| = o_p(\omega_1^{1-\eta} m_p).
\]

\[\Box\]

**Proof of Theorem 2.3.2:** \( \| \hat{\Sigma}^T - \Sigma \|_{\max} \)

*Proof.* We first bound \( \| \hat{\Lambda}' - BB' \|_{\max} \). Repeatedly using the triangular inequality yields

\[
\| \hat{\Lambda}' - BB' \|_{\max} = \max_{i,j \leq p} | \hat{\sigma}_{ij} - \sigma_{ij} | = O_p(\omega_T).
\]

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

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

Hence \( \| \hat{\Sigma}^T_{u,K} - \Sigma_u \|_{\max} = O_p(\omega_T) \). The result then follows immediately. \[\Box\]
Bibliography


Favero, C. and Marcellino, M. (2001). Large datasets, small models and monetary policy in Europe. IEP, Bocconi University, and CEPR.


ZHANG, Y. and EL GHAIU, L. (2011) Large-scale sparse principal component analysis with application to text data. *NIPS*. 

127