Autocorrelation-based Factor Analysis and Nonlinear Shrinkage Estimation of Large Integrated covariance matrix

Qilin Hu

A Thesis Submitted for the Degree of

Doctor of Philosophy

Department of Statistics
London School of Economics and Political Science

Supervisors: Assoc Prof. Clifford Lam and Prof. Piotr Fryzlewicz

London, Aug 2016

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgement is made. This thesis may not be reproduced without my prior written consent.

I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I declare that my thesis consists of 4 chapters.

Statement of conjoint work

I confirm that chapter 3 was jointly co-authored with Assoc Prof. Clifford Lam.

Acknowledgements

I would like to express my whole-hearted gratitude to my Ph.D. supervisor, Assoc Prof. Clifford Lam. For his tremendous support and encouragement during my study. I would also like to thank Prof. Piotr Fryzlewicz for his support.

In addition, I want to express my appreciation to the Department of Statistics at LSE for providing me a fantastic research environment and partial financial support. Furthermore, I am grateful for the generosity of China Scholarship Council (CSC) who provides me financial support during the last three years.

Abstract

The first part of my thesis deals with the factor modeling for high-dimensional time series based on a dimension-reduction viewpoint. we allow the dimension of time series N to be as large as, or even larger than the sample size of the time series. The estimation of the factor loading matrix and subsequently the factors are done via an eigenanalysis on a non-negative definite matrix constructed from autocorrelation matrix. The method is dubbed as AFA. We give explicit comparison of the convergence rates between AFA with PCA. We show that AFA possesses the advantage over PCA when dealing with small dimension time series for both one step and two step estimations, while at large dimension, the performance is still comparable.

The second part of my thesis considers large integrated covariance matrix estimation. While the use of intra-day price data increases the sample size substantially for asset allocation, the usual realized covariance matrix still suffers from bias contributed from the extreme eigenvalues when the number of assets is large. We introduce a novel nonlinear shrinkage estimator for the integrated volatility matrix which shrinks the extreme eigenvalues of a realized covariance matrix back to acceptable level, and enjoys a certain asymptotic efficiency at the same time, all at a high dimensional setting where the number of assets can have the same order as the number of data points. Compared to a time-variation adjusted realized covariance estimator and the usual realized covariance matrix, our estimator demonstrates favorable performance in both simulations and a real data analysis in portfolio allocation. This include a novel maximum exposure bound and an actual risk bound when our estimator is used in constructing the minimum variance portfolio.

Contents

Chapter 1

Introduction

My thesis consists of two pieces of work, the first piece of work deals with factor modeling for time series data, and the second piece of work deals with the estimation of integrated covariance matrix. In this chapter, we give the motivations, question formulation, research hypotheses to investigate, and proposed solutions for both pieces of work.

1.1 Motivation for the first piece of work

In economic or financial time series data, it is typical that on top of serial correlations driven by several common factors, some components of the observed time series can have a specific correlation structure. For instance, world wide performance index of health sector can be driven up by diseases spreading over a certain region on earth, while the global market index can actually be going down suffering from general economic downturn. The disease factor is only affecting the performance index which is specific to the index itself and cannot be explain by other factors driving many other economic indicators. Therefore, it is natural to distinguish the more pervasive factors from the local factors.

In chapter 2, we try to estimate the factors using the so called autocorrelation factor analysis (AFA). By assuming certain regularity conditions, we propose theorems concerning the convergence rate of the methods. It is clear from the theoretical results, AFA method has a big advantage when the noise is very heteroscedatic. For comparison, we have give the counterpart rates for principle components methods (PCA). When we have a large panel of time series where there are many small categories, for example, a panel of macroeconomic indicators, we can use AFA to firstly estimate the more pervasive factors, and remove the effect of the pervasive factors, and use a two step procedure to estimate the local factors. The

convergence rates of the two step procedure with known and unknown grouping structures are also proved. We also design numerical experiments and using a set of real macro economic data to demonstrate our methods.

The contribution of my part is that I explicitly proved the theorems in the chapter and do the numerical and real data examples.

1.2 Motivation for the second piece of work

The second piece of work is concerned with the estimation of the so-called Integrated Covariance. This is a particular type of covariance matrix that arises in financial econometrics from data sets of high-frequency stock returns. The fundamental difference with standard covariance matrix estimation is that the intra-day variance patterns are extremely time-varying, and must be taken into account. Intuitively speaking, the ICV matrix is best understood as the average over a certain time period of instantaneous covariance matrices. In the framework of large-dimensional asymptotics, the largest (smallest) estimated eigenvalues tend to be too high (low) and have to be pushed downwards (upwards).

In chapter 3, we propose a nonparametric estimation method which 'cross fertilize' the nonlinear shrinkage estimation of the covariance matrix onto the large-dimensional ICV matrix estimation problem. The properties of proposed methods are studied and numerical examples are given to demonstrate its usefulness.

The contribution of my part is that following the idea of Dr Lam, I explicitly give the estimator and complete the proofs with Dr Lam together, the simulation and numerical studies were also my work.

Chapter 2

Autocorrelation-based factor analysis

2.1 Relevant methods in the literature

The study of multivariate time series data becomes more and more important in many different fields, including macroeconomics, finance, and environment studies. In practice, due to the number of parameters needed to estimate is often too many, the method is rarely used without proper dimension reduction or regularization.

Factor modeling is one of the main techniques used in order to achieve dimension reduction for multivariate time series. The goal is to try to find factors which drive the dynamics of the time series. To this end, Lam, Yao and Bathia [12] propose to decompose a times series into two parts: a dynamic part which we expect to be driven by a lower dimension factor time series, and a static part which is a white noise vector. The estimation of the factor loading matrix and subsequently the factors are done via an eigenanalysis on a non-negative definite matrix constructed from autocorrelation matrix. We refer to this method as Autocorrelation-based Factor Analysis (AFA).

The vast majority of existing literature deal with factor models using a different decomposition y_t than AFA method does. Most of those factor models decompose differently in a way that they try to identify the common factors that affect the dynamics of most of the p components in the original time series and separate the so-called idiosyncratic noise components, in which each of them may affect at most a few original time series, from the common factors. This decomposition has its own merits in econometrics and finance applications. An important example utilizing this decomposition is by Bai and Ng [5] which was the method of PCA to estimate the factors and factor loadings.

However, technical difficulties arise in both model identification and inference: such decomposition requires that the dimension of time series goes to infinity for the common factors and idiosyncratic noises to be identified. In contrast, the decomposition by Lam, Yao and Bathia [12] is exempted from this identifiability problem. AFA works in both high and low dimension circumstances.

In [12], there is no direct comparison of the convergence rates for AFA and PCA. In this chapter, we give an explicit comparison of the convergence rates between AFA with PCA. We show that AFA possesses the advantage over PCA when dealing with small dimension time series for both one step and two step estimations, while at large dimension, the performance is still comparable. At the end of this chapter, we also include a discussion paper to a journal article in section 2.8.

2.2 Models and estimation

2.2.1 The Models and assumptions

Suppose we want to analyze the linear dynamic structure of time series y_t, we may decompose y_t into two parts: a static part (a white noise), and a common component which is driven by a low-dimensional process. Therefore, for $t = 1, ..., n$, we may consider the model

$$y_t = \mathbf{A}x_t + \epsilon_t, \tag{2.2.1}$$

where y_t is the observed time series of dimension p, x_t is the unobserved factor time series of dimension r, and it is assumed to be weakly stationary with finite first two moments. Here we assume both means of y_t and x_t are removed. The matrix $\mathbf{A}$ denotes the *unkown* constant factor loading matrix of size $p \times r$. Here we assume the number of factors r is much smaller than p. Finally, ϵ_t is a $p \times 1$ white noise vector with mean zero and some covariance matrix Σ_ϵ.

By noting the fact that the RHS is unchanged if we replace the pair $(\mathbf{A}, x_t)$ by $(\mathbf{AH}, \mathbf{H}^{-1}x_t)$ for any invertible $r \times r$ matrix $\mathbf{H}$, we can always find an $\mathbf{A}$ such that the columns of $\mathbf{A} = \{a_1, \cdots, a_r\}$ are orthonormal, therefore, we may assume $\mathbf{A}'\mathbf{A} = \mathbf{I}_r$, where $\mathbf{I}_r$ denotes the $r \times r$ identity matrix and $\mathbf{A}'$ denotes the transpose of $\mathbf{A}$. Let $\mathbf{B}$ be a $p \times (p - r)$ matrix for which $(\mathbf{A}, \mathbf{B})$ forms $p \times p$ orthogonal matrix. Although the factor loading matrix is not uniquely defined, the *factor loading space* $\mathcal{M}(\mathbf{A})$, which is the $r-$dimensional linear space spanned by columns of $\mathbf{A}$ is uniquely defined. Also, it is sensible to assume that any white

noise linear combination of x_t are absorbed into ϵ_t, and the rank of $\mathbf{A}$ is r as otherwise we may express (2.2.1) equivalently in terms of lower-dimensional factors.

In our model, the only observable series is y_t. How well we can recover x_t from y_t thus depends on the factor strength reflected by the coefficients in loading matrix A. For example, in the case of $\mathbf{A} = 0$, y_t carries no information on x_t, Therefore, it is intuitive to characterize the 'strength' of a factor using the number of non-zero elements in a column of $\mathbf{A}$. Intuitively, pervasive factors are those factors that affect most part of the series, and local factors are those affect only part of the series. Now assume we have r_1 pervasive factors (the numbers of non-zero elements in the corresponding columns are of order $\asymp p$), and r_2 local factors with strength $\frac{p_i}{p}$. Then the model can be written as

$$y_t = \mathbf{A}x_t + \epsilon_t = \mathbf{A}_s x_{ts} + \mathbf{A}_w x_{tw} + \epsilon_t, \tag{2.2.2}$$

where $\mathbf{A}_s'\mathbf{A}_w = 0$. In addition, for $\mathbf{A}_w$, we assume it adopts a known factor structure group:

$$A_w = \begin{pmatrix} A_{w_1} & 0 & \cdots & 0 \\ 0 & A_{w_2} & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & A_{w_{r_2}}, \end{pmatrix}, \ \mathbf{A}_{w_i} \in \mathbb{R}^{p_i}. \tag{2.2.3}$$

Here is some more notations. Define $\mathbf{\Sigma}_s(k) = \mathrm{Cov}(x_{ts}, x_{(t-k),s})$, $\mathbf{\Sigma}_{w_j}(k) = \mathrm{Cov}(x_{t,w_j}, x_{(t-k),w_j})$ and $\mathbf{\Sigma}_{x\epsilon}(k) = \mathrm{Cov}(x_{t+k}, \epsilon_t)$. For $k = 0, 1, 2, \cdots, k_0$. Also put

$$\Sigma_y(k) = \mathrm{Cov}(y_t, y_{t-k}),\ \Sigma_x(k) = \mathrm{Cov}(x_t, x_{t-k}),$$
$$\Sigma_{x\epsilon}(k) = \mathrm{Cov}(x_t, \epsilon_{t-k}).$$

Now, we have some regularity conditions.

(C1) In model (2.2.1), no linear combination of x_t is white noise and $\Sigma_x(k)$ is of full rank for $k = 0, \cdots, k_0$, where $k_0 \geq 1$ is a positive integer. In addition, $\mathbf{A}'\mathbf{A} = \mathbf{I}_r$.

(C2) The covariance matrix $\mathrm{Cov}(\epsilon_t, x_{t-s}) = 0$ for all $s \geq 0$.

(C3) The observable series y_t is strictly stationary and Ψ-mixing with mixing coefficient $\Psi(\cdot)$ satisfying that $\sum_{t\geq 1} t\Psi(t)^{\frac{1}{2}} < \infty$, and $E(|y_t|^4) < \infty$ element-wise.

(C4) For pervasive factors, it holds that $\|\mathbf{\Sigma}_s(k)\| \asymp p \asymp \|\mathbf{\Sigma}_s(k)\|_{\min}$, and for a local factor w_j, $\|\mathbf{\Sigma}_{w_j}(k)\| \asymp p_j \asymp \|\mathbf{\Sigma}_{w_j}(k)\|_{\min}$. Here, p_j denotes the number of non-zero elements in a specific local factor w_j.

(C5) For $k = 0, 1, \cdots, k_0$, $\|\mathbf{\Sigma}_{x\epsilon}(k)\| = o(p_j)$.

2.2.2 Estimation for the factor loading space

For the purpose of estimation, we want to find an estimator $\widehat{\mathbf{A}}$ for the $p \times r$ factor loading matrix $\mathbf{A}$. Recall that we have $\mathbf{A}$ being orthonormal, then the factor process x_t can be estimated by $\widehat{\mathbf{A}}' y_t$ and the resulting residual is $(\mathbf{I}_p - \widehat{\mathbf{A}}\widehat{\mathbf{A}}')y_t$.

In [27], they propose a method to estimate the factor loading matrix by performing an eigenanalysis on a non-negative semi-definite matrix. The method is outlined here.

(C2) implies that

$$\Sigma_y(k) = \mathbf{A}\Sigma_x(k)\mathbf{A}' + \mathbf{A}\Sigma_{x\epsilon}(k), \ k \geq 1.$$

Define the $p \times p$ nonnegative definite matrix $\mathbf{M}$ by

$$\mathbf{M} = \sum_{k=1}^{k_0} \Sigma_y(k)\Sigma_y(k)', \text{ where } k_0 \text{ is a prescribed integer.}$$

This matrix is constructed to accumulate the information from different time lags. Since

$$\Sigma_y(k) = \mathbf{A}\Sigma_x(k)\mathbf{A}' + \mathbf{A}\Sigma_{x\epsilon}(k), \ k \geq 1,$$

we have MB = 0, i.e. the columns of B are the eigenvectors of M corresponding to zero-eigenvalues. We use this non-negative definite matrix to avoid the cancellation of the information from different lags. Therefore, the value of k is taken from 1 rather than 0. In practice, small lag would be enough, as the autocorrelation is often at its strongest at the small time lags, large k_0 would not make a significant effect on the estimation.

The factor loading space is then spanned by the eigenvectors of $\mathbf{M}$ corresponding to its nonzero eigenvalues. We take the r orthonormal eigenvectors corresponding to non-zero eigenvalues of $\mathbf{M}$ as the columns of $\mathbf{A}$.

Finally, $\widehat{\mathbf{A}}$ is found by performing an eigenanalysis on the sample version $\widehat{\mathbf{M}}$:

$$\widehat{\mathbf{M}} = \sum_{k=1}^{k_0} \widehat{\Sigma}_y(k)\widehat{\Sigma}_y(k)',$$

where $\widehat{\Sigma}_y(k)$ denotes the sample covariance matrix of y_t at lag k.

2.3 Theoretical properties

In our notation, we use $a \asymp b$ to denote $a = O_P(b)$ and $b = O_P(a)$, and for any matrix $\mathbf{G}$, $\|\mathbf{G}\|$ is the L_2 norm, $\|\mathbf{G}\|_F$ is the Frobenius norm and $\|\mathbf{G}\|_{\min}$ is the square root of the smallest non-zero eigenvalue of $\mathbf{G}\mathbf{G}'$.

In Bai and Ng (2002), they propose a method (PCA) to find the factor loading matrix by doing an eigenanalysis on the matrix $\Sigma_y = \text{Cov}(y_t, y_t)$. The following theorem gives the convergence rates for PCA and AFA respectively.

Theorem 1 *Let conditions (C1)-(C5) hold. Further if we assume* $\|\Sigma_\epsilon\| = O(p^\gamma)$, $\gamma \in (0,1)$, *and denote the PCA estimator for fact loading matrix by* $\widehat{\mathbf{A}}_{PCA}$, *it holds for PCA that:*

$$\|\widehat{\mathbf{A}}_{PCA} - \mathbf{A}\| = Op(\left(\frac{p}{\min_j p_j}\right)^{1/2} n^{-\frac{1}{2}} \|\mathbf{B}'\mathbf{\Sigma}_\epsilon \mathbf{B}\|^{\frac{1}{2}} (1 + \left(\frac{p}{\min_j p_j}\right)^{1/2} p^{\gamma/2}) + \min_j p_j^{-1} \|\mathbf{B}'\mathbf{\Sigma}_\epsilon \mathbf{A}\|),$$

and for AFA,

$$\|\widehat{\mathbf{A}} - \mathbf{A}\| = O_P(\left(\frac{p}{\min_j p_j}\right)^{1/2} n^{-\frac{1}{2}} \|\mathbf{B}'\mathbf{\Sigma}_\epsilon \mathbf{B}\|^{\frac{1}{2}} (1 + \left(\frac{p}{\min_j p_j}\right)^{1/2} p^{\gamma/2}).$$

From theorem 1, we can see the difference of those two rates is that PCA has an extra term $\min_j p_j^{-1} \|\mathbf{B}'\mathbf{\Sigma}_\epsilon \mathbf{A}\|$ than AFA. For large p_j, this term tends to zero, and PCA and AFA have the same rates asymptotically. However, when p_j is small, $\min p_j^{-1} \|\mathbf{B}'\mathbf{\Sigma}_\epsilon \mathbf{A}\|$ may be dominant and PCA may even be inconsistent. This manifests the advantage AFA has when we deal with groups with small dimension.

In Lam and Yao (2012), they introduce a two-step estimation procedure, which is superior to the one-step estimation. The method is outlined here.

In equation (2.2.2), we first obtain the estimator $\widehat{\mathbf{A}} \equiv (\widehat{\mathbf{A}}_s, \widehat{\mathbf{A}}_w)$ for the factor loading matrix $\mathbf{A} = (\mathbf{A}_s, \mathbf{A}_w)$, then the effects of pervasive factors can be removed from the data using

$$y_t^* = y_t - \widehat{\mathbf{A}}_s \widehat{\mathbf{A}}_s' y_t.$$

Next, we perform the same estimation for the new data $\{y_t^*\}$, and obtain the estimated factor loading matrix $\widetilde{\mathbf{A}}_w$ for the local factors. The final estimator is then

$$\widetilde{\mathbf{A}} = (\widehat{\mathbf{A}}_s, \widetilde{\mathbf{A}}_w).$$

To highlight the benefits of the two-step estimation procedure, we replace condition (C5) by a stronger condition:

(C5)' $\text{Cov}(x_t, \epsilon_s) = 0$ for any t, s.

The convergence rate for two step estimation is presented in Theorem 2 below.

Theorem 2 (Two step rates) *Let conditions (C1)-(C4) and (C5)' hold. Denote* $p_m = \min_j p_j$ *and* $p_M = \max_j p_j$. *Let* $n = O(p)$ *and if we do not know the structure of the factor loading matrix, then for PCA, we have*

$$\|\widetilde{\mathbf{A}}_{PCA} - \mathbf{A}\| = O_P(r_2^{\frac{1}{2}} p^{\frac{1}{2}} p_M^{\frac{1}{2}} p_m^{-1} n^{-\frac{1}{2}} + p_m^{-1} \|\mathbf{B}'\mathbf{\Sigma}_\epsilon \mathbf{A}_w\|),$$

and the counterpart rate for AFA is

$$\|\widetilde{\mathbf{A}} - \mathbf{A}\| = O_P(r_2^{\frac{1}{2}} p^{\frac{1}{2}} p_M^{\frac{1}{2}} p_m^{-1} n^{-\frac{1}{2}}).$$

If we know the local factor structure group, and denote $\mathbf{\Sigma}_\epsilon^{(i)}$ *as the covariance for* i*th group of noise, then for local factors* i*th group* $\mathbf{A}_w^{(i)}$,

$$\|\widetilde{\mathbf{A}}_{PCA}^{(i)} - \mathbf{A}_w^{(i)}\| = O_P(n^{-\frac{1}{2}} + p_i^{-1} \|\mathbf{B}'\mathbf{\Sigma}_\epsilon^{(i)} \mathbf{A}_w^{(i)}\|),$$

and

$$\|\widetilde{\mathbf{A}}^{(i)} - \mathbf{A}_w^{(i)}\| = O_P(n^{-\frac{1}{2}}).$$

2.4 Simulations

We conduct the following simulation examples to illustrate our method. The comparison with principal components method of Bai & Ng [**?**] is also reported.

We set $r = 11$ as the number of factors in model (2.2.2) , including one pervasive factors and 10 local factors. And the 100×11 factor loading matrix has the following structure: (hence the factor structure group is known as we assumed.)

$$\mathbf{A} = \begin{pmatrix} A_{s_1} & A_{w_1} & 0 & \cdots & 0 \\ A_{s_2} & 0 & A_{w_2} & 0 & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ A_{s_{10}} & 0 & \cdots & 0 & A_{w_{10}} \end{pmatrix} \quad A_{w_i}, A_{s_i} \in \mathbb{R}^{10} \tag{2.4.1}$$

Every non-zero entry in $\mathbf{A}$ comes from $U(0,1)$ distribution and the first column is orthogonal to every other column (This is done by QR decomposition, the local factors are automatically orthogonal as non-zero entries are not overlapping). The factors are defined by $x_t = 0.9x_{t-1} + \eta_t$, where is η_t are independent $N(0,1)$ random variables. The performance of estimator is measured by

$$d(\widehat{\mathbf{A}}) = \|(\mathbf{I} - \mathbf{A}\mathbf{A}')\widehat{\mathbf{A}}\|, \tag{2.4.2}$$

which measures the 'distance' between the estimated factor loading matrix and the true fact loading matrix and a smaller number means a better estimation of the factor loading matrix. We consider two scenarios for noise structure ϵ_t.

Scenario I

In scenario I, $\epsilon_t \sim N(0, \Sigma_1)$, where the (i, j)th element of Σ_1 is defined as

$$\sigma_{ij} = \frac{1}{2}\{(|i-j|+1)^{2H} - 2|i-j|^{2H} + (|i-j|-1)^{2H}\}, \tag{2.4.3}$$

and H is the Hurst parameter which takes value in $[0.5, 1]$. Larger H_p means stronger cross-correlations for the components in ϵ_t and larger H_n accounts for stronger autocorrelations for ϵ_t's.

Table 2.1 reports the results obtained for scenario I. The experiments is conducted with $n = 1000$ and 100 repetitions. AFA1 and PCA1 are the results obtained without the knowledge of data structure and AFA2 and PCA2 are the results obtained with known data structure.

Table 2.1: Scenario I

		Errors/s.d.			
H_p	H_n	AFA1	PCA1	AFA2	PCA2
0.5	0.5	$26_{(22)}$	$22_{(15)}$	$27_{(60)}$	$21_{(47)}$
0.7	0.5	$26_{(22)}$	$41_{(61)}$	$26_{(70)}$	$20_{(50)}$
0.9	0.5	$26_{(36)}$	$86_{(64)}$	$26_{(69}$	$29_{(30)}$
0.5	0.7	$42_{(28)}$	$32_{(17)}$	$27_{(58)}$	$21_{(41)}$
0.7	0.7	$43_{(37)}$	$47_{(62)}$	$27_{(65)}$	$22_{(44)}$
0.9	0.7	$46_{(61)}$	$86_{(70)}$	$26_{(57)}$	$32_{(35)}$
0.5	0.9	$89_{(91)}$	$71_{(123)}$	$29_{(58)}$	$22_{(41)}$
0.7	0.9	$88_{(89)}$	$72_{(124)}$	$31_{(61)}$	$23_{(45)}$
0.9	0.9	$81_{(115)}$	$83_{(91)}$	$29_{(65)}$	$27_{(49)}$

Table: Means (true value multiplied by 100)

and standard deviation (in brackets, true value multiplied by 1000)

The results show that when both serial and cross-correlations are not too strong, our methods perform similar to PCA, with cross-correlations are strong, AFA has better accuracy in estimating the factor loading space. However, When the serial correlation is strong as well, PCA outperforms our methods.

Scenario II

In scenario II, $\epsilon_t \sim N(0, \Sigma_2)$, where $\Sigma_2 = \begin{pmatrix} \mathbf{H} & 0 & \cdots & 0 \\ 0 & \mathbf{H} & 0 & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & ... & 0 & \mathbf{H}, \end{pmatrix}$, and each $\mathbf{H}$ is a 10×10 diagonal matrix. In this case, the white noise ϵ_t adopts a *heteroscedastic and diagonal* covariance structure.

To obtain ϵ_t in scenario II, we first generate $p \times n$ matrix $\mathbf{Z}$ with entries being independently identically distributed standard normal random variable. And let $\mathbf{H} = diag\{\sigma_1^2, \sigma_1^2, \sigma_2^2, \sigma_2^2, \sigma_2^2, \sigma_3^2, \sigma_3^2, \sigma_3^2, \sigma_3^2, \sigma_3^2\}$. Therefore $\Sigma_2^{\frac{1}{2}}\mathbf{Z}$ would give the required noise structure for scenario II.

The performance is reported for both AFA and PCA, AFA1 and PCA1 are done without the knowledge of factor structure group, AFA2 and PCA2 are the results obtained with a known factor structure group. For each of the following cases, we replicate the simulation 100 times.

Firstly, we conduct experiment for homogeneous noise, i.e., Σ_2 has all diagonal entries equal one.

Table 2.2: Scenario II Homogeneous noise

	Error/s.d.			
n	AFA1	PCA1	AFA2	PCA2
200	$91_{(79)}$	$74_{(92)}$	$52_{(115)}$	$42_{(95)}$
500	$48_{(64)}$	$36_{(30)}$	$35_{(78)}$	$28_{(64)}$
1000	$26_{(21)}$	$21_{(17)}$	$26_{(70)}$	$20_{(58)}$

Table: Means (true value multiplied by 100)

and standard deviation (in brackets, true value multiplied by 1000)

From Table 2.2, we can see that by knowing the factor structure group, both methods perform better. And PCA outperforms our method in general. This is consistent with our asymptotic results.

However, the situation changes once we introduce heteroscedastic noise.

Table 2.3: Scenario II Heteroscedastic noise

	Error/s.d.			
$\sigma_1/\sigma_2/\sigma_3$	AFA1	PCA1	AFA2	PCA2
1/1/1	$26_{(22)}$	$22_{(16)}$	$26_{(68)}$	$19_{(52)}$
1/1.2/1.3	$32_{(32)}$	$30_{(23)}$	$25_{(59)}$	$20_{(43)}$
1/1.3/1.5	$35_{(29)}$	$37_{(29)}$	$26_{(75)}$	$23_{(46)}$
1/1.4/1.8	$43_{(50)}$	$53_{(53)}$	$28_{(70)}$	$29_{(42)}$
1/1.5/2	$47_{(49)}$	$62_{(60)}$	$27_{(61)}$	$34_{(35)}$

Table: Means (true value multiplied by 100)
and standard deviation (in brackets, true value multiplied by 1000)

From Table 2.3, The performance of our method catches up with PCA as the noise gradually becomes larger and more heteroscedastic. In addition, AFA2 does not change as we increase the variance of the noise. This is consistent with our theoretical results: when we estimate with the knowledge of factor loading matrix, the convergent rate is only dependent on n for our estimator, which is robust against the change in noise structure. On the other hand, the performance with PCA would deteriorate if the noise becomes more heteroscedastic.

To further demonstrate the performances for those two methods, we choose $\{\sigma_1, \sigma_2, \sigma_3\} = \{1, 1.4, 1.8\}$ as the parameters in $\mathbf{H}$ and combine with the Hurst noise matrix with $H_p = 0.8$.

Table 2.4: Scenario II Heteroscedastic noise 2

	Error			
n	AFA1	PCA1	AFA2	PCA2
100	99	99	77	91
200	99	99	55	87
300	99	99	48	81
400	99	99	40	67
500	98	99	37	66
1000	60	99	28	52
2000	28	98	20	49
5000	16	99	13	37

Table: Means (true value multiplied by 100)

From Table 2.4 , it is clear that when noise becomes more heteroscedastic, PCA simply

does not work in estimating the factor loading. However, our method can give a sensible estimation. As we increase the sample size, the estimation becomes very good.

2.5 Real data example

We compare AFA and PCA methods using a macroeconomic data set obtained from Stock and Watson (2005). The data consists of monthly data from January 1959 to December 2003. There are 132 U.S. macroeconomic time series in total, and they are categorized into 14 categories (numbers in the brackets denote the number of variables in each category): personal income (2); consumption (1); real retail, manufacturing and trade sales (2); industrial production indices (15); employment and hours (30); housing starts and sales (10); orders and real inventories (10); money and credit quantity aggregates (11); stock prices (4); interest rates and spreads (17); exchange rates (5); price indices (21); average hourly earnings (3); and miscellaneous (1).

Amongst those 14 categories, two of them have over 20 variables, five have 10 to 17 variables, another five categories have 2 to 5 variables, and the rest have only 1 variable each. Therefore, the data has a natural grouping, and if on top of some pervasive factors, there are some categories specific factors, then forecast may not be done well as the size of the categories are small compared to sample size.

Effectively, we are working with y_t of dimension 132, and $t = 1, 2, \cdots, 526$. We perform the factor modeling on each of 36 rolling windows with length of 490 each. We applied AFA and PCA to the data for each window. We use an autoregressive model of order 3 to forecast the $(i + 490)$th value of the estimated factor series $x^{(1)}_{i+490}$, so as to obtain a one-step ahead forecast $\widehat{y}^{(1)}_{i+490} = \widehat{\mathbf{A}} x_{i+490}$ for y_{i+490}. For comparison, we calculated the forecast error for the $(i + 490)$th month for each method, defined by

$$\text{Forecast error} = p^{-\frac{1}{2}} \|\widehat{y}^{(1)}_{i+490} - y_{i+490}\|. \tag{2.5.1}$$

Firstly, we need to estimate the number of factors. We plot the average forecast error for different number of factors r in Figure 2.1. We estimate the total number of factors to be around 20, and the number of more pervasive factors to be 3.

Figure 2.2 shows that the cumulative forecast errors obtained without assuming specific data structure are roughly the same for both PCA and AFA. Since we have a natural grouping structure, this suggests us to first estimate the three pervasive factors, and we estimated the

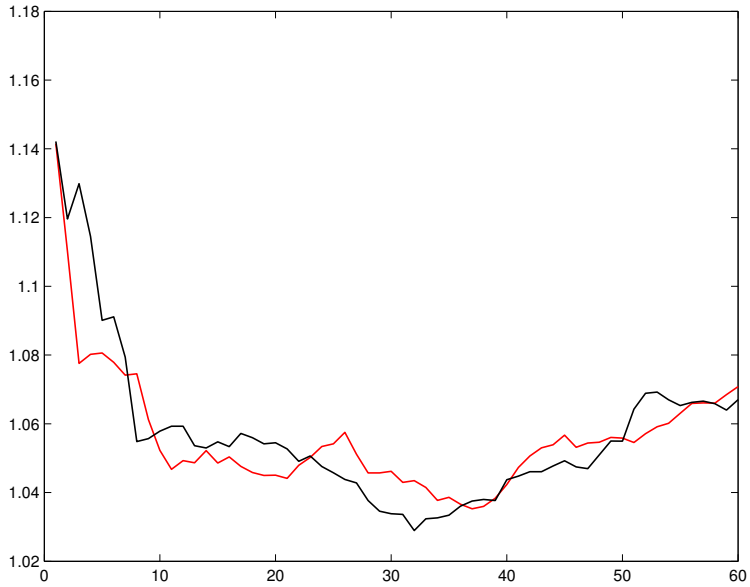

Figure 2.1: red line for AFA, black line for PCA.

local factors for categories with more than 4 variables. The cumulative forecast errors for estimation with known factor structure groupare plotted in Figure 2.3.

From the graph, we can see that the performance of AFA is better than PCA, this is expected as when estimating the local factors, PCA is not as good as AFA as suggested by Theorem 2.

2.6 Summary of this chapter

In this chapter, we explore the factor modeling for high-dimensional time series based on a dimension-reduction viewpoint. we allow the dimension of time series to be as large as, or even larger than the sample size of the time series. The estimation of the factor loading matrix and subsequently the factors are done via an eigenanalysis on a non-negative definite matrix constructed from autocorrelation matrix. Under the condition of PCA, We give explicit comparison of the convergence rates between AFA with PCA. We show that AFA possesses the advantage over PCA when dealing with small dimension time series for both one step and two step estimations, while at large dimension, the performance is still comparable. We also demonstrate in numerical examples and real data from macro economic data, that our method would have a better performance when the noise level is high.

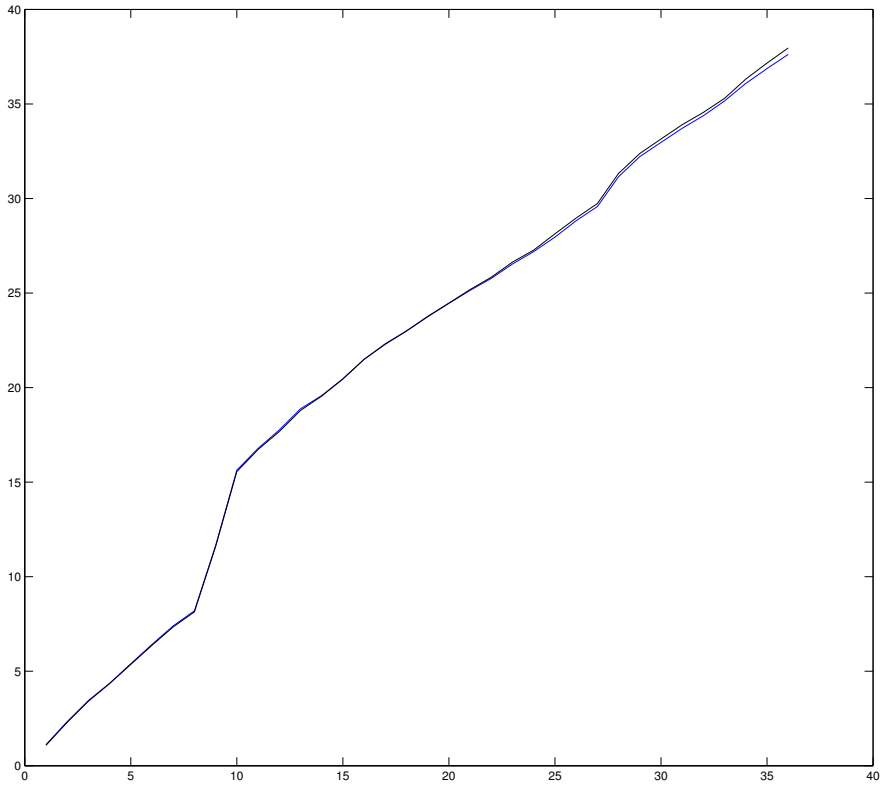

Figure 2.2: cumulative forecast error for estimation without assuming a known structure, blue line for AFA, black line for PCA

2.7 Proofs for this chapter

The idea of proofs for both theorem 1 and 2 is based on the following lemma which is Theorem 8.1.10 of Golub and Van Loan (2013).

Lemma 1 *Suppose A and $A+E$ are n-by-n symmetric matrices and that $Q = [Q_1\ Q_2]$ is an orthogonal matrix such that span(Q_1) is an invariant subspace for A. Here Q_1 has size $n \times r$ and Q_2 has size $n \times (n-r)$. Partition the matrices $Q'AQ$ and $Q'EQ$ as follows:*

$$Q'AQ = \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix}, \; Q'EQ = \begin{pmatrix} E_{11} & E_{21}' \\ E_{21} & E_{22} \end{pmatrix}.$$

If $sep(D_1, D_2) = \min_{\lambda_1 \in \lambda(D_1), \lambda_2 \in \lambda(D_2)} |\lambda_1 - \lambda_2| > 0$, where $\lambda(M)$ denotes the set of eigenvalues of the matrix M, and $\|E\| \leq sep(D_1, D_2)/5$, then there exists a matrix $P \in \mathbb{R}^{(n-r)\times r}$ with

$$\|P\| \leq \frac{4}{sep(D_1, D_2)} \|E_{21}\|.$$

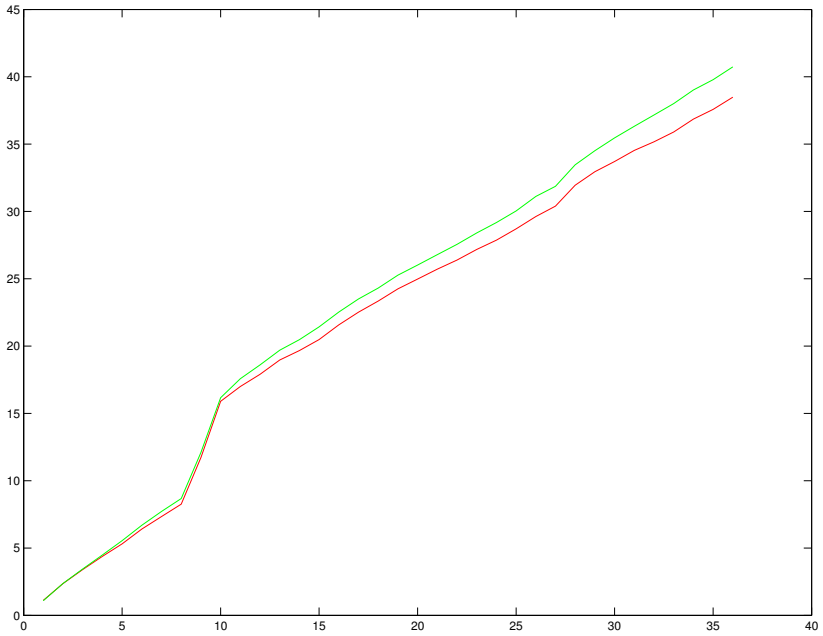

Figure 2.3: cumulative forecast error for estimation using known structure, red line for AFA, green line for PCA.

such that the columns of $\widehat{Q}_1 = (Q_1 + Q_2 P)(I + P'P)^{-1/2}$ *define an orthonormal basis for a subspace that is invariant for* $A + E$.

Proof of Theorem 1. For PCA, we do eigenanalysis on

$$\mathbf{\Sigma}_y = \mathbf{A}\mathbf{\Sigma}_x\mathbf{A}' + \mathbf{\Sigma}_\epsilon,$$

Its corresponding sample version is

$$\begin{aligned} \widehat{\mathbf{\Sigma}}_y &= n^{-1}\mathbf{\Sigma}_{t=1}^{n} y_t y_t' \\ &= \mathbf{A}\widehat{\mathbf{\Sigma}}_x\mathbf{A} + \mathbf{A}\widehat{\mathbf{\Sigma}}_{x\epsilon} + \widehat{\mathbf{\Sigma}}_{\epsilon x}\mathbf{A}' + \widehat{\mathbf{\Sigma}}_\epsilon. \end{aligned}$$

Define $\widehat{L} = \mathbf{A}\mathbf{\Sigma}_x\mathbf{A}' + (\widehat{\mathbf{\Sigma}}_y - \mathbf{A}\mathbf{\Sigma}_x\mathbf{A}')$, where we refer to the first term as L, and the second term as E. It follows that

$$\begin{pmatrix} \mathbf{A}' \\ \mathbf{B}' \end{pmatrix} L(\mathbf{A}\ \mathbf{B}) = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}.$$

Hence,

$$\text{sep}(D, 0) = \lambda_{\min}(D)$$

$$\begin{aligned} &= \lambda_{\min}(\mathbf{\Sigma}_x) \\ &\asymp p_j. \end{aligned}$$

Now we need to find $\|E_{21}\|$, where E_{21} is defined through

$$Q'EQ = \begin{pmatrix} E_{11} & E'_{21} \\ E_{21} & E_{22} \end{pmatrix}.$$

In our case,

$$Q'EQ = \begin{pmatrix} \mathbf{A}' \\ \mathbf{B}' \end{pmatrix} (\widehat{\mathbf{\Sigma}}_y - \mathbf{A}\mathbf{\Sigma}_x\mathbf{A}')(\mathbf{A}\ \mathbf{B}),$$

So we have $E_{21} = \mathbf{B}'\widehat{\mathbf{\Sigma}}_{\epsilon x} + \mathbf{B}'\widehat{\mathbf{\Sigma}}_{\epsilon}\mathbf{A}$.

Let $x_t = \mathbf{\Sigma}_x^{\frac{1}{2}} z_t$ and $\epsilon_t = \mathbf{\Sigma}_\epsilon^{\frac{1}{2}} u_t$, where z_t and u_t are vectors consist of i.i.d. random variables with mean 0 and variance 1. Also denote $\epsilon = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)$ and $u = (u_1, u_2, ..., u_n)$.

Therefore,

$$\|E_{21}\| \leq \|\mathbf{B}'\widehat{\mathbf{\Sigma}}_{\epsilon x}\| + \|\mathbf{B}'\mathbf{\Sigma}_\epsilon\mathbf{A}\| + \|\mathbf{B}'(n^{-1}\epsilon\epsilon' - \mathbf{\Sigma}_\epsilon)\mathbf{A}\|.$$

$$\begin{aligned} \mathbf{B}'\widehat{\mathbf{\Sigma}}_{\epsilon x} &= \mathbf{B}' n^{-1}\epsilon_t x_t' \\ &= \mathbf{B}'\mathbf{\Sigma}_\epsilon^{\frac{1}{2}} n^{-1} u_t z_t' \mathbf{\Sigma}_x^{\frac{1}{2}}. \end{aligned}$$

This implies,

$$\|\mathbf{B}'\widehat{\mathbf{\Sigma}}_{\epsilon x}\| \leq \|\mathbf{B}'\mathbf{\Sigma}_\epsilon^{\frac{1}{2}}\| \cdot \|\widehat{\mathbf{\Sigma}}_{uz}\| \cdot \|\mathbf{\Sigma}_x^{\frac{1}{2}}\|.$$

Using $\|\widehat{\mathbf{\Sigma}}_{uz}\| = O_P(p^{\frac{1}{2}} n^{-\frac{1}{2}})$, and $\|\mathbf{\Sigma}_x^{\frac{1}{2}}\| = O_P(p_j^{\frac{1}{2}})$, we have

$$\|\mathbf{B}'\widehat{\mathbf{\Sigma}}_{\epsilon x}\| = O_P(\|\mathbf{B}'\mathbf{\Sigma}_\epsilon\mathbf{B}\|^{\frac{1}{2}} \cdot p_j^{\frac{1}{2}} \cdot p^{\frac{1}{2}} n^{-\frac{1}{2}})$$

Then for the last term,

$$\begin{aligned} \mathbf{B}'(n^{-1}\epsilon\epsilon' - \mathbf{\Sigma}_\epsilon)\mathbf{A} &= \mathbf{B}'(\mathbf{\Sigma}_\epsilon^{\frac{1}{2}} n^{-1} uu' \mathbf{\Sigma}_\epsilon^{\frac{1}{2}} - \mathbf{\Sigma}_\epsilon^{\frac{1}{2}}\mathbf{\Sigma}_\epsilon^{\frac{1}{2}})\mathbf{A} \\ &= \mathbf{B}'\mathbf{\Sigma}_\epsilon^{\frac{1}{2}}(n^{-1}uu' - I)\mathbf{\Sigma}_\epsilon^{\frac{1}{2}}\mathbf{A}. \end{aligned}$$

To analyse the expression on the right hand side term by term, $\|n^{-1}uu' - I\| = O_P(pn^{-\frac{1}{2}})$, $\|\mathbf{\Sigma}_\epsilon^{\frac{1}{2}}\| = O_P(p^{\gamma/2})$ and $\|\mathbf{A}\| = 1$.

Put everything together, we have

$$\|\mathbf{B}'(n^{-1}\epsilon\epsilon' - \mathbf{\Sigma}_\epsilon)\mathbf{A}\| \leq \|\mathbf{B}'\mathbf{\Sigma}_\epsilon\mathbf{B}\|^{\frac{1}{2}} O_P(pn^{-\frac{1}{2}}) O_P(p^{\gamma/2})$$

$$= O_P(\|\mathbf{B}'\boldsymbol{\Sigma}_\epsilon\mathbf{B}\|^{\frac{1}{2}}pn^{-\frac{1}{2}}p^{\gamma/2}).$$

This leads to

$$\begin{aligned}\|E_{21}\| &= O_P(\|\mathbf{B}'\boldsymbol{\Sigma}_\epsilon\mathbf{B}\|^{\frac{1}{2}}p_j^{\frac{1}{2}}\cdot p^{\frac{1}{2}}n^{-\frac{1}{2}} + \|\mathbf{B}'\boldsymbol{\Sigma}_\epsilon\mathbf{B}\|^{\frac{1}{2}}pn^{-\frac{1}{2}}p^{\gamma/2} + \|\mathbf{B}'\boldsymbol{\Sigma}_\epsilon\mathbf{A})\|\\ &= O_P((pp_j)^{\frac{1}{2}}n^{-\frac{1}{2}}\|\mathbf{B}'\boldsymbol{\Sigma}_\epsilon\mathbf{B}\|^{\frac{1}{2}}((\frac{p}{p_j})^{\frac{1}{2}}p^{\gamma/2}+1) + \|\mathbf{B}'\boldsymbol{\Sigma}_\epsilon\mathbf{A})\|).\end{aligned}$$

Finally,

$$\begin{aligned}\|\widehat{\mathbf{A}}_{\text{PCA}} - \mathbf{A}\| &= \frac{\|E_{21}\|}{\text{sep}(D,0)}\\ &= O_P((pp_j)^{\frac{1}{2}}n^{-\frac{1}{2}}\|\mathbf{B}'\boldsymbol{\Sigma}_\epsilon\mathbf{B}\|^{\frac{1}{2}}((\frac{p}{p_j})^{\frac{1}{2}}p^{\gamma/2}+1) + \|\mathbf{B}'\boldsymbol{\Sigma}_\epsilon\mathbf{A})\|)/p_j\\ &= O_P(\frac{p}{p_j})^{\frac{1}{2}}n^{-\frac{1}{2}}\|\mathbf{B}'\boldsymbol{\Sigma}_\epsilon\mathbf{B}\|^{\frac{1}{2}}(1+(\frac{p}{p_j})^{\frac{1}{2}}p^{\gamma/2}) + p_j^{-1}\|\mathbf{B}'\boldsymbol{\Sigma}_\epsilon\mathbf{A})\|).\end{aligned}$$

This proves the PCA rate in Theorem 1.

Now, for AFA, define

$$\begin{aligned}L &= \sum_{k=1}^{k_0}\boldsymbol{\Sigma}_y(k)\boldsymbol{\Sigma}_y(k)'\\ &= \mathbf{A}\left(\sum_{k=1}^{k_0}(\Sigma_x(k)\mathbf{A}'+\Sigma_{x\epsilon}(k))(\Sigma_x(k)\mathbf{A}'+\Sigma_{x\epsilon}(k))'\right)\mathbf{A}'.\end{aligned}$$

and

$$\begin{aligned}\widehat{L} &= \sum_{k=1}^{k_0}\widehat{\boldsymbol{\Sigma}}_y(k)\widehat{\boldsymbol{\Sigma}}_y(k)'\\ &= \sum_{k=1}^{k_0}(\mathbf{A}\widehat{\boldsymbol{\Sigma}}_x(k)\mathbf{A}' + \mathbf{A}\widehat{\boldsymbol{\Sigma}}_{x\epsilon}(k) + \widehat{\boldsymbol{\Sigma}}_{\epsilon x}(k)\mathbf{A}' + \widehat{\boldsymbol{\Sigma}}_\epsilon(k))\\ &\quad (\mathbf{A}\widehat{\boldsymbol{\Sigma}}_x(k)\mathbf{A}' + \mathbf{A}\widehat{\boldsymbol{\Sigma}}_{x\epsilon}(k) + \widehat{\boldsymbol{\Sigma}}_{\epsilon x}(k)\mathbf{A}' + \widehat{\boldsymbol{\Sigma}}_\epsilon(k))'.\end{aligned}$$

Define $E = \widehat{L} - L$, and from proof of Theorem 1, Lam et al. (2011), by assuming $\kappa_{\max} = o(p_j)$, we have $\text{sep}(D,0) \asymp p_j^2$.

Since $\mathbf{B}'L = 0$ and noting that $\mathbf{B}'\mathbf{A} = 0$, we have

$$\begin{aligned}E_{21} &= \mathbf{B}'(\widehat{L} - L)\mathbf{A}\\ &= \mathbf{B}'\widehat{L}\mathbf{A}.\end{aligned}$$

Then,

$$\mathbf{B}'\sum_{k=1}^{k_0}(\mathbf{A}\widehat{\boldsymbol{\Sigma}}_x(k)\mathbf{A}' + \mathbf{A}\widehat{\boldsymbol{\Sigma}}_{x\epsilon}(k) + \widehat{\boldsymbol{\Sigma}}_{\epsilon x}(k)\mathbf{A}' + \widehat{\boldsymbol{\Sigma}}_\epsilon(k))$$

$$(\mathbf{A}\widehat{\mathbf{\Sigma}}_x(k)\mathbf{A}' + \mathbf{A}\widehat{\mathbf{\Sigma}}_{x\epsilon}(k) + \widehat{\mathbf{\Sigma}}_{\epsilon x}(k)\mathbf{A}' + \widehat{\mathbf{\Sigma}}_{\epsilon}(k))'\mathbf{A}$$
$$= \sum_{k=1}^{k_0}(\mathbf{B}'\widehat{\mathbf{\Sigma}}_{\epsilon x}(k)\mathbf{A}' + \mathbf{B}'\widehat{\mathbf{\Sigma}}_{\epsilon}(k))(\mathbf{A}\widehat{\mathbf{\Sigma}}_x(k)\mathbf{A}' + \mathbf{A}\widehat{\mathbf{\Sigma}}_{x\epsilon}(k) + \widehat{\mathbf{\Sigma}}_{\epsilon x}(k)\mathbf{A}' + \widehat{\mathbf{\Sigma}}_{\epsilon}(k))'\mathbf{A}.$$

We can write,

$$\begin{aligned}\mathbf{A}\widehat{\mathbf{\Sigma}}_x(k)\mathbf{A}' + \mathbf{A}\widehat{\mathbf{\Sigma}}_{x\epsilon}(k) + \widehat{\mathbf{\Sigma}}_{\epsilon x}(k)\mathbf{A}' + \widehat{\mathbf{\Sigma}}_{\epsilon}(k) &= \widehat{\mathbf{\Sigma}}_y(k) \\ &= \widehat{\mathbf{\Sigma}}_y(k) - \mathbf{\Sigma}_y(k) + \mathbf{\Sigma}_y(k)\end{aligned}$$

.

From proof of Theorem 1 and 2, Lam et al. (2011), if conditions 8 and 9 are satisfied, we have

$$\begin{aligned}\|\widehat{\mathbf{\Sigma}}_y(k) - \mathbf{\Sigma}_y(k)\| &= O_P((pp_j)^{\frac{1}{2}}n^{-\frac{1}{2}}) \\ \|\mathbf{\Sigma}_y(k)\| &= O(p_j).\end{aligned}$$

Therefore if we assume $(\frac{p}{p_j})^{\frac{1}{2}}n^{-\frac{1}{2}} = o(1)$, we have

$$\begin{aligned}\|\mathbf{A}\widehat{\mathbf{\Sigma}}_x(k)\mathbf{A}' + \mathbf{A}\widehat{\mathbf{\Sigma}}_{x\epsilon}(k) + \widehat{\mathbf{\Sigma}}_{\epsilon x}(k)\mathbf{A}' + \widehat{\mathbf{\Sigma}}_{\epsilon}(k)\| &\leq \|\widehat{\mathbf{\Sigma}}_y(k) - \mathbf{\Sigma}_y(k)\| + \|\mathbf{\Sigma}_y(k)\| \\ &= O_P((\frac{p}{p_j})^{\frac{1}{2}}n^{-\frac{1}{2}} + p_j) \\ &= O_P(p_j).\end{aligned}$$

Or if we do not assume $(\frac{p}{p_j})^{\frac{1}{2}}n^{-\frac{1}{2}} = o(1)$, then $\|\widehat{\mathbf{\Sigma}}_y(k) - \mathbf{\Sigma}_y(k)\| = O_P(pn^{-\frac{1}{2}})$, then it may be too restrictive for the inclusion of local factors, for instance, we need $\log(\frac{p_j}{p}) > \frac{1}{2}$ when $p \asymp n$ to get the same rate of $O_P(p_j)$.

What is left is to find the order of $\|\mathbf{B}'\widehat{\mathbf{\Sigma}}_{\epsilon x}(k)\mathbf{A}' + \mathbf{B}'\widehat{\mathbf{\Sigma}}_{\epsilon}(k)\|$.

Recall we have defined $\epsilon_t = \mathbf{\Sigma}_{\epsilon}^{\frac{1}{2}}u_t$ and $x_t = \mathbf{\Sigma}_{\epsilon}^{\frac{1}{2}}z_t$, so we have

$$\begin{aligned}\mathbf{B}'\widehat{\mathbf{\Sigma}}_{\epsilon x}(k)\mathbf{A}' &= \mathbf{B}'\frac{1}{n-k}\mathbf{\Sigma}_{\epsilon}^{\frac{1}{2}}u_t z'_{t-k}\mathbf{\Sigma}_x^{\frac{1}{2}}\mathbf{A}^T \\ &= \mathbf{B}'\mathbf{\Sigma}_{\epsilon}^{\frac{1}{2}}\widehat{\mathbf{\Sigma}}_{uz}(k)\mathbf{\Sigma}_x^{\frac{1}{2}}\mathbf{A}^T.\end{aligned}$$

This implies,

$$\begin{aligned}\|\mathbf{B}'\widehat{\mathbf{\Sigma}}_{\epsilon x}(k)\mathbf{A}'\| &\leq \|\mathbf{B}'\mathbf{\Sigma}_{\epsilon}\mathbf{B}\|^{\frac{1}{2}} \cdot \|\widehat{\mathbf{\Sigma}}_{uz}(k)\| \cdot \|\mathbf{\Sigma}_x^{\frac{1}{2}}\| \\ &= O_P(\|\mathbf{B}'\mathbf{\Sigma}_{\epsilon}\mathbf{B}\|^{\frac{1}{2}}p^{\frac{1}{2}}n^{-\frac{1}{2}}p_j^{\frac{1}{2}})\end{aligned}$$

$$= O_P(\|\mathbf{B}'\mathbf{\Sigma}_\epsilon\mathbf{B}\|^{\frac{1}{2}}(pp_j)^{\frac{1}{2}}n^{-\frac{1}{2}}).$$

And (recall that $\mathbf{\Sigma}_\epsilon(k) = 0$),

$$\begin{aligned}\|\mathbf{B}'\widehat{\mathbf{\Sigma}}_\epsilon(k)\| &= \|\mathbf{B}'(\mathbf{\Sigma}_\epsilon^{\frac{1}{2}}(n-k)^{-1}u_t u_{t-k}'\mathbf{\Sigma}_\epsilon^{\frac{1}{2}})\| \\ &\leq \|\mathbf{B}'\mathbf{\Sigma}_\epsilon\mathbf{B}\|^{\frac{1}{2}} \cdot \|\widehat{\mathbf{\Sigma}}_u(k)\| \cdot \|\mathbf{\Sigma}_\epsilon^{\frac{1}{2}}\| \\ &= O_P(\|\mathbf{B}'\mathbf{\Sigma}_\epsilon\mathbf{B}\|^{\frac{1}{2}}pn^{-\frac{1}{2}}p^{\frac{\gamma}{2}}).\end{aligned}$$

Therefore,

$$\begin{aligned}\|\mathbf{B}'\widehat{\mathbf{\Sigma}}_{\epsilon x}(k)\mathbf{A}' + \mathbf{B}'\widehat{\mathbf{\Sigma}}_\epsilon(k)\| &\leq \|\mathbf{B}'\widehat{\mathbf{\Sigma}}_{\epsilon x}(k)\mathbf{A}'\| + \|\mathbf{B}'\widehat{\mathbf{\Sigma}}_\epsilon(k)\| \\ &= O_P((pp_j)^{\frac{1}{2}}n^{-\frac{1}{2}}\|\mathbf{B}'\mathbf{\Sigma}_\epsilon\mathbf{B}\|^{\frac{1}{2}}(1 + (\frac{p}{p_j})^{\frac{1}{2}}p^{\frac{\gamma}{2}})).\end{aligned}$$

Finally, we have

$$\begin{aligned}\|\widehat{\mathbf{A}} - \mathbf{A}\| &= \frac{O_P((pp_j)^{\frac{1}{2}}n^{-\frac{1}{2}}\|\mathbf{B}'\mathbf{\Sigma}_\epsilon\mathbf{B}\|^{\frac{1}{2}}(1 + (\frac{p}{p_j})^{\frac{1}{2}}p^{\frac{\gamma}{2}}) \cdot p_j)}{\text{sep}(D, 0)} \\ &= \frac{O_P(p^{\frac{1}{2}}p_j^{\frac{3}{2}}n^{-\frac{1}{2}}\|\mathbf{B}'\mathbf{\Sigma}_\epsilon\mathbf{B}\|^{\frac{1}{2}}(1 + (\frac{p}{p_j})^{\frac{1}{2}}p^{\frac{\gamma}{2}}))}{p_j^2} \\ &= O_P((\frac{p}{p_j})^{\frac{1}{2}}n^{\frac{1}{2}}\|\mathbf{B}'\mathbf{\Sigma}_\epsilon\mathbf{B}\|^{\frac{1}{2}}(1 + (\frac{p}{p_j})^{\frac{1}{2}}p^{\frac{\gamma}{2}}).\end{aligned}$$ □

Proof of Theorem 2. For PCA, we perform eigen-analysis on

$$\begin{aligned}\widehat{\mathbf{\Sigma}}_y =& n^{-1}\Sigma_{t=1}^n y_t y_t' \\ =& \mathbf{A}_s\widehat{\mathbf{\Sigma}}_s\mathbf{A}_s' + \mathbf{A}_s\widehat{\mathbf{\Sigma}}_{sw}\mathbf{A}_w' + \mathbf{A}_s\widehat{\mathbf{\Sigma}}_{s\epsilon} \\ &+ \mathbf{A}_w\widehat{\mathbf{\Sigma}}_{ws}\mathbf{A}_s' + \mathbf{A}_w\widehat{\mathbf{\Sigma}}_w\mathbf{A}_w' + \mathbf{A}_w\widehat{\mathbf{\Sigma}}_{w\epsilon} \\ &+ \widehat{\mathbf{\Sigma}}_{\epsilon s}\mathbf{A}_s' + \widehat{\mathbf{\Sigma}}_{\epsilon w}\mathbf{A}_w' + \widehat{\mathbf{\Sigma}}_\epsilon.\end{aligned}$$

Now, since factors are uncorrelated with noise, and pervasive factors are uncorrelated with local factors, we have

$$\mathbf{\Sigma}_y = \mathbf{A}_s\mathbf{\Sigma}_s\mathbf{A}_s' + \mathbf{A}_w\mathbf{\Sigma}_w\mathbf{A}_w' + \mathbf{\Sigma}_\epsilon.$$

We write $\widehat{L} = \widehat{\mathbf{\Sigma}}_y = \mathbf{A}_s\mathbf{\Sigma}_s\mathbf{A}_s' + \mathbf{A}_w\mathbf{\Sigma}_w\mathbf{A}_w' + (\widehat{\mathbf{\Sigma}}_y - \mathbf{A}_s\mathbf{\Sigma}_s\mathbf{A}_s' - \mathbf{A}_w\mathbf{\Sigma}_w\mathbf{A}_w')$.

Firstly, we need to find $\text{sep}(D_s, D_w)$. We first do this pervasive factors, then D_s is of size $r_1 \times r_1$.

Clearly, $\mathbf{A}_s$ and $\mathbf{A}_w$ contain r_1+r_2 eigenvectors of L. Let $\mathbf{B}$ be the orthogonal complement of $(\mathbf{A}_s\ \mathbf{A}_w)$, that is $\mathbf{B}'\mathbf{B} = I_{p-r_1-r_2}$, and $\mathbf{B}'\mathbf{A}_s = \mathbf{B}'\mathbf{A}_w = 0$. Then

$$\begin{pmatrix} \mathbf{A}_s' \\ \mathbf{A}_w' \\ \mathbf{B}' \end{pmatrix} L(\mathbf{A}_s\ \mathbf{A}_w\ \mathbf{B}) = \begin{pmatrix} D_s & 0 & 0 \\ 0 & D_w & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

where $\mathbf{A}_s' L \mathbf{A}_s = D_s$. To find the order of $\|\widehat{\mathbf{A}}_s - \mathbf{A}_s\|$, we need to know $\text{sep}\left(D_s, \begin{pmatrix} D_w & 0 \\ 0 & 0 \end{pmatrix}\right)$.

$$\begin{aligned} \text{sep}\left(D_s, \begin{pmatrix} D_w & 0 \\ 0 & 0 \end{pmatrix}\right) =& \lambda_{\min}(D_s) - \lambda_{\max}(D_w) \\ =& \lambda_{\min}(\mathbf{\Sigma}_s) - \lambda_{\max}(\mathbf{\Sigma}_w) \\ \asymp& p - \max_{1\le j\le r_2} p_j \\ \asymp& p. \end{aligned}$$

$$\begin{aligned} Q'EQ &= \begin{pmatrix} \mathbf{A}_s' \\ \mathbf{A}_w' \\ \mathbf{B}' \end{pmatrix} (\widehat{\mathbf{\Sigma}}_y - \mathbf{A}_s\mathbf{\Sigma}_s\mathbf{A}_s' - \mathbf{A}_w\mathbf{\Sigma}_w\mathbf{A}_w')(\mathbf{A}_s\ \mathbf{A}_w\ \mathbf{B}) \\ &= \begin{pmatrix} E_{11} & E_{21}' \\ E_{21} & E_{22} \end{pmatrix}. \end{aligned}$$

In our case, $Q = [Q_1\ Q_2]$, where we take $Q_1 = \mathbf{A}_s$ and $Q_2 = [\mathbf{A}_w\ \mathbf{A}_s]$,

$$E_{21} = (\begin{pmatrix} \mathbf{A}' \\ \mathbf{B}' \end{pmatrix} (\widehat{\mathbf{\Sigma}}_y - \mathbf{A}_s\mathbf{\Sigma}_s\mathbf{A}_s' - \mathbf{A}_w\mathbf{\Sigma}_w\mathbf{A}_w')\mathbf{A}_s = \begin{pmatrix} \mathbf{A}_w'\widehat{\mathbf{\Sigma}}_y\mathbf{A}_s \\ \mathbf{B}'\widehat{\mathbf{\Sigma}}_y\mathbf{A}_s \end{pmatrix}.$$

Let $x_t = \mathbf{\Sigma}_x^{\frac{1}{2}} z_t$ and $\epsilon_t = \mathbf{\Sigma}_\epsilon^{\frac{1}{2}} u_t$, where x_t and ϵ_t are vectors with $\text{Var}(x_t) = \mathbf{I}$ and $\text{Var}(\epsilon_t) = \mathbf{I}$ and $\mathbf{\Sigma}_x^{\frac{1}{2}}$ and $\mathbf{\Sigma}_\epsilon^{\frac{1}{2}}$ are square roots of matrices $\mathbf{\Sigma}_x$ and $\mathbf{\Sigma}_\epsilon$, Then

$$\begin{aligned} \|\mathbf{A}_w'\widehat{\mathbf{\Sigma}}_y\mathbf{A}_s\| =& \|\widehat{\mathbf{\Sigma}}_{ws} + \widehat{\mathbf{\Sigma}}_{w\epsilon}\mathbf{A}_s + \mathbf{A}_w'\widehat{\mathbf{\Sigma}}\epsilon s + \mathbf{A}_w'\widehat{\mathbf{\Sigma}}\epsilon\mathbf{A}_s\| \\ \le& \|\mathbf{\Sigma}_w^{\frac{1}{2}}(n^{-1}z_w z_s')\mathbf{\Sigma}_s^{\frac{1}{2}}\| + \|\mathbf{\Sigma}_w^{\frac{1}{2}}(n^{-1}z_w u')\mathbf{\Sigma}_\epsilon^{\frac{1}{2}}\mathbf{A}_s\| \\ &+ \|\mathbf{A}_w'\mathbf{\Sigma}_\epsilon^{\frac{1}{2}}(n^{-1}u z_s')\mathbf{\Sigma}_s^{\frac{1}{2}}\| + \|\mathbf{A}_w'\mathbf{\Sigma}_\epsilon^{\frac{1}{2}}(n^{-1}uu')\mathbf{\Sigma}_\epsilon^{\frac{1}{2}}\mathbf{A}_s\|. \end{aligned}$$

Using the facts that pervasive factors are uncorrelated with local factors and factors are uncorrelated with noise and $\|\mathbf{\Sigma}_\epsilon\| = O(p^\gamma)$, after some algebra,

$$\|\mathbf{\Sigma}_w^{\frac{1}{2}}\mathbf{\Sigma}_{z_w z_s}\mathbf{\Sigma}_s^{\frac{1}{2}}\| + \|\mathbf{\Sigma}_w^{\frac{1}{2}}(n^{-1}z_w z_s' - \mathbf{\Sigma}_{z_w z_s})\mathbf{\Sigma}_s^{\frac{1}{2}}\| = O_P(\max_j p_j^{\frac{1}{2}} \cdot p^{\frac{1}{2}} \cdot n^{-\frac{1}{2}}),$$

$$\|\mathbf{\Sigma}_w^{\frac{1}{2}}\mathbf{\Sigma}_{z_w\epsilon}\mathbf{\Sigma}_\epsilon^{\frac{1}{2}}\mathbf{A}_s\| + \|\mathbf{\Sigma}_w^{\frac{1}{2}}(n^{-1}z_w u' - \mathbf{\Sigma}_{z_w\epsilon})\mathbf{\Sigma}_\epsilon^{\frac{1}{2}}\mathbf{A}_s\| = O_P(\max_j p_j^{\frac{1}{2}} \cdot p^{\frac{\gamma}{2}} \cdot p^{\frac{1}{2}}n^{-\frac{1}{2}}),$$
$$\|\mathbf{A}_w'\mathbf{\Sigma}_\epsilon^{\frac{1}{2}}\mathbf{\Sigma}_{\epsilon z_s}\mathbf{\Sigma}_s^{\frac{1}{2}}\| + \|\mathbf{A}_w'\mathbf{\Sigma}_\epsilon^{\frac{1}{2}}(n^{-1}uz_s' - \mathbf{\Sigma}_{\epsilon z_s})\mathbf{\Sigma}_s^{\frac{1}{2}}\| = O_P(p^{\frac{\gamma}{2}} \cdot p^{\frac{1}{2}} \cdot p^{\frac{1}{2}}n^{-\frac{1}{2}}),$$
$$\|\mathbf{A}_w'\mathbf{\Sigma}_\epsilon\mathbf{A}_s\| + \|\mathbf{A}_w'\mathbf{\Sigma}_\epsilon^{\frac{1}{2}}(n^{-1}uu' - \mathbf{I})\mathbf{\Sigma}_\epsilon^{\frac{1}{2}}\mathbf{A}_s\| = O_P(p^\gamma + pn^{-\frac{1}{2}}).$$

This gives

$$\|\mathbf{A}_w'\widehat{\mathbf{\Sigma}}_y\mathbf{A}_s\| = O_P(p^{\frac{\gamma}{2}}pn^{-\frac{1}{2}} + p^\gamma + pn^{-\frac{1}{2}}).$$

$$\begin{aligned}
\|\mathbf{B}'\widehat{\mathbf{\Sigma}}_y\mathbf{A}_s\| =& \|\mathbf{B}'\widehat{\mathbf{\Sigma}}_{\epsilon s} + \mathbf{B}'\widehat{\mathbf{\Sigma}}_\epsilon\mathbf{A}_s\| \\
\leq& \|\mathbf{B}'\widehat{\mathbf{\Sigma}}_{\epsilon s}\| + \|\mathbf{B}'\widehat{\mathbf{\Sigma}}_\epsilon\mathbf{A}_s\| \\
=& O_P(\|\mathbf{B}'\mathbf{\Sigma}_\epsilon\mathbf{B}\|^{\frac{1}{2}}pn^{-\frac{1}{2}} + \|\mathbf{B}'\mathbf{\Sigma}_\epsilon\mathbf{B}\|^{\frac{1}{2}}p^{\frac{\gamma}{2}}pn^{-\frac{1}{2}} + \|\mathbf{B}'\mathbf{\Sigma}_\epsilon\mathbf{A}\|) \\
=& O_P(\|\mathbf{\Sigma}_\epsilon\|^{\frac{1}{2}}pn^{-\frac{1}{2}} + \|\mathbf{\Sigma}_\epsilon\|^{-\frac{1}{2}}p^{\frac{\gamma}{2}}pn^{-\frac{1}{2}} + \|\mathbf{\Sigma}_\epsilon\|) \\
=& O_P(p^{\frac{\gamma}{2}}pn^{-\frac{1}{2}} + p^{\frac{\gamma}{2}}p^{\frac{\gamma}{2}}pn^{-\frac{1}{2}} + p^\gamma), \\
\|\mathbf{B}'\widehat{\mathbf{\Sigma}}_y\mathbf{A}_s\| =& O_P(p^{\frac{\gamma}{2}}pn^{-\frac{1}{2}} + p^\gamma + p^\gamma pn^{-\frac{1}{2}}).
\end{aligned}$$

Therefore, if $p^\gamma \geq 1$,

$$\|E_{21}\| = O_P(p^\gamma + p^\gamma pn^{-\frac{1}{2}}).$$

Finally,

$$\begin{aligned}
\|\widehat{\mathbf{A}}_s - \mathbf{A}_s\| =& \frac{\|E_{21}\|}{\operatorname{sep}\left(D_s, \begin{pmatrix} D_w & 0 \\ 0 & 0 \end{pmatrix}\right)} \\
=& \frac{O_P(p^\gamma + p^\gamma pn^{-\frac{1}{2}})}{p} \\
=& O_P(p^{\gamma-1} + p^\gamma n^{-\frac{1}{2}}).
\end{aligned}$$

The overall rate is determined by the local factor convergence speed. If we remove the effect of pervasive factors, then we have

$$\mathbf{\Sigma}_{y^*} = \mathbf{A}_w\mathbf{\Sigma}_w\mathbf{A}_w' + \mathbf{A}_s\mathbf{\Sigma}_s\mathbf{A}_s' + \mathbf{A}_s\widehat{\mathbf{\Sigma}}_{w\epsilon}\mathbf{H}_s + \mathbf{H}_s\mathbf{\Sigma}_\epsilon\mathbf{H}_s, \text{ where } \mathbf{H}_s = \mathbf{I} - \mathbf{A}_s\mathbf{A}_s'.$$
$$\widehat{\mathbf{\Sigma}}_{y^*} = \widehat{\mathbf{H}}_s\widehat{\mathbf{\Sigma}}_y\widehat{\mathbf{H}}_s, \text{ where } \widehat{\mathbf{H}}_s = \mathbf{I} - \widehat{\mathbf{A}}_s\widehat{\mathbf{A}}_s'.$$

therefore,

$$\begin{aligned}\widehat{\mathbf{\Sigma}}_{y^*} &= \widehat{\mathbf{H}}_s\mathbf{A}_s\widehat{\mathbf{\Sigma}}_s\mathbf{A}_s'\widehat{\mathbf{H}}_s + \widehat{\mathbf{H}}_s\mathbf{A}_s\widehat{\mathbf{\Sigma}}_{sw}\mathbf{A}_w'\widehat{\mathbf{H}}_s + \widehat{\mathbf{H}}_s\mathbf{A}_s\widehat{\mathbf{\Sigma}}_{s\epsilon}\widehat{\mathbf{H}}_s \\ &+ \widehat{\mathbf{H}}_s\mathbf{A}_w\widehat{\mathbf{\Sigma}}_{ws}\mathbf{A}_s'\widehat{\mathbf{H}}_s + \widehat{\mathbf{H}}_s\mathbf{A}_w\widehat{\mathbf{\Sigma}}_w\mathbf{A}_w'\widehat{\mathbf{H}}_s + \widehat{\mathbf{H}}_s\mathbf{A}_w\widehat{\mathbf{\Sigma}}_{w\epsilon}\widehat{\mathbf{H}}_s \\ &+ \widehat{\mathbf{H}}_s\widehat{\mathbf{\Sigma}}_{\epsilon s}\mathbf{A}_s'\widehat{\mathbf{H}}_s + \widehat{\mathbf{H}}_s\widehat{\mathbf{\Sigma}}_{\epsilon w}\mathbf{A}_w'\widehat{\mathbf{H}}_s + \widehat{\mathbf{H}}_s\widehat{\mathbf{\Sigma}}_{\epsilon}\widehat{\mathbf{H}}_s\end{aligned}$$

We want to find $\|\widetilde{\mathbf{A}}_w - \mathbf{A}_w\| = O_P(\frac{\|E_{21}^*\|}{\text{sep}(D_w^*,0)})$.

$$\begin{pmatrix}\mathbf{A}_w' \\ \mathbf{B}'\end{pmatrix}(\mathbf{A}_w\mathbf{\Sigma}_w\mathbf{A}_w')(\mathbf{A}_w\ \mathbf{B}) = \begin{pmatrix}D_w^* & 0 \\ 0 & 0\end{pmatrix}.$$

Hence,

$$\begin{aligned}\text{sep}(D_w^*, 0) &= \lambda_{\min}(D_w^*) \\ &\asymp \min_{1\le j\le r_2} p_j.\end{aligned}$$

$$\begin{aligned}Q'EQ &= \begin{pmatrix}\mathbf{A}_w' \\ \mathbf{B}'\end{pmatrix}(\widehat{\mathbf{\Sigma}}_{y^*} - \mathbf{A}_w\mathbf{\Sigma}_w\mathbf{A}_w')(\mathbf{A}_w\ \mathbf{B}) \\ &= \begin{pmatrix}E_{11}^* & E_{21}^{*T} \\ E_{21}^* & E_{22}^*\end{pmatrix}.\end{aligned}$$

After some algebra, we have

$$\begin{aligned}\|E_{21}^*\| &= O_P(\|\mathbf{B}'\widehat{\mathbf{\Sigma}}_{y^*}\mathbf{A}_w\|) \\ &= O_P(\|\mathbf{B}'\widehat{\mathbf{H}}_s\widehat{\mathbf{\Sigma}}_{\epsilon w}\mathbf{A}_w'\widehat{\mathbf{H}}_s\mathbf{A}_w\| + \|\mathbf{B}'\widehat{\mathbf{H}}_s\widehat{\mathbf{\Sigma}}_{\epsilon}\widehat{\mathbf{H}}_s\mathbf{A}_w\|) \\ &= O_P(r_2^{\frac{1}{2}}(p\cdot\max p_j)^{\frac{1}{2}}n^{-\frac{1}{2}} + \|\mathbf{B}'\mathbf{\Sigma}_{\epsilon}\mathbf{A}_w\|),\end{aligned}$$

and finally, the two step estimation rate for PCA is

$$\begin{aligned}\|\widetilde{\mathbf{A}}_w - \mathbf{A}_w\| &= O_P(\frac{\|E_{21}^*\|}{\text{sep}(D_w^*, 0)}) \\ &= O_P(\frac{r_2^{\frac{1}{2}}(p\cdot\max p_j)^{\frac{1}{2}}n^{-\frac{1}{2}} + \|\mathbf{B}'\mathbf{\Sigma}_{\epsilon}\mathbf{A}_w\|}{\min p_j}) \\ &= O_P(r_2^{\frac{1}{2}}p^{\frac{1}{2}}\max p_j^{\frac{1}{2}}\min p_j^{-1}n^{-\frac{1}{2}} + \min p_j^{-1}\|\mathbf{B}'\mathbf{\Sigma}_{\epsilon}\mathbf{A}_w\|)\end{aligned}$$

The proof for AFA method is in principle the same as that of Theorem 3 in Lam and Yao (2012), and thus omitted.

For the second part of the theorem, suppose we know the structure of the data as defined in (2.2.3), then, for ith local factor,

$$\|\widetilde{\mathbf{A}}_{w_i} - \mathbf{A}_{w_i}\| = O_P(\frac{\|\mathbf{B}_i'E_i\mathbf{A}_{w_i}\|}{\text{sep}(\mathbf{\Sigma}_w^{(i)}, 0)}),$$

where $\mathbf{B}_i$ is the orthogonal complement of $\mathbf{A}_{w_i}$, and $\mathbf{\Sigma}_w^{(i)}$ is the covariance matrix for the ith group local factors. For ith group of local factors, we data $y_{t,i}^*$, that is the corresponding rows of y_t^*. Then we have

$$E_i = \widehat{\mathbf{\Sigma}}_{y_i^*} - \mathbf{A}_{w_i}\mathbf{\Sigma}_w^{(i)}\mathbf{A}'_{w_i}.$$

Therefore,

$$\begin{aligned}
\mathbf{B}_i' E_i \mathbf{A}_{w_i} &= \mathbf{B}_i'\widehat{\mathbf{H}}_s^{(i)}\mathbf{A}_s\widehat{\mathbf{\Sigma}}_s\mathbf{A}_s'\widehat{\mathbf{H}}_s^{(i)}\mathbf{A}_w + \mathbf{B}_i'\widehat{\mathbf{H}}_s^{(i)}\mathbf{A}_s\widehat{\mathbf{\Sigma}}_{sw}^{(i)}\mathbf{A}_w'\widehat{\mathbf{H}}_s^{(i)}\mathbf{A}_w + \mathbf{B}_i'\widehat{\mathbf{H}}_s^{(i)}\mathbf{A}_s\widehat{\mathbf{\Sigma}}_{s\epsilon}^{(i)}\widehat{\mathbf{H}}_s^{(i)}\mathbf{A}_w \\
&+ \mathbf{B}_i'\widehat{\mathbf{H}}_s^{(i)}\mathbf{A}_w\widehat{\mathbf{\Sigma}}_{ws}^{(i)}\mathbf{A}_s'\widehat{\mathbf{H}}_s^{(i)}\mathbf{A}_w + \mathbf{B}_i'\widehat{\mathbf{H}}_s^{(i)}\mathbf{A}_w\widehat{\mathbf{\Sigma}}_w^{(i)}\mathbf{A}_w'\widehat{\mathbf{H}}_s^{(i)}\mathbf{A}_w + \mathbf{B}_i'\widehat{\mathbf{H}}_s^{(i)}\mathbf{A}_w\widehat{\mathbf{\Sigma}}_{w\epsilon}^{(i)}\widehat{\mathbf{H}}_s^{(i)}\mathbf{A}_w \\
&+ \mathbf{B}_i'\widehat{\mathbf{H}}_s^{(i)}\widehat{\mathbf{\Sigma}}_{\epsilon s}^{(i)}\mathbf{A}_s'\widehat{\mathbf{H}}_s^{(i)}\mathbf{A}_w + \mathbf{B}_i'\widehat{\mathbf{H}}_s^{(i)}\widehat{\mathbf{\Sigma}}_{\epsilon w}^{(i)}\mathbf{A}_w'\widehat{\mathbf{H}}_s^{(i)}\mathbf{A}_w + \mathbf{B}_i'\widehat{\mathbf{H}}_s^{(i)}\widehat{\mathbf{\Sigma}}_{\epsilon}^{(i)}\widehat{\mathbf{H}}_s^{(i)}\mathbf{A}_w \\
&= \Sigma_{j=1}^9 I_j.
\end{aligned}$$

Also we have,

$$\begin{aligned}
&\|\widehat{\mathbf{\Sigma}}_s^{(i)}\| = O_P(p), \\
&\|\widehat{\mathbf{\Sigma}}_w^{(i)}\| = O_P(p_i), \\
&\|\widehat{\mathbf{\Sigma}}_{sw}^{(i)}\| = \|\widehat{\mathbf{\Sigma}}_{ws}^{(i)}\| = O_P(r_2^{\frac{1}{2}} p_i n^{-\frac{1}{2}}), \\
&\|\widehat{\mathbf{\Sigma}}_{s\epsilon}^{(i)}\| = \|\widehat{\mathbf{\Sigma}}_{\epsilon s}\| = O_P(p^{\frac{1}{2}} n^{-\frac{1}{2}}), \\
&\|\widehat{\mathbf{\Sigma}}_{w\epsilon}^{(i)}\| = \|\widehat{\mathbf{\Sigma}}_{\epsilon w}\| = O_P(p_i n^{-\frac{1}{2}}), \\
&\|\widehat{\mathbf{\Sigma}}_{\epsilon}^{(i)} - \mathbf{\Sigma}_{\epsilon}^{(i)}\| = O_P(p_i n^{-\frac{1}{2}}).
\end{aligned}$$

Hence the dominating terms are $\|I_8\| + \|I_9\|$, therefore,

$$\mathbf{B}_i' E_i \mathbf{A}_{w_i} = O_P(p_i n^{-\frac{1}{2}} + \|\mathbf{B}_i'\mathbf{\Sigma}_{\epsilon}^{(i)}\mathbf{A}_w^{(i)}\|). \tag{2.7.1}$$

Finally, if we know the local factor structure group, for ith group of local factors,

$$\begin{aligned}
\|\widetilde{\mathbf{A}}_w^{(i)} - \mathbf{A}_w^{(i)}\| &== O_P(\frac{p_i n^{-\frac{1}{2}} + \|\mathbf{B}_i'\mathbf{\Sigma}_{\epsilon}^{(i)}\mathbf{A}_w^{(i)}\|}{p_i}) \\
&= O_P(n^{-\frac{1}{2}} + p_i^{-1}\|\mathbf{B}'\mathbf{\Sigma}_{\epsilon}^{(i)}\mathbf{A}_w^{(i)}\|).
\end{aligned} \tag{2.7.2}$$

We can use a similar argument to prove the rate for AFA. □

2.8 Discussion to "Large covariance estimation by thresholding principal orthogonal complements"

We congratulate the authors for this insightful paper.[1] Here we suggest a method to address two concerns:

1. The potential underestimation of the number of factors K;

2. The potential non-sparseness of the estimated principal orthogonal complement.

The first point is addressed by using a larger K. With pervasive factors assumed in the paper, it is relatively easy to find such K. However, in an analysis of macroeconomic data for example, there can be a mix of pervasive factors and many weaker ones; see [15, 12, 27], for a general definition of local factors. In [39], a monthly data of $p = 132$ U.S. macroeconomic time series from 1959 to 2003 ($n = 5261$) is analyzed. Using principal component analysis (PCA) [**?**], the method in [12] and a modified version called the autocovariance-based factor modeling (AFA) (details omitted), we compute the average forecast errors of 30 monthly forecasts using a vector autoregressive model VAR(3) on the estimated factors from these methods with different number of factors r (Figure 2.4). While 3 pervasive factors decrease forecast errors sharply, including more factors, up to $r = 35$, decrease forecast errors slower, showing the existence of many weaker factors. Hence it is not always possible to have

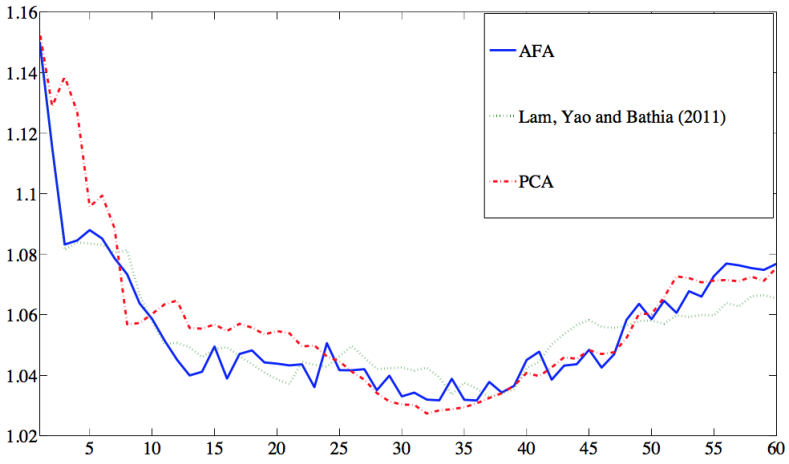

Figure 2.4: Average forecast errors for different number of factors r.

[1]This section is the discussion paper [26] to the article [22] by Jianqing Fan, Yuan Liao and Martina Mincheva.

"enough" factors for accurate thresholding of the principal orthogonal complement, which can still include contribution from many local factors and is not sparse. Points 1 and 2 can therefore be closely related, and can be addressed if we regularize the condition number of the orthogonal complement instead of thresholding. While [7] restrict the extreme eigenvalues with a tuning parameter to be chosen, we use the idea of [1] (properties are not investigated enough unfortunately). We simulate 100 times from the panel regression model

$$y_t = X_t\beta + \epsilon_t, \beta = (-0.5, 0.5, 0.3, -0.6)', \tag{2.8.1}$$

with x_{it} being independent AR(1) processes and ϵ_t the standardized macroeconomic data in [6] plus independent $N(0, 0.2)$ noise. Following Example 5 of the paper, we estimate $\mathbf{\Sigma}_\epsilon^{-1}$ using different methods and plot the sum of absolute bias for estimating β using generalized least square (GLS) against the number of factors r used in Figure 2.5. Clearly regularizing on condition number leads to stabler estimators.

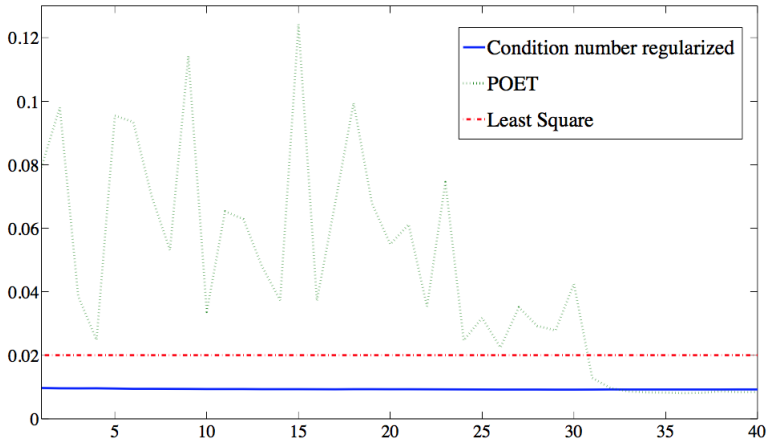

Figure 2.5: Sum of absolute bias (averaged over 100 simulations) for estimating β using GLS against the number of factors r used in POET (C=0.5) and the condition number regularized estimator. Bias for least square method is constant throughout.

Parallel to section 7.2 in [22], we compare the risk of portfolios created using POET and the method above. Again Figure 2.6 shows stabler performance of regularization on condition number.

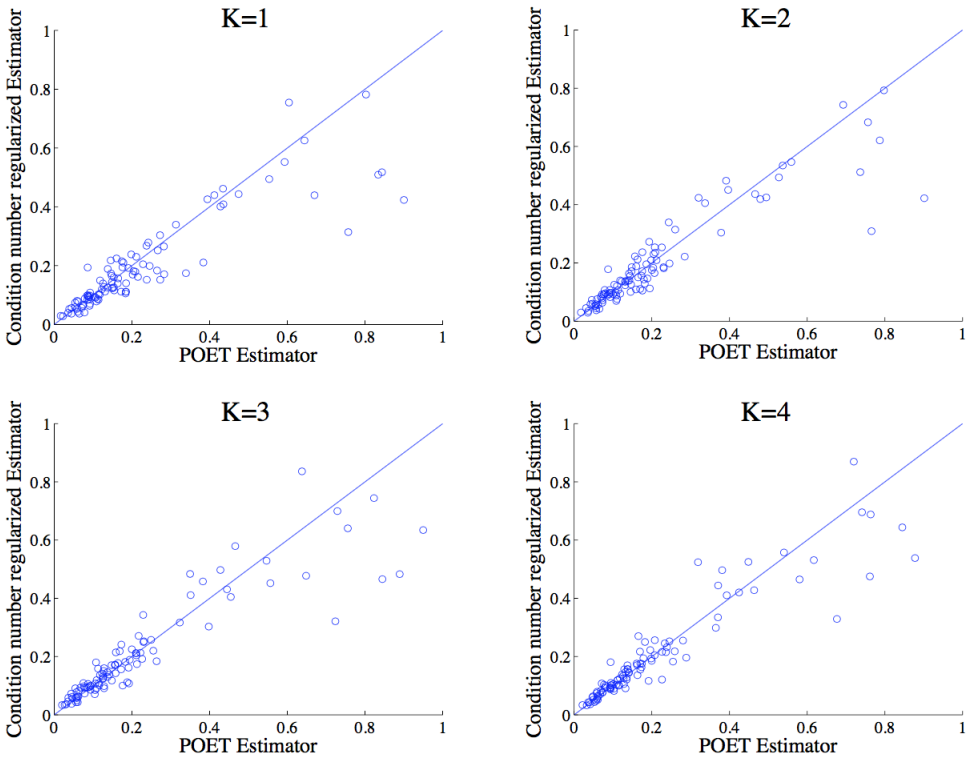

Figure 2.6: Risk of portfolios created with POET (C=0.5) and condition number regularized estimator.

Chapter 3

Nonlinear shrinkage of large integrated covariance matrix

3.1 An overview of relevant estimation methods in the literature

3.1.1 The problem of covariance matrix estimation

In data analysis, one of the most widely used entity is the covariance matrix. It has many applications in different fields of studies, including principle component analysis, network analysis, linear discriminant analysis and so on. In particular, covariance matrix plays an important role in Markowitz's mean-variance optimization [34], the covariance matrix of stock returns concerns risk management. Suppose we denote $\mathbf{Y}$ to be the $p \times n$ data matrix consists of n independent and identically distributed sample that we can observe,[1] sample covariance matrix which is defined by

$$\mathbf{S}_n = \frac{1}{n}\mathbf{Y}\mathbf{Y}', \tag{3.1.1}$$

is an often used estimator due to its simplicity.

However, when the dimensionality of the data is high, the number of parameter needs to be estimated grows very fast. To be specific, suppose we have a $p \times p$ matrix, the number of parameters needs to be estimated is of order p^2 ($\frac{1}{2}p(p+1)$). When the dimensionality p grows, the number of parameters to be estimated increases proportional to p^2, but the

[1] Without loss of generality we assume that the mean of each row vector of $\mathbf{Y}$ is zero.

number of observations available only grow proportional to p. In parameter estimation for a structured covariance matrix, simulation results show that parameter estimation becomes very poor when the number of parameters is more than four [6]. This renders the sample covariance matrix unreliable, the coefficients in the sample covariance contain an extreme amount of error and when the dimension is larger than the sample size, the sample covariance matrix even becomes singular. In this case, improved estimator methods for the underlying true covariance matrix are needed.

The most commonly used methods for covariance estimation can be categorized into two broad classes. The first one is by imposing certain structural assumptions on the underlying matrices, and develop certain convergence results. Additional knowledge assumed including sparseness, a graph model or a factor model

Nonlinear shrinkage

Another class of estimator does not assume a specific structure of the underlying covariance matrix. Since the sample covariance accumulates errors through its coefficients, and the most extreme estimated coefficients contribute a lot of errors. The idea of *shrinkage* arises to rectify the problem. In [29], Ledoit and Wolf demonstrated that the largest sample eigenvalues are systematically biased upwards, and the smallest ones downwards. Intuitively, we want to find a way to 'pull' those extreme coefficients towards the center.

Stein [38] advocated the use of the class of rotation-equivariant estimator[2] that shrink the eigenvalues of sample covariances and keep the eigenvectors intact.

Many methods are proposed along this line. The goal is to find an estimator that minimize the frobenius norm of the difference between the estimator and true underlying matrix:

$$\min_{D} \|PDP' - \mathbf{\Sigma}\|_F. \tag{3.1.2}$$

The solution to this optimization problem is $D = diag(d_1, \cdots, d_p)$, where

$$d_i = p_i' \mathbf{\Sigma} p_i, i = 1, \cdots, p, p_i istheeigenvector. \tag{3.1.3}$$

However, the quantity d_i is not readily available and needs to be estimated. Ledoit and Wolf [31] propose a nonlinear shrinkage formula to do this. Whereas Lam [23] introduces

[2] An estimator is rotation-equivariant if and only if it has the same eigenvectors as the sample covariance matrix so a rotation-equivariant can be written as PDP', where matrix P consists of eigenvector of sample covariance matrix and D is a diagonal matrix .

the so-called Nonparametric Eigenvalue-Regularized COvariance Matrix Estimator (NERCOME) to achieve the goal of shrinking eigenvalues by subsampling the data. Those methods are asymptotically equivalent[23].

3.1.2 Large dimension integrated covariance matrix estimation

In financial econometrics, when it comes to high-frequency data regime, the independent assumption of data is no longer appropriate. For instance, if one looks at intra-day high-frequency stock returns, the variance patterns are highly time-varying and cannot be assumed to be independent. In this case, instead of using the sample covariance matrice, one can use the so-called *Integrated Covariance matrix* (ICV) which essentially can be thought as the average over a certain time period of instantaneous covariance matrices.

The ICV matrix is difficult to estimate due to the integrated nature of the estimation problem. By considering a particular class of underlying multivariate process, Zheng and Li [43] manage to develop the so-called *Time-Variation Adjusted Realized Covariance* (TVARCV) matrix. Intuitively speaking, it serves as a counterpart of the standard sample covariance matrix in the ICV framework.

As in the standard covariance matrix estimation case, in the large dimension asymptotic setting, the use of sample eigenvalues is not appropriate. Neither the use of population eigenvalues is optimal due to the estimation errors in the sample eigenvector.

In fact, the eigenvalues should be shrunk nonlinearly to a set of less dispersed values. Therefore, by applying nonlinear shrinkage method, one should expect to obtain a better estimator for large dimension ICV estimation. Chapter 3 explicitly propose an estimator for the ICV, reports its theoretical properties and numerical studies.

3.2 Introduction

With the easily obtainable intra-day trading data nowadays, financial market analysts and academic researchers enjoy more accurate return or volatility matrix estimation through the substantial increase in sample size. Yet, with respect to the integrated covariance matrix estimation for asset returns, there are several well-known challenges using such intra-day price data. For instance, when tick-by-tick price data is used, the contamination by market microstructure noise [2, 4] can hugely bias the realized covariance matrix. Non-synchronous

trading times presents another challenge when there are more than one asset to consider.

To present a further challenge, it is well-documented that with independent and identically distributed random vectors, random matrix theories imply that there are biased extreme eigenvalues for the corresponding sample covariance matrix when the dimension of the random vectors p has the same order as the sample size n, i.e., $p/n \to c > 0$ for some constant $c > 0$. See for instance [7] for more details. This suggests that the realized covariance matrix, which is essentially a sample covariance matrix when all covolatilities are constants and all log prices have zero drift with equally-spaced observation times (see the diffusion process for the log price defined in (3.3.1) for more details), can have biased extreme eigenvalues under the high dimensional setting $p/n \to c > 0$. The resulting detrimental effects to risk estimation or portfolio allocation are thoroughly demonstrated in [8] when inter-day price data is used.

To rectify this bias problem, many researchers focused on regularized estimation of covariance or precision matrices with special structures. These go from banded [11] or sparse covariance matrix [10, 13, 25, 37], sparse precision matrix [20, 35], sparse modified Cholesky factor [36], to a spiked covariance matrix from a factor model [16, 18], or combinations of these [19].

Recently, Ledoit and Wolf [30] proposed a nonlinear shrinkage formula for shrinking the extreme eigenvalues in a sample covariance matrix without assuming a particular structure of the true covariance matrix. The method is generalized in [32] for portfolio allocation with remarkable results. However, such a nonlinear shrinkage formula is only applicable to the independent and identically distributed random vector setting. It is not applicable to intra-day price data since the volatility within a trading day is highly variable, so that asset returns at different time periods, albeit independent theoretically, are not identically distributed.

Lam [23] proves that by splitting the data into two independent portions of certain sizes, one can achieve the same nonlinear shrinkage asymptotically without the need to evaluate a shrinkage formula as in [30], which can be computationally expensive. At the same time, such a data splitting approach can be generalized to adapt to different data settings. In this chapter, we modify the method proposed in [23] to achieve nonlinear shrinkage of eigenvalues in the realized covariance matrix using intra-day price data. We use the same assumption as in [43] (see Assumption 3.3.1 in Section 3.3 and the details therein) to overcome the difficulty of time-varying volatilities for all underlying stocks. Ultimately, our method produces a

positive definite integrated covariance matrix asymptotically almost surely with shrinkage of eigenvalues achieved nonlinearly, while local integrated covolatilities are adapted and estimated accurately. Our method is fast since it involves only eigen-decompositions of matrices of size $p \times p$, which is not computationally expensive when p is of order of hundreds. This is usually the typical order for p in the case of portfolio allocation. We also present the maximum exposure bound and the actual risk bound for portfolio allocation using our estimator as an input for the minimum variance portfolio. These bounds are important in practice as seen in the real data analysis results in Section 3.5.2, when our portfolio do not over-invest in individual assets, and the actual risk is small compared to other methods.

The rest of the chapter is organized as follows. We first present the framework for the data together with the notations and the main assumptions to be used in Section 3.3. Our method of estimation is detailed in Section 3.3.1, while Section 3.4 presents all related theories. Simulation results are given in Section 3.5, with the theorem concerning the maximum exposure bound and the actual risk bound in portfolio allocation using our method presented in Section 3.5.1. A real data example of portfolio allocation is presented in Section 3.5.2. All proofs are presented at the end of the chapter.

3.3 Framework and Methodology

Let $\mathbf{X}_t = (X_t^{(1)}, \cdots, X_t^{(p)})^{\mathrm{T}}$ be a p-dimensional log-price process which is modeled by the diffusion process

$$d\mathbf{X}_t = \boldsymbol{\mu}_t dt + \boldsymbol{\Theta}_t d\mathbf{W}_t, \ t \in [0, 1], \tag{3.3.1}$$

where $\boldsymbol{\mu}_t$ is the drift, $\boldsymbol{\Theta}_t$ is a $p \times p$ matrix called the (instantaneous) covolatility process, and $\mathbf{W}_t = (W_t^{(1)}, \ldots, W_t^{(p)})^{\mathrm{T}}$ is a p-dimensional standard Brownian motion. We want to estimate the integrated covariance matrix

$$\boldsymbol{\Sigma}_p = \int_0^1 \boldsymbol{\Theta}_t \boldsymbol{\Theta}_t^{\mathrm{T}} dt. \tag{3.3.2}$$

It is well-known that the log-price process $\mathbf{X}_t$ is contaminated by market microstructure noise; see [42] for instance when the tick-by-tick high-frequency trading data is used to calculate an integrated covariance estimator. In this paper, instead of using the tick-by-tick data which has the highest observation frequency possible, we use sparsely sampled data synchronized by refresh times [3, 9], so that the theory in our paper should be readily applicable. Hence in the sequel, we assume that we can observe the price $\mathbf{X}_t$ at synchronous

time points $\tau_{n,\ell}$, $\ell = 0, 1, \ldots, n$. The realized covariance matrix is then defined as

$$\mathbf{\Sigma}_p^{\mathrm{RCV}} = \sum_{\ell=1}^{n} \Delta\mathbf{X}_\ell \Delta\mathbf{X}_\ell^{\mathrm{T}}, \quad \text{where } \Delta\mathbf{X}_\ell := \mathbf{X}_{\tau_{n,\ell}} - \mathbf{X}_{\tau_{n,\ell-1}}. \tag{3.3.3}$$

[21] shows that as n goes to infinity, the above estimator converges weakly to the true one defined in (3.3.2). Hence the realized covariance matrix is one of the most frequently used estimator for the integrated covariance matrix.

While the intra-day volatility can change hugely within a short time period, it is not unreasonable to assume that the correlation of any two price processes stays constant within such a period, say within a trading day. Following [43], for $j = 1, \ldots, p$, write

$$dX_t^{(j)} = \mu_t^{(j)} + \sigma_t^{(j)} dZ_t^{(j)},$$

where $\mu_t^{(j)}, \sigma_t^{(j)}$ are assumed to be càdlàg over $[0, 1]$, and the $Z_t^{(j)}$'s are one dimensional standard Brownian motions. Both the $\sigma_t^{(j)}$'s and the $Z_t^{(j)}$'s are related to $\mathbf{\Theta}_t$ and $\mathbf{W}_t$ in (3.3.1). We assume further, defining $\langle X, Y\rangle_t$ to be the quadratic covariation between the processes X and Y:

Assumption 3.3.1 The correlation matrix process of $\mathbf{Z}_t = (Z_t^{(1)}, \ldots, Z_t^{(p)})^{\mathrm{T}}$, defined by

$$R_t = \big(\langle Z^{(j)}, Z^{(k)}\rangle_t / t\big)_{1\le j,k\le p},$$

is constant and non-zero on $(0, 1]$ for each j, k. Furthermore, the correlation matrix process of $\mathbf{X}_t$, defined by

$$\left(\frac{\int_0^t \sigma_s^{(j)}\sigma_s^{(k)} d\langle Z^{(j)}, Z^{(k)}\rangle_s}{\sqrt{\int_0^t (\sigma_s^{(j)})^2 ds \cdot \int_0^t (\sigma_s^{(k)})^2 ds}}\right)_{1\le j,k\le p},$$

is constant on $(0, 1]$ for each j, k.

The rest of the assumptions in this paper can be found in Section 3.4. We present this assumption first since following Proposition 4 in [43], the log-price process $\mathbf{X}_t$ defined in (3.3.1) satisfying Assumption 3.3.1 is such that there exist a càdlàg process $(\gamma_t)_{t\in[0,1]}$ and a $p \times p$ matrix $\mathbf{\Lambda}$ satisfying $\mathrm{tr}(\mathbf{\Lambda}\mathbf{\Lambda}^{\mathrm{T}}) = p$ such that

$$\mathbf{\Theta}_t = \gamma_t \mathbf{\Lambda}. \tag{3.3.4}$$

The nonlinear shrinkage estimator described in the next section is based on this property.

3.3.1 Nonlinear shrinkage estimator

When the dimension p is large relative to the sample size n, even for a sample covariance matrix constructed from independent and identically distributed random vectors, its extreme eigenvalues will be severely biased from the true ones (see chapter 5.2 of [7] for example). While various assumptions have been made on the true integrated covariance matrix like sparsity [41] or having a factor structure [40], in this paper we follow [30] and introduce nonlinear shrinkage for regularization, which does not need a particular structural assumption on the true integrated covariance matrix itself.

However, since intra-day covariance can vary hugely within a short time period, the $\Delta\mathbf{X}_\ell$'s defined in (3.3.3) are not identically distributed, and hence we cannot directly apply the nonlinear shrinkage formula in [30] to the realized covariance matrix in (3.3.3). Instead, we use the data splitting idea for nonlinear shrinkage of eigenvalues in [23], and modify their method to accommodate the intra-day volatility change base on (3.3.4), which is a condition derived from Assumption 3.3.1 as proved in [43].

To this end, observe that by (3.3.4), the integrated covariance matrix in (3.3.2) can be written as $\mathbf{\Sigma}_p = \int_0^1 \gamma_t^2 dt \cdot \mathbf{\Lambda}\mathbf{\Lambda}^{\mathrm{T}}$. [43] proposed a so-called Time-variation adjusted realized covariance matrix, defined as

$$\check{\mathbf{\Sigma}}_p := \frac{\mathrm{tr}(\mathbf{\Sigma}_p^{\mathrm{RCV}})}{p}\check{\mathbf{\Phi}}, \quad \text{where } \check{\mathbf{\Phi}} := \frac{p}{n}\sum_{\ell=1}^{n}\frac{\Delta\mathbf{X}_\ell \Delta\mathbf{X}_\ell^{\mathrm{T}}}{\left\|\Delta\mathbf{X}_\ell\right\|^2}, \tag{3.3.5}$$

and $\left\|\cdot\right\|$ denotes the norm of a vector. They demonstrate that $\check{\mathbf{\Sigma}}_p$ is a good estimator for $\mathbf{\Sigma}_p$ by showing that $\mathrm{tr}(\mathbf{\Sigma}_p^{\mathrm{RCV}})/p$ is a good estimator for $\int_0^1 \gamma_t^2 dt$, while $\check{\mathbf{\Phi}}$ is good for $\mathbf{\Phi} = \mathbf{\Lambda}\mathbf{\Lambda}^{\mathrm{T}}$. Here $\check{\mathbf{\Phi}}$ plays the role of a sample covariance matrix for estimating $\mathbf{\Phi}$. Hence if $p/n \to c > 0$, then $\check{\mathbf{\Phi}}$ suffers from bias to the extreme eigenvalues as well.

Remark 3.3.2 An intuition of why $\check{\mathbf{\Phi}}$ is similar to a sample covariance matrix can be seen as follows. If $\boldsymbol{\mu}_t = \mathbf{0}$ in (3.3.1) and the $\tau_{n,\ell}$'s are independent of $\mathbf{W}_t$ (see Assumptions 3.4.1 and 3.4.2 respectively in Section 3.4), then by model (3.3.1), we can write

$$\Delta\mathbf{X}_\ell = \int_{\tau_{n,\ell-1}}^{\tau_{n,\ell}} \gamma_t \mathbf{\Lambda} d\mathbf{W}_t \overset{d}{=} \Big(\int_{\tau_{n,\ell-1}}^{\tau_{n,\ell}} \gamma_t^2 dt\Big)^{1/2} \mathbf{\Phi}^{1/2}\mathbf{Z}_\ell,$$

where $\overset{d}{=}$ stands for equal in distribution, and the $\mathbf{Z}_\ell$'s are independent random vectors each with $\mathbf{Z}_\ell \sim N(\mathbf{0}, \mathbf{I}_p)$. Then

$$\check{\mathbf{\Phi}} \overset{d}{=} \frac{1}{n}\sum_{\ell=1}^{n}\frac{\mathbf{\Phi}^{1/2}\mathbf{Z}_\ell\mathbf{Z}_\ell^{\mathrm{T}}\mathbf{\Phi}^{1/2}}{\mathbf{Z}_\ell^{\mathrm{T}}\mathbf{\Phi}\mathbf{Z}_\ell/p} = \mathbf{\Phi}^{1/2}\Big(\frac{1}{n}\sum_{\ell=1}^{n}\frac{\mathbf{Z}_\ell\mathbf{Z}_\ell^{\mathrm{T}}}{\mathbf{Z}_\ell^{\mathrm{T}}\mathbf{\Phi}\mathbf{Z}_\ell/p}\Big)\mathbf{\Phi}^{1/2}.$$

We can actually show that $\mathbf{Z}_\ell^{\mathrm{T}}\mathbf{\Phi}\mathbf{Z}_\ell/p$ goes to 1 almost surely, leaving the above being the sample covariance matrix constructed from the $\mathbf{Z}_\ell$'s sandwiched by $\mathbf{\Phi}^{1/2}$.

Following [23], since the $\Delta\mathbf{X}_\ell$'s are independent following model (3.3.1), we split the data $\Delta\mathbf{X} = (\Delta\mathbf{X}_1, \ldots, \Delta\mathbf{X}_n)$ into two independent parts, say $\Delta\mathbf{X} = (\Delta\mathbf{X}^1, \Delta\mathbf{X}^2)$, with $\Delta\mathbf{X}^i$ having size $p \times n_i$ for $i = 1, 2$, such that $n = n_1 + n_2$. Define

$$\widetilde{\mathbf{\Phi}}_i = \frac{p}{n_i} \sum_{\ell \in I_i} \frac{\Delta\mathbf{X}_\ell \Delta\mathbf{X}_\ell^{\mathrm{T}}}{\left\|\Delta\mathbf{X}_\ell\right\|^2}, \tag{3.3.6}$$

where $I_i = \{\ell : \Delta\mathbf{X}_\ell \in \Delta\mathbf{X}^i\}$. Carrying out an eigen-analysis on $\widetilde{\mathbf{\Phi}}_1$ defined in (3.3.6) above, suppose $\widetilde{\mathbf{\Phi}}_1 = \mathbf{P}_1\mathbf{D}_1\mathbf{P}_1^{\mathrm{T}}$. Then we introduce our estimator as

$$\widehat{\mathbf{\Sigma}}_p := \frac{\mathrm{tr}(\mathbf{\Sigma}_p^{\mathrm{RCV}})}{p}\widehat{\mathbf{\Phi}}, \quad \text{where } \widehat{\mathbf{\Phi}} := \mathbf{P}_1 \mathrm{diag}(\mathbf{P}_1^{\mathrm{T}}\widetilde{\mathbf{\Phi}}_2\mathbf{P}_1)\mathbf{P}_1^{\mathrm{T}}, \tag{3.3.7}$$

with $\mathrm{diag}(\cdot)$ setting all non-diagonal elements of a matrix to 0. The estimator $\widehat{\mathbf{\Phi}}$ above belongs to a class of rotation equivariant estimator $\mathbf{\Phi}(\mathbf{D}) = \mathbf{P}_1\mathbf{D}\mathbf{P}_1^{\mathrm{T}}$, where $\mathbf{D}$ is a diagonal matrix, and $\mathbf{P}_1$ is the matrix containing all the eigenvectors of $\widetilde{\mathbf{\Phi}}_1$. The choice of $\mathbf{D} = \mathrm{diag}(\mathbf{P}_1^{\mathrm{T}}\widetilde{\mathbf{\Phi}}_2\mathbf{P}_1)$ comes from solving

$$\min_{\mathbf{D}} \left\|\mathbf{P}_1\mathbf{D}\mathbf{P}_1^{\mathrm{T}} - \widetilde{\mathbf{\Phi}}_2\right\|_F,$$

where $\left\|\mathbf{A}\right\|_F = \mathrm{tr}^{1/2}(\mathbf{A}\mathbf{A}^{\mathrm{T}})$ is the Frobenius norm of a matrix. Similar to [23], regularization of the eigenvalues in $\mathbf{D} = \mathrm{diag}(\mathbf{P}_1^{\mathrm{T}}\widetilde{\mathbf{\Phi}}_2\mathbf{P}_1)$ comes from the independence between $\mathbf{P}_1$ and $\widetilde{\mathbf{\Phi}}_2$, since $\Delta\mathbf{X}^1$ is independent of $\Delta\mathbf{X}^2$.

3.4 Asymptotic Theory and Practical Implementation

We introduce two more assumptions needed for our results to hold. Assumption 3.3.1 is presented in Section 3.3.

Assumption 3.4.1 The drift in (3.3.1) satisfies $\boldsymbol{\mu}_t = \mathbf{0}$ for $t \in [0, 1]$. All eigenvalues of $\mathbf{\Theta}_t\mathbf{\Theta}_t^{\mathrm{T}}$ are bounded uniformly from 0 and infinity in $t \in [0, 1]$.

Assumption 3.4.2 The observation times $\tau_{n,\ell}$'s are independent of the log-price $\mathbf{X}_t$, and there exists a constant $C > 0$ such that for all positive integer n,

$$\max_{1 \le \ell \le n} n(\tau_{n,\ell} - \tau_{n,\ell-1}) \le C.$$

We set $\boldsymbol{\mu}_t = \mathbf{0}$ in Assumption 3.4.1 for the ease of proofs and presentation. If $\boldsymbol{\mu}_t$ is slowly varying locally, the results to be presented are still valid at the expense of longer and more complex proofs. The uniform bounds on the eigenvalues of $\boldsymbol{\Theta}_t\boldsymbol{\Theta}_t^{\mathrm{T}}$ are needed so that individual volatility process for each $X_t^{(i)}$ are bounded uniformly. Also, $\int_0^1 \gamma_t^2 dt > 0$ uniformly, and finally, $\left\|\boldsymbol{\Sigma}_p\right\| = O(1)$ uniformly as a result, which are all needed for our results to hold. These assumptions essentially treat γ_t as non-random. Extension to γ_t being stochastic can follow the lines of [43], but we keep it non-random for the ease of presentation and proofs as well.

Lemma 2 *Let Assumptions 3.3.1, 3.4.1 and 3.4.2 hold for the log-price process $\mathbf{X}_t$ in (3.3.1). Then for the estimator $\widehat{\boldsymbol{\Phi}}$ in (3.3.7), writing $\mathbf{P}_1 = (\mathbf{p}_{11}, \ldots, \mathbf{p}_{1p})$, if $p/n \to c > 0$ and $\sum_{n_2 \geq 1} pn_2^{-5} < \infty$, we have*

$$\max_{1 \leq i \leq p} \left| \frac{\mathbf{p}_{1i}^{\mathrm{T}} \widetilde{\boldsymbol{\Phi}}_2 \mathbf{p}_{1i} - \mathbf{p}_{1i}^{\mathrm{T}} \boldsymbol{\Phi} \mathbf{p}_{1i}}{\mathbf{p}_{1i}^{\mathrm{T}} \boldsymbol{\Phi} \mathbf{p}_{1i}} \right| \overset{a.s.}{\to} 0,$$

where $\overset{a.s.}{\to}$ represents almost sure convergence.

Since the eigenvalues of $\widehat{\boldsymbol{\Phi}}$ are the $\mathbf{p}_{1i}^{\mathrm{T}} \widetilde{\boldsymbol{\Phi}}_2 \mathbf{p}_{1i}$'s, the above Lemma shows that they are regularized to $\mathbf{p}_{1i}^{\mathrm{T}} \boldsymbol{\Phi} \mathbf{p}_{1i}$ asymptotically almost surely, which has values bounded by $\lambda_{\min}(\boldsymbol{\Phi})$ and $\lambda_{\max}(\boldsymbol{\Phi})$, the minimum and maximum eigenvalues of $\boldsymbol{\Phi}$ respectively. Assumption 3.4.1 ensures that these eigenvalues are uniformly bounded away from 0 and infinity, and hence $\widehat{\boldsymbol{\Phi}}$ is asymptotically almost surely positive definite. This is true even when the constant $c > 1$, i.e., when p is larger than n as they grow together to infinity.

With this result, we can present the following theorem.

Theorem 3 *Let all the assumptions in Lemma 2 hold. Then as $p, n \to \infty$ such that $p/n \to c > 0$, $\widehat{\boldsymbol{\Sigma}}_p$ defined in (3.3.7) is almost surely positive definite.*

This is an important result since $\boldsymbol{\Sigma}_p$ is always assumed to be positive definite, and we want our estimator to be so too. This is certainly not the case for a sample covariance matrix when $p > n$, and is still not the case for $\check{\boldsymbol{\Sigma}}_p$ defined in (3.3.5) by [43], which is demonstrated in our simulation results in Section 3.5.

Remark 3.4.3 Both Lemma 2 and Theorem 3 requires $\sum_{n_2 \geq 1} pn_2^{-5} < \infty$. Following [23], we set $n_2 = an^{1/2}$ where a is a constant, so that when $p/n \to c > 0$, the condition is satisfied. See Section 3.4.1 for more details on how to find n_2 with finite sample.

To present the rest of the results, we introduce a benchmark estimator for comparisons to our estimator. This estimator is called the ideal estimator, defined by

$$\mathbf{\Sigma}_{\text{ideal}} = \int_0^1 \gamma_t^2 dt \cdot \mathbf{P}\text{diag}(\mathbf{P}^{\text{T}}\mathbf{\Phi}\mathbf{P})\mathbf{P}^{\text{T}}.$$

This is similar to the proposed estimator defined in (3.3.7), except that the estimator $\text{tr}(\mathbf{\Sigma}_p^{\text{RCV}})/p$ is replaced by the population counterpart $\int_0^1 \gamma_t^2 dt$, while $\widetilde{\mathbf{\Phi}}_2$ is replaced by the population counterpart $\mathbf{\Phi}$. Also, $\mathbf{P}_1$ is replaced by $\mathbf{P}$, which is the matrix containing all orthonormal eigenvectors for the covariance-type matrix $\check{\mathbf{\Phi}}$ defined in (3.3.5) using all data points. This is in line with the ideal estimator defined in [30] and [23] which utilizes all data points for calculating the eigenmatrix $\mathbf{P}$, and assumes the knowledge of $\mathbf{\Phi}$ and $\int_0^1 \gamma_t^2 dt$. With this, we define the efficiency loss of any estimator $\widehat{\mathbf{\Sigma}}$ as

$$EL(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}}) := 1 - \frac{L(\mathbf{\Sigma}_p, \mathbf{\Sigma}_{\text{Ideal}})}{L(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}})},$$

where $L(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}})$ is a loss function for estimating $\mathbf{\Sigma}_p$ by $\widehat{\mathbf{\Sigma}}$. We consider the Frobenius loss

$$L(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}}) = \left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}_p\right\|_F^2, \tag{3.4.1}$$

and the inverse Stein's loss function in this paper,

$$L(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}}) = \text{tr}(\mathbf{\Sigma}_p\widehat{\mathbf{\Sigma}}^{-1}) - \log\det(\mathbf{\Sigma}_p\widehat{\mathbf{\Sigma}}^{-1}) - p. \tag{3.4.2}$$

The class of rotation-equivariant estimator $\mathbf{\Sigma}(\mathbf{D}) = \mathbf{P}\mathbf{D}\mathbf{P}^{\text{T}}$ minimizes the Frobenius norm exactly at $\mathbf{\Sigma}_{\text{Ideal}}$, while similar to Proposition 2 in [23], $\mathbf{\Sigma}_{\text{Ideal}}$ also minimizes the inverse Stein's loss within such a class of estimator. Hence it is intuitive that our estimator $\widehat{\mathbf{\Sigma}}_p$ will be relatively less efficient in the sense that $EL(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}}_p) > 0$. It turns out that asymptotically, $\widehat{\mathbf{\Sigma}}_p$ is doing as good as $\mathbf{\Sigma}_{\text{Ideal}}$, as shown in the Theorem 4 below. To present this theorem, we need to make two more assumptions:

Assumption 3.4.4 Let $v_{n,1} \geq \cdots \geq v_{n,p}$ be the p eigenvalues of $\mathbf{\Phi}$. Define $H_n(v) = p^{-1}\sum_{i=1}^p \mathbf{1}_{\{v_{n,i}\leq v\}}$ the empirical distribution function of the population eigenvalues. We assume $H_n(v)$ converges to some non-random limit H at every point of continuity of H.

Assumption 3.4.5 The support of H defined above is the union of a finite number of compact intervals bounded away from zero and infinity. Also, there exists a compact interval in $(0, +\infty)$ that contains the support of H_n for each n.

These two assumptions are essentially Assumptions (A3) and (A4) in [23] applied on $\mathbf{\Phi}$.

Theorem 4 *Let all the assumptions in Lemma 2 hold, together with Assumption 3.4.4 and 3.4.5. Then as $p, n \to \infty$ such that $p/n \to c > 0$, we have $EL(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}}_p) \leq 0$ almost surely with respect to both the Frobenius and the inverse Stein's loss functions, as long as $p^{-1}L(\mathbf{\Sigma}_p, \mathbf{\Sigma}_{Ideal})$ does not tend to 0 almost surely.*

The requirement $p^{-1}L(\mathbf{\Sigma}_p, \mathbf{\Sigma}_{\text{Ideal}})$ not going to 0 almost surely eliminates the case $\mathbf{\Sigma}_p = \int_0^1 \gamma_t^2 dt \cdot \mathbf{I}_p$, when both the loss functions will attain 0 for the the ideal estimator. Our estimator will still do a good job in such a case since $\text{tr}(\mathbf{\Sigma}_p^{\text{RCV}})/p$ will still be a good estimator for $\int_0^1 \gamma_t^2 dt$ by the proof of Theorem 3, while $\widehat{\mathbf{\Phi}}$ can still do a fine job when permutation of the data is allowed as demonstrated in the simulation results in [23]. Improvement by averaging and permutation will be described in Section 3.4.1.

3.4.1 Practical Implementation

Following Assumption 3.3.1, $\Delta\mathbf{X}_\ell \Delta\mathbf{X}_\ell^{\text{T}} / \left\|\Delta\mathbf{X}_\ell\right\|^2$ is independent of γ_t and is similar to a data point in constructing a sample covariance matrix, which is independent of each others for different ℓ; see Remark 3.3.2 in Section 3.3.1. This observation permits us to permute the data beforehand, say at the jth permutation, we form a data matrix $\Delta\mathbf{X}^{(j)} = (\Delta\mathbf{X}_1^{(j)}, \Delta\mathbf{X}_2^{(j)})$, with $\Delta\mathbf{X}_i^{(j)}$ having size $p \times n_i$ for $i = 1, 2$, such that $n = n_1 + n_2$. Then we construct

$$\widetilde{\mathbf{\Phi}}_i^{(j)} = \frac{p}{n_i} \sum_{\ell \in I_i^{(j)}} \frac{\Delta\mathbf{X}_\ell \Delta\mathbf{X}_\ell^{\text{T}}}{\left\|\Delta\mathbf{X}_\ell\right\|^2}, \tag{3.4.3}$$

where $I_i^{(j)} = \{\ell : \Delta\mathbf{X}_\ell \in \Delta\mathbf{X}_i^{(j)}\}$, and perform eigen-analysis on $\widetilde{\mathbf{\Phi}}_1^{(j)}$, say $\widetilde{\mathbf{\Phi}}_1^{(j)} = \mathbf{P}_1^{(j)} \mathbf{D}_1^{(j)} \mathbf{P}_1^{(j)\text{T}}$. The we can form the jth estimator as

$$\widehat{\mathbf{\Sigma}}_p^{(j)} := \frac{\text{tr}(\mathbf{\Sigma}_p^{\text{RCV}})}{p} \widehat{\mathbf{\Phi}}^{(j)}, \quad \text{where } \widehat{\mathbf{\Phi}}^{(j)} := \mathbf{P}_1^{(j)} \text{diag}(\mathbf{P}_1^{(j)\text{T}} \widetilde{\mathbf{\Phi}}_2^{(j)} \mathbf{P}_1^{(j)}) \mathbf{P}_1^{(j)\text{T}}. \tag{3.4.4}$$

If we perform M permutations and get M estimators as above, we can define the averaged estimator as

$$\widehat{\mathbf{\Sigma}}_{p,M} := \frac{1}{M} \sum_{j=1}^{M} \widehat{\mathbf{\Sigma}}_p^{(j)}. \tag{3.4.5}$$

Note that in all M estimators, we are only using one split location, n_1, for the data, instead of using several of them and then average the results similar to the grand average estimator in [1]. To find the best split location empirically, we minimize the following function:

$$g(m) = \left\| \frac{1}{M} \sum_{j=1}^{M} (\widehat{\mathbf{\Phi}}_p^{(j)} - \widetilde{\mathbf{\Phi}}_2^{(j)}) \right\|_F^2, \tag{3.4.6}$$

where $\widetilde{\mathbf{\Phi}}_2^{(j)}$ is defined in (3.4.3) and $\widehat{\mathbf{\Phi}}^{(j)}$ in (3.4.4). A very similar function is also used to determine the split location for nonlinear shrinkage of a covariance matrix in [23].

We now show that the averaged estimator defined in (3.4.5) also enjoys good asymptotic efficiency.

Theorem 5 *Let all the assumptions in Lemma 2 hold, together with Assumption 3.4.4 and 3.4.5. Suppose the number of permutations M is finite in (3.4.5). Then as $p, n \to \infty$ such that $p/n \to c > 0$, we have $EL(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}}_{p,M}) \leq 0$ almost surely with respect to both the Frobenius and the inverse Stein's loss functions, as long as $p^{-1}L(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}}_{Ideal})$ does not tend to 0 almost surely.*

In practice, the estimator $\widehat{\mathbf{\Sigma}}_{p,M}$, with a good choice of m, performs much better than using just $M = 1$. We use $M = 50$ which provides a good trade-off between computational complexity and estimation accuracy with respect to the Frobenius or the inverse Stein's loss functions. For minimizing $g(m)$ defined in (3.4.6), we search the following split locations:

$$m = [2n^{1/2}, 0.2n, 0.4n, 0.6n, 0.8n, n - 2.5n^{1/2}, n - 1.5n^{1/2}].$$

Except for the case $\mathbf{\Sigma}_p = \int_0^1 \gamma_t^2 dt \mathbf{I}_p$ which needs n to be as small as possible (see the arguments provided in [23]), the two split locations $[n - 2.5n^{1/2}]$ and $[n - 1.5n^{1/2}]$ are those satisfying the condition $\sum_{n_2 \geq 1} p n_2^{-5} < \infty$ needed in all theorems presented when $p/n \to c > 0$. We include $0.2n$ to $0.8n$ for accommodating finite sample performance.

3.5 Empirical Results

We carry out simulation studies to compare the performances of our estimator in (3.3.7), the time variation-adjusted realized covariance matrix in (3.3.5) and the realized covariance matrix in (3.3.3) by comparing their Frobenius and inverse Stein's losses defined in (3.4.1) and (3.4.2) respectively. Then in Section 3.5.1, we consider a trading exercise using simulated market data and compare the risks associated with the minimum variance portfolios constructed using these three different estimators. Finally, in Section 3.5.2, we consider real data from the New York Stock Exchange.

Consider two different scenarios for the diffusion process $\{\mathbf{X}_t\}$ defined in (3.3.1), with $\boldsymbol{\mu}_t = \mathbf{0}$ and $\mathbf{\Theta}_t = \gamma_t \mathbf{\Lambda}$ as in (3.3.4). One has γ_t being piecewise constant, the other has γ_t being continuous, detailed as follows:

Design I: Piecewise constants. We take γ_t to be

$$\gamma_t = \begin{cases} \sqrt{0.0007}, & t \in [0, 1/4) \cup [3/4, 1], \\ \sqrt{0.0001}, & t \in [1/4, 3/4). \end{cases}$$

Design II: Continuous path. We take γ_t to be

$$\gamma_t = \sqrt{0.0009 + 0.0008\cos(2\pi t)}, t \in [0, 1].$$

We assume $\mathbf{\Lambda} = \mathbf{I}_p$ and the observation times are taken to be equidistant, where $\tau_{n,\ell} = \ell/n, \ell = 1, \ldots, n$. We generate $\{\mathbf{X}_t\}$ using model (3.3.1) and get $n = 200$ discrete observations, and consider $p = 100, 200$. For each design and each (n, p) combination, we repeat 1000 times the simulations, and compare the mean Frobenius and inverse Stein's losses for our proposed estimator, the time variation-adjusted realized covariance matrix and the realized covariance matrix.

Table 3.1 presents the simulation results. It is clear that overall, our proposed estimator performs the best. In particular, since the realized covariance or the time variation-adjusted realized covariance matrices are singular when $p = 200$, their inverses do not exist. In contrast, our proposed estimator is always non-singular and stable even in this case, which is in line with Theorem 3.

3.5.1 A market trading exercise

As an application in finance, we simulate market trading data in this section and construct minimum variance portfolio using the three different estimators compared in the previous section. Given an integrated covariance matrix $\mathbf{\Sigma}_p$, the minimum variance portfolio solves $\min_{\mathbf{w}:\mathbf{w}^{\mathrm{T}}\mathbf{1}_p=1} \mathbf{w}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{w}$, where $\mathbf{1}_p$ is a vector of p ones. The solution to the above is given by

$$\mathbf{w}_{\mathrm{opt}} = \frac{\mathbf{\Sigma}_p^{-1}\mathbf{1}_p}{\mathbf{1}_p^{\mathrm{T}}\mathbf{\Sigma}_p^{-1}\mathbf{1}_p}. \tag{3.5.1}$$

Before presenting our simulation settings, we present a theorem concerning the minimum variance portfolio (3.5.1) constructed using our integrated covariance matrix estimator $\widehat{\mathbf{\Sigma}}_{p,M}$. In the sequel, we denote $\|\cdot\|_{\max}$ the maximum absolute value of a vector, and define the condition number of a positive semi-definite matrix $\mathbf{A}$ to be $\mathrm{Cond}(\mathbf{A}) = \lambda_{\max}(\mathbf{A})/\lambda_{\min}(\mathbf{A})$.

Theorem 6 *Let all the assumptions in Lemma 2 hold. Suppose the number of permutations M is finite in (3.4.5). Then as $p, n \to \infty$ such that $p/n \to c > 0$, almost surely,*

$$p^{1/2}\|\widehat{\mathbf{w}}_{\mathrm{opt}}\|_{\max} \leq \mathrm{Cond}(\mathbf{\Phi}), \quad pR(\widehat{\mathbf{w}}_{\mathrm{opt}}) \leq \mathrm{Cond}^2(\mathbf{\Phi})\lambda_{\max}(\mathbf{\Sigma}_p),$$

Design I	Proposed	Time variation-adjusted	Realized covariance
	Frobenius loss		
$p = 100$	$.13_{(.02)}$	$2.8_{(.04)}$	$3.6_{(.06)}$
$p = 200$	$.55_{(.17)}$	$69_{(31)}$	$1564_{(63)}$
	Inverse Stein's loss		
$p = 100$	$.17_{(.014)}$	$5.63_{(.058)}$	$7.08_{(.08)}$
$p = 200$	$88_{(15)}$	-	-

Design II	Proposed	Time variation-adjusted	Realized covariance
	Frobenius loss		
$p = 100$	$.29_{(.03)}$	$6.32_{(.09)}$	$7.55_{(.1)}$
$p = 200$	$.38_{(.03)}$	$13_{(10)}$	$15_{(20)}$
	Inverse Stein's loss		
$p = 100$	$.54_{(.1)}$	$693_{(31)}$	$1232_{(53.8)}$
$p = 200$	$88_{(16)}$	-	-

Table 3.1: Mean and standard deviation (in bracket) of losses for different methods. All values reported in this table are multiplied by 1000. Upper table: results for Design I. Lower table: results for Design II. For $p = 200$, the time variation-adjusted and realized covariance matrices are always singular, and hence inverse Stein's loss are at infinity.

where $\widehat{\mathbf{w}}_{\text{opt}}$ *is the weight in (3.5.1) with* $\mathbf{\Sigma}_p$ *substituted by* $\widehat{\mathbf{\Sigma}}_{p,M}$ *defined in (3.4.5). The function* $R(\widehat{\mathbf{w}}_{\text{opt}}) = \widehat{\mathbf{w}}_{\text{opt}}^{\text{T}} \mathbf{\Sigma}_p \widehat{\mathbf{w}}_{\text{opt}}$ *represents the actual risk when investing using* $\widehat{\mathbf{w}}_{\text{opt}}$ *as the portfolio weights.*

This theorem shows that the maximum absolute weight, we name it the maximum exposure of the portfolio, is decaying at a rate $p^{-1/2}$, while the actual risk is decaying at a rate p^{-1}. It is an important quality of our estimator when applied to portfolio allocation, since we do not want to over-invest into a single asset when there are many other assets for risk diversification. Such over-investment can result in huge loss if the involved asset suddenly drops in price due to random events over the investment period. At the same time, the actual risk of the portfolio is decaying linearly as we have more and more assets for diversification. We demonstrate the results in Theorem 6 in Table 3.2 and Section 3.5.2.

For simulating the price data, following [9] and [17], we simulate $p = 100$ stock prices for 200 days using $X_t^{o(i)} = X_t^{(i)} + \epsilon_t^{(i)}$, where $X_t^{(i)}$ is the underlying log-price, and $\epsilon_t^{(i)}$ models the market microstructure noise, with $\epsilon_t^{(i)} \sim N(0, 0.0005^2)$ and are assumed to be independent of each other. The underlying log-price $X_t^{(i)}$ is generated by the stochastic volatility model. For $i = 1, \ldots, 100$,

$$dX_t^{(i)} = \mu^{(i)} dt + \rho^{(i)} \sigma_t^{(i)} dB_t^{(i)} + \sqrt{1 - (\rho^{(i)})^2} \sigma_t^{(i)} dW_t + \nu^{(i)} dZ_t,$$

where $\{W_t\}$, $\{Z_t\}$ and the $\{B_t^{(i)}\}$'s are all independent standard Brownian motions. The process $\{Z_t\}$ plays the role of a pervasive factor, which is usually the market factor in asset returns. The spot volatility $\sigma_t^{(i)} = \exp(\varrho_t^{(i)})$ follows the independent Ornstein-Uhlenbeck process

$$d\varrho_t^{(i)} = \alpha^{(i)}(\beta_0^{(i)} - \varrho_t^{(i)})dt + \beta_1^{(i)} dU_t^{(i)},$$

where the $\{U_t^{(i)}\}$'s are independent standard Brownian motions. We use $(\mu^{(i)}, \beta_0^{(i)}, \beta_1^{(i)}, \alpha^{(i)}, \rho^{(i)}) = (0.03x_1^{(i)}, -x_2^{(i)}, 0.75x_3^{(i)}, -1/40x_4^{(i)}, -0.7)$ and $\nu^{(i)} = exp(\beta_0^{(i)})$, where the $x_j^{(i)}$'s are independent and uniformly distributed on the interval $[0.7, 1.3]$. The initial value of each log-price is set at $X_0^{(i)} = 1$ and the starting spot volatility $\varrho_0^{(i)} = 0$.

We simulate the trading times independently from the price data assuming the transaction times for each stock follow independent Poisson processes with rates $\lambda_1, \cdots, \lambda_{100}$ respectively, where $\lambda_i = 0.01i \times 23400$. We set this because a normal trading time for one day is 23400 seconds.

After simulating the data, we split a trading day into 15-minute intervals and set the price data for each stock at the end of each interval as the price observed at the trade right

Theoretical risk	.735		
	Proposed	Time variation-adjusted	Realized covariance
Actual risk	.922	3.918	4.115
Perceived risk	.753	3.869	4.034

Table 3.2: Mean of theoretical risk $R(\mathbf{w}_{\text{opt}})$, actual risk $R(\widehat{\mathbf{w}}_{\text{opt}})$ and perceived risk $\widehat{R}(\widehat{\mathbf{w}}_{\text{opt}})$.

before the end of the interval. The data is used to calculate various integrated covariance estimators, including our proposed one.

At the start, we invest 1 unit of capital using the minimum variance allocations (3.5.1) constructed from using different estimators of the integrated covariance matrix. Each time we use a 60-day training window (so the first trade starts on day 61, and the last one on day 200) and we re-evaluate our portfolio weights every 5 days, using the past 60 days of data as a training set until we reach day 195.

In Table 3.2, we report the mean of three risks. The first one is the theoretical risk $R(\mathbf{w}_{\text{opt}}) = \mathbf{w}_{\text{opt}}^{\text{T}}\boldsymbol{\Sigma}_p\mathbf{w}_{\text{opt}}$, where $\mathbf{w}_{\text{opt}}$ is calculated as in (3.5.1) using the true integrated covariance matrix of the underlying log-return over the past 60-day training period and $\boldsymbol{\Sigma}_p$ is the true integrated covariance matrix over the 5-day investment period. The second one is the actual risk $R(\widehat{\mathbf{w}}_{\text{opt}}) = \widehat{\mathbf{w}}_{\text{opt}}^{\text{T}}\boldsymbol{\Sigma}_p\widehat{\mathbf{w}}_{\text{opt}}$, where $\widehat{\mathbf{w}}_{\text{opt}}$ is calculated using different integrated covariance matrix estimators. Finally the perceived risk is defined by $\widehat{R}(\widehat{\mathbf{w}}_{\text{opt}}) = \widehat{\mathbf{w}}_{\text{opt}}^{\text{T}}\widehat{\boldsymbol{\Sigma}}_p\widehat{\mathbf{w}}_{\text{opt}}$.

We can see from Table 3.2 that our method has the best performance among all three different methods, and has the risk closest to the theoretical one. In particular, our method has the smallest actual risk, which is the most relevant risk in practice. Such a small actual risk for our method is also consistent with the actual risk bound in Theorem 6, when the number of assets is $p = 100$ which is relatively large.

3.5.2 Portfolio allocation on NYSE data

We consider $p = 45$ stocks from the New York Stock Exchange from January 1 of 2013 to December 31 of 2013 (245 trading days). We choose the stocks from mid-cap energy sector stocks. We downloaded all the trades of these stocks from Wharton Research Data Services (WRDS, https://wrds-web.wharton.upenn.edu/). The raw data are of high-frequency nature. As mentioned before, the stocks have non-synchronous trading times and all the

log-prices are contaminated by market microstructure noise.

Like the market trading exercise in Section 3.5.1, we consider trades in 15-minute intervals on every trading day from 9:30 to 16:00, with each log-price being the observed one from a trade right before a 15-minute interval ends. This results in a total of 6732 observations over the 245 trading days. Hence on average there are around 27 observations per day.

We consider two settings. For the first one, we consider 20-day training windows and re-evaluate portfolio weights every 5 days. Another setting use 5-day training windows and re-evaluate portfolio weights everyday. We use the annualized out-of-sample standard deviation $\widehat{\sigma}$, together with the annualized portfolio return $\widehat{\mu}$ and the Sharpe ratio $\widehat{\mu}/\widehat{\sigma}$ to gauge the performance of each method. For 20-day training windows and 5 day re-evaluation period, $\widehat{\mu}$ and $\widehat{\sigma}$ are defined by

$$\widehat{\mu} = 52 \times \frac{1}{45} \sum_{i=5}^{49} \mathbf{w}_i^{\mathrm{T}} \mathbf{r}_i, \ \widehat{\sigma} = \Big(52 \times \frac{1}{45} \sum_{i=5}^{49} (\mathbf{w}_i^{\mathrm{T}} \mathbf{r}_i - \widehat{\mu})^2 \Big)^{1/2}.$$

We use the annualized out-of-sample standard deviation since we do not know the true underlying integrated covariance matrix, and hence the actual risk cannot be calculated. For 5-day training windows with daily re-evaluation of portfolio weights, $\widehat{\mu}$ and $\widehat{\sigma}$ are defined by

$$\widehat{\mu} = 252 \times \frac{1}{240} \sum_{i=6}^{245} \mathbf{w}_i^{\mathrm{T}} \mathbf{r}_i, \ \widehat{\sigma} = \Big(252 \times \frac{1}{240} \sum_{i=6}^{245} (\mathbf{w}_i^{\mathrm{T}} \mathbf{r}_i - \widehat{\mu})^2 \Big)^{1/2}.$$

On top of all the above, we also report the mean maximum exposure $\big\|\widehat{\mathbf{w}}_{\mathrm{opt}}\big\|_{\max}$ over all investment periods for the portfolios constructed under different methods.

Table 3.3 shows the results. For both settings, we see that the annualized out-of-sample standard deviation is the smallest for our method. The difference between different methods for the setting with 20-day training windows is less pronounced. This is because for a 20-day window, we have on average $27 \times 20 = 540$ data points for calculating different integrated covariance estimators. The time variation-adjusted and realized covariance matrices will not be too much disadvantaged since the bias problem from having high dimension is not too serious as we have $p/n = 45/540 = 0.083$. On the other hand, the second setting has on average $27 \times 5 = 135$ data points and so $p/n = 45/135 = 0.33$, so that the bias from having high dimension for the two estimators is much more serious than the first case. Our estimator has clearly done a good job in minimizing the risk in both situations. It would seem that we can solve the problem simply by using a longer training window. However, the problem of using a longer training window is that the true integrated covariance matrix

20-day training window Re-evaluate every 5 days	Proposed (%)	Time variation adjusted(%)	Realized covariance(%)
Annualized return	5.50	5.87	10.59
Annualized out-of-sample SD	11.07	11.16	12.67
Sharpe ratio	49.66	52.62	83.59
Mean maximum exposure	$15.19_{(3.26)}$	$19.82_{(4.74)}$	$28.86_{(11.05)}$

5-day training window Re-evaluate everyday	Proposed (%)	Time variation adjusted(%)	Realized covariance(%)
Annualized return	7.40	8.25	7.39
Annualized out-of-sample SD	10.78	11.66	13.89
Sharpe ratio	68.69	73.92	53.19
Mean maximum exposure	$11.47_{(2.19)}$	$24.06_{(7.11)}$	$31.27_{(18.76)}$

Table 3.3: Percentage annualized return, out-of-sample standard deviation, Sharpe ratio and mean maximum exposure (standard deviation in bracket) for different methods. Upper table: Use 20-day training windows and re-evaluate portfolio weights every 5 days. Lower table: Use 5-day training windows and re-evaluate portfolio weights everyday.

from the training period will likely to be different from the investment period ahead if such a window is too long.

Observe also the mean maximum exposure for our method is always smaller than other methods, with smaller standard deviations. For the case of 5-day training window, the differences in maximum exposure among the three different methods are huge, with our method having this at 11.5% which is reasonable, and compatible with the results in Theorem 6. At around 25% or even 30% for the other two methods, there is a much bigger risk of suffering a loss due to random events from the heavily invested individual assets.

3.6 Summary of this chapter

In this chapter, we first review the relevant estmiation methods in the literature, and then introduce a novel nonlinear shrinkage estimator for the integrated volatility matrix. The properties of our estimation methods including that the estimator shrinks the extreme eigenvalues of a realized covariance matrix back to acceptable level, and enjoys a certain asymptotic efficiency at the same time, all at a high dimensional setting where the number of assets can have the same order as the number of data points. By using some numerical examples and real data from NYSE trading data, we demostrated that our estimator has a favorable performance compared to a time-variation adjusted realized covariance estimator and the usual realized covariance matrix. This includes a novel maximum exposure bound and an actual risk bound when our estimator is used in constructing the minimum variance portfolio.

3.7 Proofs for this chapter

Proof of Lemma 2. Using Remark 3.3.2, we can write

$$\Delta \mathbf{X}_\ell = \Big(\int_{\tau_{n,\ell-1}}^{\tau_{n,\ell}} \gamma_t^2 dt \Big)^{1/2} \mathbf{\Phi}^{1/2} \mathbf{Z}_\ell,$$

where the $\mathbf{Z}_\ell$'s are independent of each other and each $\mathbf{Z}_\ell$ is a p dimensional vector with independent standard normal entries. Then we can decompose

$$\frac{\mathbf{p}_{1i}^{\mathrm{T}} \widetilde{\mathbf{\Phi}} \mathbf{p}_{1i} - \mathbf{p}_{1i}^{\mathrm{T}} \mathbf{\Phi} \mathbf{p}_{1i}}{\mathbf{p}_{1i}^{\mathrm{T}} \mathbf{\Phi} \mathbf{p}_{1i}} = I_{i1} + I_{i2},$$

where

$$I_{i1} = \frac{\mathbf{p}_{1i}^{\mathrm{T}} \widetilde{\mathbf{\Phi}}_2 \mathbf{p}_{1i} - n_2^{-1} \sum_{\ell \in I_2} (\mathbf{p}_{1i}^{\mathrm{T}} \mathbf{\Phi}^{1/2} \mathbf{Z}_\ell)^2}{\mathbf{p}_{1i}^{\mathrm{T}} \mathbf{\Phi} \mathbf{p}_{1i}}, \quad I_{i2} = \frac{n_2^{-1} \sum_{\ell \in I_2} (\mathbf{p}_{1i}^{\mathrm{T}} \mathbf{\Phi}^{1/2} \mathbf{Z}_\ell)^2 - \mathbf{p}_{1i}^{\mathrm{T}} \mathbf{\Phi} \mathbf{p}_{1i}}{\mathbf{p}_{1i}^{\mathrm{T}} \mathbf{\Phi} \mathbf{p}_{1i}}.$$

First of all, since we can write

$$n_2^{-1}\sum_{\ell\in I_2}(\mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Phi}^{1/2}\mathbf{Z}_\ell)^2=\mathbf{p}_{1i}^{\mathrm{T}}\big(n_2^{-1}\sum_{\ell\in I_2}\mathbf{\Phi}^{1/2}\mathbf{Z}_\ell\mathbf{Z}_\ell^{\mathrm{T}}\mathbf{\Phi}^{1/2}\big)\mathbf{p}_{1i},$$

with $n_2^{-1}\sum_{\ell\in I_2}\mathbf{\Phi}^{1/2}\mathbf{Z}_\ell\mathbf{Z}_\ell^{\mathrm{T}}\mathbf{\Phi}^{1/2}$ a proper sample covariance matrix for estimating $\mathbf{\Phi}$, we can use Lemma 1 of [23] to conclude that

$$\max_{1\le i\le p}|I_{i2}|\overset{a.s.}{\to}0.$$

Hence it remains to show that $\max_{1\le i\le p}|I_{i1}|\overset{a.s.}{\to}0$. To this end, consider

$$\begin{aligned}\max_{1\le i\le p}|I_{i1}|&=\max_{1\le i\le p}\left|\frac{n_2^{-1}\sum_{\ell\in I_2}\frac{(\mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Phi}^{1/2}\mathbf{Z}_\ell)^2}{\mathbf{Z}_\ell^{\mathrm{T}}\mathbf{\Phi}\mathbf{Z}_\ell/p}-\mathbf{p}_{1i}^{\mathrm{T}}\big(n_2^{-1}\sum_{\ell\in I_2}\mathbf{\Phi}^{1/2}\mathbf{Z}_\ell\mathbf{Z}_\ell^{\mathrm{T}}\mathbf{\Phi}^{1/2}\big)\mathbf{p}_{1i}}{\mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Phi}\mathbf{p}_{1i}}\right|\\&=\max_{1\le i\le p}\left|\frac{n_2^{-1}\sum_{\ell\in I_2}\big(\frac{1}{\mathbf{Z}_\ell^{\mathrm{T}}\mathbf{\Phi}\mathbf{Z}_\ell/p}-1\big)\mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Phi}^{1/2}\mathbf{Z}_\ell\mathbf{Z}_\ell^{\mathrm{T}}\mathbf{\Phi}^{1/2}\mathbf{p}_{1i}}{\mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Phi}\mathbf{p}_{1i}}\right|\\&\le\max_{\ell\in I_2}\left|\frac{1}{\mathbf{Z}_\ell^{\mathrm{T}}\mathbf{\Phi}\mathbf{Z}_\ell/p}-1\right|\cdot(1+\max_{1\le i\le p}|I_{i2}|)\overset{a.s.}{\to}0,\end{aligned}$$

if we can show further that $\max_{\ell\in I_2}\left|\frac{1}{\mathbf{Z}_\ell^{\mathrm{T}}\mathbf{\Phi}\mathbf{Z}_\ell/p}-1\right|\overset{a.s.}{\to}0$.

To show this, using Lemma 2.7 of [8], we have

$$E(\mathbf{Z}_\ell^{\mathrm{T}}\mathbf{\Phi}\mathbf{Z}_\ell-\mathrm{tr}(\mathbf{\Phi}))^6\le K_6(E^3|z_{\ell,1}|^4\mathrm{tr}^3(\mathbf{\Phi}^2)+E|z_{\ell,1}|^{12}\mathrm{tr}(\mathbf{\Phi}^6)),$$

where K_6 is a constant independent of ℓ, n and p. This implies that, since $\mathrm{tr}(\mathbf{\Phi})=p$,

$$\begin{aligned}E\bigg(\max_{\ell\in I_2}\left|\frac{\mathbf{Z}_\ell^{\mathrm{T}}\mathbf{\Phi}\mathbf{Z}_\ell}{p}-1\right|^6\bigg)&\le n_2\cdot K_6\Big(E^3|z_{\ell,1}|^4\frac{\mathrm{tr}^3(\mathbf{\Phi}^2)}{p^6}+E|z_{\ell,1}|^{12}\frac{\mathrm{tr}(\mathbf{\Phi}^6)}{p^6}\Big)\\&=O(n_2p^{-3}).\end{aligned}$$

The rate in the last line comes from Assumption 3.4.1 that $\mathbf{\Theta}_t\mathbf{\Theta}_t^{\mathrm{T}}=\gamma_t^2\mathbf{\Phi}$ has all its eigenvalues uniformly bounded away from 0 and infinity, so that $\mathrm{tr}^3(\mathbf{\Phi}^2)=O(p^3)$ and $\mathrm{tr}(\mathbf{\Phi}^6)=O(p)$, and the fact that the higher order moments of the $z_{\ell,1}$'s are all finite since they are all normally distributed.

Finally, since $n_2=O(n^{1/2})$ and p has the same order as n, we have $O(n_2p^{-3})=O(n^{-5/2})$. Since $\sum_{n\ge1}n^{-5/2}<\infty$, through the Borel-Cantelli lemma, we have proved that $\max_{\ell\in I_2}\left|\frac{\mathbf{Z}_\ell^{\mathrm{T}}\mathbf{\Phi}\mathbf{Z}_\ell}{p}-1\right|\overset{a.s.}{\to}0$, meaning that $\max_{\ell\in I_2}\left|\frac{1}{\mathbf{Z}_\ell^{\mathrm{T}}\mathbf{\Phi}\mathbf{Z}_\ell/p}-1\right|\overset{a.s.}{\to}0$ as well. This completes the proof of the lemma. □

Proof of Theorem 3. Firstly, by Lemma 2, $\widehat{\mathbf{\Phi}}$ defined in (3.3.7) is almost surely positive definite since all its eigenvalues are almost surely $\mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Phi}\mathbf{p}_{1i}$ for some i as $n,p\to\infty$ such

that $p/n \to c > 0$, and Assumption 3.3.1 ensures that these values are uniformly bounded away from 0 and infinity. Hence, the proof of the theorem is complete if we can show that $\mathrm{tr}(\mathbf{\Sigma}_p^{\mathrm{RCV}})/p > 0$ uniformly almost surely.

To this end, consider

$$\begin{aligned}\left|\frac{\mathrm{tr}(\mathbf{\Sigma}_p^{\mathrm{RCV}})}{p} - \int_0^1 \gamma_t^2 dt\right| &= \left|\frac{\sum_{\ell=1}^n \Delta\mathbf{X}_\ell^{\mathrm{T}}\Delta\mathbf{X}_\ell}{p} - \int_0^1 \gamma_t^2 dt\right| \\ &= \left|\frac{\sum_{\ell=1}^n \int_{\tau_{n,\ell-1}}^{\tau_{n,\ell}} \gamma_t^2 dt \cdot \mathbf{Z}_\ell^{\mathrm{T}}\mathbf{\Phi}\mathbf{Z}_\ell}{p} - \int_0^1 \gamma_t^2 dt\right| \\ &\le \max_{1\le\ell\le n}\left|\frac{\mathbf{Z}_\ell^{\mathrm{T}}\mathbf{\Phi}\mathbf{Z}_\ell}{p} - 1\right| \int_0^1 \gamma_t^2 dt.\end{aligned}$$

From the proof of $\max_{\ell\in I_2}\left|\frac{\mathbf{Z}_\ell^{\mathrm{T}}\mathbf{\Phi}\mathbf{Z}_\ell}{p} - 1\right| \xrightarrow{a.s.} 0$ in the last part of the proof of Lemma 2, we can replace n_2 there by n and conclude that

$$\begin{aligned}E\left(\max_{1\le\ell\le n}\left|\frac{\mathbf{Z}_\ell^{\mathrm{T}}\mathbf{\Phi}\mathbf{Z}_\ell}{p} - 1\right|^6\right) &\le n \cdot K_6\left(E^3|z_{\ell,1}|^4\frac{\mathrm{tr}^3(\mathbf{\Phi}^2)}{p^6} + E|z_{\ell,1}|^{12}\frac{\mathrm{tr}(\mathbf{\Phi}^6)}{p^6}\right) \\ &= O(np^{-3}) = O(n^{-2}).\end{aligned}$$

Since $\sum_{n\ge1} n^{-2} < \infty$, we have proved that $\max_{1\le\ell\le n}\left|\frac{\mathbf{Z}_\ell^{\mathrm{T}}\mathbf{\Phi}\mathbf{Z}_\ell}{p} - 1\right| \xrightarrow{a.s.} 0$. This shows that $\mathrm{tr}(\mathbf{\Sigma}_p^{\mathrm{RCV}})/p \xrightarrow{a.s.} \int_0^1 \gamma_t^2 dt$, which is uniformly larger than 0 by Assumption 3.3.1. This completes the proof of the theorem. □

To prove Theorem 4, we first present the following lemma and its proof.

Lemma 3.7.1 *Let all the assumptions in Lemma 2 hold, together with Assumption 3.4.4 and 3.4.5. Denote by $v_1^{(1)} \ge \cdots \ge v_p^{(1)}$ the eigenvalues of $\widetilde{\mathbf{\Phi}}_1$ defined in (3.3.6) with corresponding eigenvectors $\mathbf{p}_{11}, \ldots, \mathbf{p}_{1p}$, and $v_1 \ge \cdots \ge v_p$ the eigenvalues of $\check{\mathbf{\Phi}}$ defined in (3.3.5) with corresponding eigenvectors $\mathbf{p}_1, \ldots, \mathbf{p}_p$. Then there exist positive functions $\delta_1(\cdot) = \delta(\cdot)$* [3]

and distribution functions $F_1 = F$ such that

$$p^{-1}\sum_{j=1}^p \mathbf{1}_{\{x\ge v_j^{(1)}\}} \xrightarrow{a.s.} F_1(x), \quad p^{-1}\sum_{j=1}^p \mathbf{1}_{\{x\ge v_j\}} \xrightarrow{a.s.} F(x),$$

[3]The explicit form of the functions $\delta_1(\cdot)$ and $\delta(\cdot)$ are given here, since they are not important for the proof of any subsequent theorems. For any $\lambda \in \mathbf{R}$, $\delta(\lambda) = \begin{cases} \frac{\lambda}{|1-c-c\lambda\breve{m}_F(\lambda)|^2}, & \text{if } \lambda > 0; \\ \frac{1}{(c-1)\breve{m}_F(0)} & \text{if } \lambda = 0 \text{ and } c > 1; \\ 0, & \text{otherwise.} \end{cases}$

$\delta_1(\lambda) = \begin{cases} \frac{\lambda}{|1-c_1-c_1\lambda\breve{m}_{F_1}(\lambda)|^2}, & \text{if } \lambda > 0; \\ \frac{1}{(c_1-1)\breve{m}_{F_1}(0)} & \text{if } \lambda = 0 \text{ and } c > 1; \\ 0, & \text{otherwise.} \end{cases}$

$$p^{-1}\sum_{j=1}^{p}\mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Phi}\mathbf{p}_{1i}\mathbf{1}_{\{x\geq v_j^{(1)}\}}\overset{a.s.}{\to}\int_{-\infty}^{x}\delta_1(\lambda)dF_1(\lambda),\quad p^{-1}\sum_{j=1}^{p}\mathbf{p}_i^{\mathrm{T}}\mathbf{\Phi}\mathbf{p}_i\mathbf{1}_{\{x\geq v_j\}}\overset{a.s.}{\to}\int_{-\infty}^{x}\delta(\lambda)dF(\lambda).$$

Proof of Lemma 3.7.1. Write $\Delta\mathbf{X}_\ell=\big(\int_{\tau_{n,\ell-1}}^{\tau_{n,\ell}}\gamma_t^2dt\big)^{1/2}\mathbf{\Phi}^{1/2}\mathbf{Z}_\ell$ as in Remark 3.3.2. Define for $i=1,2$,

$$\widetilde{\mathbf{\Phi}}_{i,\mathrm{sam}}=n_i^{-1}\sum_{\ell\in I_i}\mathbf{\Phi}^{1/2}\mathbf{Z}_\ell\mathbf{Z}_\ell^{\mathrm{T}}\mathbf{\Phi}^{1/2},\quad \check{\mathbf{\Phi}}_{\mathrm{sam}}=n^{-1}\sum_{\ell=1}^{n}\mathbf{\Phi}^{1/2}\mathbf{Z}_\ell\mathbf{Z}_\ell^{\mathrm{T}}\mathbf{\Phi}^{1/2},$$

which are all proper sample covariance matrices. Let $v_{1,\mathrm{sam}}^{(i)}\geq\cdots\geq v_{p,\mathrm{sam}}^{(i)}$ be the eigenvalues of $\widetilde{\mathbf{\Phi}}_{i,\mathrm{sam}}$ with corresponding eigenvectors $\mathbf{p}_{i1,\mathrm{sam}},\ldots,\mathbf{p}_{ip,\mathrm{sam}}$. Also, let $v_{1,\mathrm{sam}}\geq\cdots v_{p,\mathrm{sam}}$ be the eigenvalues of $\check{\mathbf{\Phi}}_{\mathrm{sam}}$ with corresponding eigenvectors $\mathbf{p}_{1,\mathrm{sam}},\ldots,\mathbf{p}_{p,\mathrm{sam}}$. Suppose we are able to show the following. For any $z\in\mathbb{C}^+$,

$$\begin{aligned}
p^{-1}\mathrm{tr}\big((\widetilde{\mathbf{\Phi}}_{1,\mathrm{sam}}-z\mathbf{I}_p)^{-1}\big)-p^{-1}\mathrm{tr}\big((\widetilde{\mathbf{\Phi}}_1-z\mathbf{I}_p)^{-1}\big)&\overset{a.s.}{\to}0,\\
p^{-1}\mathrm{tr}\big((\check{\mathbf{\Phi}}_{\mathrm{sam}}-z\mathbf{I}_p)^{-1}\big)-p^{-1}\mathrm{tr}\big((\check{\mathbf{\Phi}}-z\mathbf{I}_p)^{-1}\big)&\overset{a.s.}{\to}0,\\
p^{-1}\mathrm{tr}\big((\widetilde{\mathbf{\Phi}}_{1,\mathrm{sam}}-z\mathbf{I}_p)^{-1}\mathbf{\Phi}\big)-p^{-1}\mathrm{tr}\big((\widetilde{\mathbf{\Phi}}_1-z\mathbf{I}_p)^{-1}\mathbf{\Phi}\big)&\overset{a.s.}{\to}0,\\
p^{-1}\mathrm{tr}\big((\check{\mathbf{\Phi}}_{\mathrm{sam}}-z\mathbf{I}_p)^{-1}\mathbf{\Phi}\big)-p^{-1}\mathrm{tr}\big((\check{\mathbf{\Phi}}-z\mathbf{I}_p)^{-1}\mathbf{\Phi}\big)&\overset{a.s.}{\to}0.
\end{aligned}\tag{3.7.1}$$

The above can in fact be written as differences of Stieltjes transforms of certain nondecreasing functions. The differences in their inverse Stieltjes transforms must then converge to 0 almost surely as well, i.e., at the point of continuity x of these nondecreasing functions,

$$\begin{aligned}
p^{-1}\sum_{j=1}^{p}\mathbf{1}_{\{x\geq v_{j,\mathrm{sam}}^{(1)}\}}-p^{-1}\sum_{j=1}^{p}\mathbf{1}_{\{x\geq v_j^{(1)}\}}&\overset{a.s.}{\to}0,\\
p^{-1}\sum_{j=1}^{p}\mathbf{1}_{\{x\geq v_{j,\mathrm{sam}}\}}-p^{-1}\sum_{j=1}^{p}\mathbf{1}_{\{x\geq v_j\}}&\overset{a.s.}{\to}0,\\
p^{-1}\sum_{j=1}^{p}\mathbf{p}_{1j,\mathrm{sam}}^{\mathrm{T}}\mathbf{\Phi}\mathbf{p}_{1j,\mathrm{sam}}\mathbf{1}_{\{x\geq v_{j,\mathrm{sam}}^{(1)}\}}-p^{-1}\sum_{j=1}^{p}\mathbf{p}_{1j}^{\mathrm{T}}\mathbf{\Phi}\mathbf{p}_{1j}\mathbf{1}_{\{x\geq v_j^{(1)}\}}&\overset{a.s.}{\to}0,\\
p^{-1}\sum_{j=1}^{p}\mathbf{p}_{j,\mathrm{sam}}^{\mathrm{T}}\mathbf{\Phi}\mathbf{p}_{j,\mathrm{sam}}\mathbf{1}_{\{x\geq v_{j,\mathrm{sam}}\}}-p^{-1}\sum_{j=1}^{p}\mathbf{p}_j^{\mathrm{T}}\mathbf{\Phi}\mathbf{p}_j\mathbf{1}_{\{x\geq v_j\}}&\overset{a.s.}{\to}0.
\end{aligned}\tag{3.7.2}$$

Interested readers are referred to [23] or [28] for the definitions of Stieltjes transform and its inversion. Theorem 4 of [28] indicates that for all $x\in\mathbb{R}$, there exist functions $\delta_1(\cdot)$ and $\delta(\cdot)$ with corresponding distribution functions F_1 and F such that

$$\begin{aligned}
p^{-1}\sum_{j=1}^{p}\mathbf{1}_{\{\lambda\geq v_{j,\mathrm{sam}}^{(1)}\}}&\overset{a.s.}{\to}F_1(\lambda),\\
p^{-1}\sum_{j=1}^{p}\mathbf{1}_{\{\lambda\geq v_{j,\mathrm{sam}}\}}&\overset{a.s.}{\to}F(\lambda),
\end{aligned}$$

$$p^{-1}\sum_{j=1}^{p}\mathbf{p}_{1j,\mathrm{sam}}^{\mathrm{T}}\mathbf{\Phi}\mathbf{p}_{1j,\mathrm{sam}}\mathbf{1}_{\{x\geq v_{j,\mathrm{sam}}^{(1)}\}}\overset{a.s.}{\to}\int_{-\infty}^{x}\delta_1(\lambda)dF(\lambda),$$
$$p^{-1}\sum_{j=1}^{p}\mathbf{p}_{j,\mathrm{sam}}^{\mathrm{T}}\mathbf{\Phi}\mathbf{p}_{j,\mathrm{sam}}\mathbf{1}_{\{x\geq v_{j,\mathrm{sam}}\}}\overset{a.s.}{\to}\int_{-\infty}^{x}\delta(\lambda)dF(\lambda).$$

At the same time, since p/n_1 and p/n both go to $c>0$, Theorem 4.1 of [7] tells us that the two limits F_1 and F are equal, and since $\delta_1(\cdot)$ and $\delta(\cdot)$ depend on F_1 and F respectively (both depend on the same c also; see equation (2.7) and (2.9) in [23] for more details), we must have $\delta_1(\cdot)=\delta(\cdot)$ also.

With the above, (3.7.2) immediately implies the results we need. Hence it remains to show (3.7.1).

To prove the first and third results of (3.7.1), consider for $k=0,1$,

$$\begin{aligned}p^{-1}\mathrm{tr}\big((\widetilde{\mathbf{\Phi}}_1-z\mathbf{I}_p)^{-1}\mathbf{\Phi}^k\big)&=p^{-1}\mathrm{tr}\big((\mathbf{I}_p-(\widetilde{\mathbf{\Phi}}_{1,\mathrm{sam}}-z\mathbf{I}_p)^{-1}(\widetilde{\mathbf{\Phi}}_{1,\mathrm{sam}}-\widetilde{\mathbf{\Phi}}_1))^{-1}(\widetilde{\mathbf{\Phi}}_{1,\mathrm{sam}}-z\mathbf{I}_p)^{-1}\mathbf{\Phi}^k\big)\\&=p^{-1}\mathrm{tr}\big((\widetilde{\mathbf{\Phi}}_{1,\mathrm{sam}}-z\mathbf{I}_p)^{-1}\mathbf{\Phi}^k\big)+R,\quad\text{with}\\R&=\sum_{j\geq 1}p^{-1}\mathrm{tr}\big([(\widetilde{\mathbf{\Phi}}_{1,\mathrm{sam}}-z\mathbf{I}_p)^{-1}(\widetilde{\mathbf{\Phi}}_{1,\mathrm{sam}}-\widetilde{\mathbf{\Phi}}_1)]^j(\widetilde{\mathbf{\Phi}}_{1,\mathrm{sam}}-z\mathbf{I}_p)^{-1}\mathbf{\Phi}^k\big),\end{aligned}$$

where R comes from a Neumann series expansion. Such an expansion is valid since

$$\begin{aligned}r:=\big\|(\widetilde{\mathbf{\Phi}}_{1,\mathrm{sam}}-z\mathbf{I}_p)^{-1}(\widetilde{\mathbf{\Phi}}_{1,\mathrm{sam}}-\widetilde{\mathbf{\Phi}}_1)\big\|&\leq\big\|(\widetilde{\mathbf{\Phi}}_{1,\mathrm{sam}}-z\mathbf{I}_p)^{-1}\big\|\cdot\big\|\widetilde{\mathbf{\Phi}}_{1,\mathrm{sam}}-\widetilde{\mathbf{\Phi}}_1\big\|\\&\leq\frac{1}{|\mathrm{Im}(z)|}\cdot\max_{\ell\in I_1}\left|\frac{1}{\mathbf{Z}_\ell^{\mathrm{T}}\mathbf{\Phi}\mathbf{Z}_\ell/p}-1\right|\big\|\widetilde{\mathbf{\Phi}}_{1,\mathrm{sam}}\big\|\\&\leq\frac{\big\|\mathbf{\Phi}^{1/2}\big\|^2}{|\mathrm{Im}(z)|}\cdot\max_{\ell\in I_1}\left|\frac{1}{\mathbf{Z}_\ell^{\mathrm{T}}\mathbf{\Phi}\mathbf{Z}_\ell/p}-1\right|\big\|n_1^{-1}\sum_{\ell\in I_1}\mathbf{Z}_\ell\mathbf{Z}_\ell^{\mathrm{T}}\big\|\\&\leq\frac{\big\|\mathbf{\Phi}^{1/2}\big\|^2(1+\sqrt{c})^2}{|\mathrm{Im}(z)|}\cdot\max_{\ell\in I_1}\left|\frac{1}{\mathbf{Z}_\ell^{\mathrm{T}}\mathbf{\Phi}\mathbf{Z}_\ell/p}-1\right|\overset{a.s.}{\to}0,\end{aligned}$$

where we used Lemma S.2 of [24] to conclude $\big\|(\widetilde{\mathbf{\Phi}}_{1,\mathrm{sam}}-z\mathbf{I}_p)^{-1}\big\|\leq 1/|\mathrm{Im}(z)|$, and we used Theorem 5.11 of [7] to conclude that $\big\|n_1^{-1}\sum_{\ell\in I_1}\mathbf{Z}_\ell\mathbf{Z}_\ell^{\mathrm{T}}\big\|\leq(1+\sqrt{c})^2$ almost surely, since p/n_1, like p/n, goes to $c>0$ also. Finally, the term $\max_{\ell\in I_1}\big|\frac{1}{\mathbf{Z}_\ell^{\mathrm{T}}\mathbf{\Phi}\mathbf{Z}_\ell/p}-1\big|\overset{a.s.}{\to}0$ by the last part of the proof of Lemma 2. With this,

$$|R|\leq\sum_{j\geq 1}r^j\big\|\mathbf{\Phi}^k\big\|\cdot\big\|(\widetilde{\mathbf{\Phi}}_{1,\mathrm{sam}}-z\mathbf{I}_p)^{-1}\big\|\leq\frac{r\big\|\mathbf{\Phi}^k\big\|}{(1-r)|\mathrm{Im}(z)|}\overset{a.s.}{\to}0,$$

so that we have proved the first and third results in (3.7.1). For the other two results, the proof follows exactly the same lines as before after replacing $\widetilde{\mathbf{\Phi}}_{1,\mathrm{sam}}$ by $\check{\mathbf{\Phi}}_{\mathrm{sam}}$ and $\widetilde{\mathbf{\Phi}}_1$ by $\check{\mathbf{\Phi}}$. This completes the proof of the lemma. □

Proof of Theorem 4. Define $\theta = \int_0^1 \gamma_t^2 dt$, $\widehat{\theta} = \mathrm{tr}(\boldsymbol{\Sigma}_p^{\mathrm{RCV}})/p$ and $\boldsymbol{\Phi}_{\mathrm{Ideal}} = \mathbf{P}\mathrm{diag}(\mathbf{P}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{P})\mathbf{P}^{\mathrm{T}}$. We can easily see that

$$EL(\boldsymbol{\Sigma}_p, \widehat{\boldsymbol{\Sigma}}_p) \leq 1 - \left(\frac{p^{-1/2}\left\|(\widehat{\theta}-\theta)\widehat{\boldsymbol{\Phi}}\right\|_F}{p^{-1/2}\theta\left\|\boldsymbol{\Phi}_{\mathrm{Ideal}}-\boldsymbol{\Phi}\right\|_F} + \frac{p^{-1/2}\left\|\widehat{\boldsymbol{\Phi}}-\boldsymbol{\Phi}\right\|_F}{p^{-1/2}\left\|\boldsymbol{\Phi}_{\mathrm{Ideal}}-\boldsymbol{\Phi}\right\|_F} \right)^{-2}, \tag{3.7.3}$$

where $p^{-1/2}\left\|(\widehat{\theta}-\theta)\widehat{\boldsymbol{\Phi}}\right\|_F \leq |\widehat{\theta}-\theta| \max_{1\leq i\leq p} \mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\boldsymbol{\Phi}}_2\mathbf{p}_{1i} \overset{a.s.}{\to} 0$ by Lemma 2 and the proof of Theorem 3. Since we are equivalently assuming that $p^{-1/2}\left\|\boldsymbol{\Sigma}_{\mathrm{Ideal}}-\boldsymbol{\Phi}\right\|_F$ is not going to 0, it remains to show that the rightmost term above is going to 1. To this end, observe that

$$\frac{p^{-1}\left\|\widehat{\boldsymbol{\Phi}}-\boldsymbol{\Phi}\right\|_F^2}{p^{-1}\left\|\boldsymbol{\Phi}_{\mathrm{Ideal}}-\boldsymbol{\Phi}\right\|_F^2} = \frac{p^{-1}\sum_{i=1}^p(\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\boldsymbol{\Phi}}_2\mathbf{p}_{1i}-\mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{p}_{1i})^2}{p^{-1}\left\|\mathbf{P}\mathrm{diag}(\mathbf{P}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{P})\mathbf{P}^{\mathrm{T}}-\boldsymbol{\Phi}\right\|_F^2} + \frac{p^{-1}\left\|\mathbf{P}_1\mathrm{diag}(\mathbf{P}_1^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{P}_1)\mathbf{P}_1^{\mathrm{T}}-\boldsymbol{\Phi}\right\|_F^2}{p^{-1}\left\|\mathbf{P}\mathrm{diag}(\mathbf{P}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{P})\mathbf{P}^{\mathrm{T}}-\boldsymbol{\Phi}\right\|_F^2}.$$

By Assumptions 3.4.4 and 3.4.5, and the results of Lemma 3.7.1, we have

$$\begin{aligned} p^{-1}\left\|\mathbf{P}_1\mathrm{diag}(\mathbf{P}_1^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{P}_1)\mathbf{P}_1^{\mathrm{T}}-\boldsymbol{\Phi}\right\|_F^2 &= p^{-1}\mathrm{tr}(\boldsymbol{\Phi}^2) - p^{-1}\sum_{i=1}^p(\mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{p}_{1i})^2 \\ &\overset{a.s.}{\to} \int \tau^2 dH(\tau) - \int \delta_1^2(\lambda)dF_1(\lambda) \\ &= \int \tau^2 dH(\tau) - \int \delta^2(\lambda)dFs(\lambda), \end{aligned}$$

which is non-zero if $\boldsymbol{\Phi} \neq \mathbf{I}_p$. This is also the almost sure limit of $p^{-1}\left\|\boldsymbol{\Phi}_{\mathrm{Ideal}}-\boldsymbol{\Phi}\right\|_F^2$, and hence the rightmost term in (3.7.3) is indeed going to 1 if we can also show that $p^{-1}\sum_{i=1}^p(\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\boldsymbol{\Phi}}_2\mathbf{p}_{1i} - \mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{p}_{1i})^2 \overset{a.s.}{\to} 0$. By Lemma 2,

$$p^{-1}\sum_{i=1}^p(\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\boldsymbol{\Phi}}_2\mathbf{p}_{1i} - \mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{p}_{1i})^2 \leq \max_{1\leq i\leq p}\left|\frac{\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\boldsymbol{\Phi}}_2\mathbf{p}_{1i} - \mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{p}_{1i}}{\mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{p}_{1i}}\right| \cdot \max_{1\leq i\leq p}\mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{p}_{1i} \overset{a.s.}{\to} 0.$$

This completes the proof for the Frobenius loss.

For the inverse Stein's loss, by Lemma 3.7.1, we have

$$\begin{aligned} p^{-1}L(\boldsymbol{\Sigma}_p, \boldsymbol{\Sigma}_{\mathrm{Ideal}}) &= p^{-1}\sum_{i=1}^p \log(\mathbf{p}_i^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{p}_i) - p^{-1}\sum_{i=1}^p \log(v_{n,i}) \\ &\overset{a.s.}{\to} \int \log(\delta(\lambda))dF(\lambda) - \int \log(\tau)dH(\tau), \end{aligned}$$

where $v_{n,i}$ is the ith largest eigenvalue of $\boldsymbol{\Phi}$. Now consider the decomposition

$$\begin{aligned} p^{-1}L(\boldsymbol{\Sigma}_p, \widehat{\boldsymbol{\Sigma}}_p) &= I_1 + I_2 + I_3 + I_4 + I_5, \quad \text{where} \\ I_1 &= \log(\widehat{\theta}/\theta), \\ I_2 &= \Big(\frac{\theta}{\widehat{\theta}} - 1\Big)p^{-1}\sum_{i=1}^p\left(\frac{\mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{p}_{1i}}{\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\boldsymbol{\Phi}}_2\mathbf{p}_{1i}}\right), \end{aligned}$$

$$I_3 = p^{-1}\sum_{i=1}^{p}\left(\frac{\mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{p}_{1i}}{\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\boldsymbol{\Phi}}_2\mathbf{p}_{1i}}\right) - 1,$$
$$I_4 = p^{-1}\sum_{i=1}^{p}\log\left(\frac{\mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{p}_{1i}}{\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\boldsymbol{\Phi}}_2\mathbf{p}_{1i}}\right),$$
$$I_5 = p^{-1}\sum_{i=1}^{p}\log(\mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{p}_{1i}) - p^{-1}\sum_{i=1}^{p}\log(v_{n,i}).$$

We can prove easily that I_1, I_2, I_3 and I_4 are all almost surely 0 by the proof of Theorem 3, and the result of Lemma 2. By Lemma S.3.7.1, we can show that

$$I_5 \overset{a.s.}{\to} \int \log(\delta(\lambda))dF(\lambda) - \int \log(\tau)dH(\tau),$$

so that $p^{-1}L(\boldsymbol{\Sigma}_p, \boldsymbol{\Sigma}_{\text{Ideal}})/p^{-1}L(\boldsymbol{\Sigma}_p, \widehat{\boldsymbol{\Sigma}}_p) \overset{a.s.}{\to} 1$, showing $EL(\boldsymbol{\Sigma}_p, \widehat{\boldsymbol{\Sigma}}_p) \overset{a.s.}{\to} 0$. This completes the proof of the theorem. □

Proof of Theorem 5. For the Frobenius loss,

$$\|\widehat{\boldsymbol{\Sigma}}_{p,M} - \boldsymbol{\Sigma}_p\|_F^2 = \|\frac{1}{M}\sum_{i=1}^{M}(\widehat{\boldsymbol{\Sigma}}_p^{(i)} - \boldsymbol{\Sigma}_p)\|_F^2 \leq (\frac{1}{M}\sum_{i=1}^{M}\|\widehat{\boldsymbol{\Sigma}}_p^{(i)} - \boldsymbol{\Sigma}_p\|_F)^2$$
$$\leq \frac{1}{M}\sum_{i=1}^{M}\|\widehat{\boldsymbol{\Sigma}}_p^{(i)} - \boldsymbol{\Sigma}_p\|_F^2,$$

so that

$$EL(\boldsymbol{\Sigma}_p, \widehat{\boldsymbol{\Sigma}}_{p,M}) \leq 1 - \frac{\|\widehat{\boldsymbol{\Sigma}}_{\text{Ideal}} - \boldsymbol{\Sigma}_p\|_F^2}{\frac{1}{M}\sum_{i=1}^{M}\|\boldsymbol{\Sigma}_p^{(i)} - \boldsymbol{\Sigma}_p\|_F^2} = 1 - \frac{1}{\frac{1}{M}\boldsymbol{\Sigma}_{i=1}^{M}\frac{1}{1-EL(\boldsymbol{\Sigma}_p, \widehat{\boldsymbol{\Sigma}}_p^{(i)})}}$$
$$\overset{a.s.}{\to} 0,$$

since $EL(\boldsymbol{\Sigma}_p, \widehat{\boldsymbol{\Sigma}}_p^{(i)}) \overset{a.s.}{\to} 0$ by Theorem 4.

For the inverse Stein's loss, we can follow exactly the same lines as in the proof of Theorem 6 in [23] to prove the result. □

Proof of Theorem 6. We have

$$p^{1/2}\left\|\mathbf{w}_{\text{opt}}\right\|_\infty \leq \frac{p^{1/2}\left\|\widehat{\boldsymbol{\Sigma}}_p^{-1}\right\|_1}{\mathbf{1}_p^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_p^{-1}\mathbf{1}_p} = \frac{p^{1/2}\left\|\widehat{\boldsymbol{\Phi}}^{-1}\right\|_\infty}{\mathbf{1}_p^{\mathrm{T}}\widehat{\boldsymbol{\Phi}}^{-1}\mathbf{1}_p} \leq \frac{p^{1/2}\cdot p^{1/2}\lambda_{\min}^{-1}(\widehat{\boldsymbol{\Phi}})}{p\lambda_{\min}(\widehat{\boldsymbol{\Phi}}^{-1})}$$
$$= \frac{\lambda_{\max}(\widehat{\boldsymbol{\Phi}})}{\lambda_{\min}(\widehat{\boldsymbol{\Phi}})} \leq \frac{\max_{1\leq i\leq p}\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\boldsymbol{\Phi}}_2\mathbf{p}_{1i}}{\min_{1\leq i\leq p}\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\boldsymbol{\Phi}}_2\mathbf{p}_{1i}} \overset{a.s.}{\to} \frac{\max_{1\leq i\leq p}\mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{p}_{1i}}{\min_{1\leq i\leq p}\mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{p}_{1i}}$$
$$\leq \frac{\lambda_{\max}(\boldsymbol{\Phi})}{\lambda_{\min}(\boldsymbol{\Phi})} = \text{Cond}(\boldsymbol{\Phi}),$$

where we used the results of Lemma 2 for the almost sure convergence. For the actual risk bound,

$$
\begin{aligned}
pR(\widehat{\boldsymbol{\Sigma}}_p) = \frac{p\mathbf{1}_p^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_p^{-1}\boldsymbol{\Sigma}_p\widehat{\boldsymbol{\Sigma}}_p^{-1}\mathbf{1}_p}{(\mathbf{1}_p^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_p^{-1}\mathbf{1}_p)^2} &\leq \frac{p\cdot p\lambda_{\max}^2(\widehat{\boldsymbol{\Sigma}}_p^{-1})\lambda_{\max}(\boldsymbol{\Sigma}_p)}{p^2\lambda_{\min}^2(\widehat{\boldsymbol{\Sigma}}_p^{-1})} = \left(\frac{\lambda_{\max}(\widehat{\boldsymbol{\Phi}})}{\lambda_{\min}(\widehat{\boldsymbol{\Phi}})}\right)^2\lambda_{\max}(\boldsymbol{\Sigma}_p) \\
&\left(\frac{\max_{1\leq i\leq p}\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\boldsymbol{\Phi}}_2\mathbf{p}_{1i}}{\min_{1\leq i\leq p}\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\boldsymbol{\Phi}}_2\mathbf{p}_{1i}}\right)^2\lambda_{\max}(\boldsymbol{\Sigma}_p) \overset{a.s.}{\to} \left(\frac{\max_{1\leq i\leq p}\mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{p}_{1i}}{\min_{1\leq i\leq p}\mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{p}_{1i}}\right)^2\lambda_{\max}(\boldsymbol{\Sigma}_p) \\
&\leq \left(\frac{\lambda_{\max}(\boldsymbol{\Phi})}{\lambda_{\min}(\boldsymbol{\Phi})}\right)^2\lambda_{\max}(\boldsymbol{\Sigma}_p) = \mathrm{Cond}^2(\boldsymbol{\Phi})\lambda_{\max}(\boldsymbol{\Sigma}_p).
\end{aligned}
$$

This completes the proof of the theorem. □

Bibliography

[1] Karim M. Abadir, Walter Distaso, and Filip Žikeš, *Model-free estimation of large variance matrices*, The Rimini Centre for Economic Analysis,, WP 10-17 (2010).

[2] Yacine Aït-Sahalia, Per A. Mykland, and Lan Zhang, *How often to sample a continuous-time process in the presence of market microstructure noise*, Review of Financial Studies, 18 (2005), pp. 351–416.

[3] Torben Andersen, Tim Bollerslev, Francis Diebold, and Labys P., *The distribution of realized exchange rate volatility*, Journal of the American Statistical Association, 96 (2001), pp. 42–55.

[4] ELENA Asparouhova, HENDRIK Bessembinder, and IVALINA Kalcheva, *Noisy prices and inference regarding returns*, The Journal of Finance, 68 (2013), pp. 665–714.

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

[6] Z. Bai and J. Silverstein, *Spectral analysis of large dimensional random matrices*, Springer, 2010.

[7] Zhidong Bai and Jack Silverstein, *Spectral Analysis of Large Dimensional Random Matrices*, Springer Series in Statistics, New York, 2 ed., 2010.

[8] Z. D. Bai and J.W. Silverstein, *No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices*, The Annals of Probability, 26 (1998), pp. 316–345.

[9] Ole E. Barndorff-Nielsen, Peter Reinhard Hansen, Asger Lunde, and Neil Shephard, *Multivariate realised kernels: Consistent positive semi-definite es-*

timators of the covariation of equity prices with noise and non-synchronous trading, Journal of Econometrics, 162 (2011), pp. 149 – 169.

[10] Peter J. Bickel and Elizaveta Levina, *Covariance regularization by thresholding*, The Annals of Statistics, 36 (2008), pp. 2577–2604.

[11] Peter J. Bickel and Elizaveta Levina, *Regularized estimation of large covariance matrices*, Ann. Statist., 36 (2008), pp. 199–227.

[12] C. Lam, Q. Yao and N. Bathia, *Estimation of latent factors for high-dimensional time series*, Biometrika, (2011).

[13] T. Cai and Harrison H. Zhou, *Optimal rates of convergence for sparse covariance matrix estimation*, The Annals of Statistics, 40 (2012), pp. 2389–2420.

[14] G. Chamberlain and M. Rothschild, *Arbitrage, factor structure, and mean variance analysis on large asset markets*, Econometrica, (1983), pp. 1305–1324

[15] Pesaran M.H. Chudik, A. and E. Tosetti, *Weak and strong cross-section depedence and estimation of large panels*, The econometrics journal, (2011).

[16] Jianqing Fan, Yingying Fan, and Jinchi Lv, *High dimensional covariance matrix estimation using a factor model*, Journal of Econometrics, 147 (2008), pp. 186–197.

[17] Jianqing Fan, Yingying Li, and Ke Yu, *Vast volatility matrix estimation using high- frequency data for portfolio selection*, Journal of the American Statistical Association, 107 (2012), pp. 412–428.

[18] Jianqing Fan, Yuan Liao, and Martina Mincheva, *High-dimensional covariance matrix estimation in approximate factor models*, The Annals of Statistics, 39 (2011), pp. 3320–3356.

[19] Jianqing Fan, Yuan Liao, and Martina Mincheva, *Large covariance estimation by thresholding principal orthogonal complements*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 75 (2013), pp. 603–680.

[20] Jerome Friedman, Trevor Hastie, and Robert Tibshirani, *Sparse inverse covariance estimation with the graphical lasso*, Biostatistics, 9 (2008), pp. 432–441.

[21] Jean Jacod and Philip Protter, *Asymptotic error distributions for the euler method for stochastic differential equations*, Ann. Probab., 26 (1998), pp. 267–307.

[22] Yuan Liao Jianqing Fan and Martina Mincheva, *Large covariance estimation by thresholding principal orthogonal complements*, JRS B, (2012).

[23] C. Lam, *Nonparametric eigenvalue-regularized precision or covariance matrix estimator*, Ann. Statist., (2016). To appear.

[24] Clifford Lam, *Supplement to "Nonparametric eigenvalue-regularized precision or covariance matrix estimator"*, Ann. Stats., (2016).

[25] Clifford Lam and Jianqing Fan, *Sparsistency and rates of convergence in large covariance matrix estimation*, Ann. Statist., 37 (2009), pp. 4254–4278.

[26] C. Lam and C. Hu, *Discussion of large covariance estimation by thresholding principal orthogonal complements by fan, liu and mincheva.*, Journal of Royal Statistical Society B., (2013).

[27] C. Lam and Q. Yao, *Factor modeling for high-dimensional time series: inference for the number of factors*, The Annals of Statistics, (2012).

[28] Olivier Ledoit and Sandrine Péché, *Eigenvectors of some large sample covariance matrix ensembles*, Probability Theory and Related Fields, 151 (2011), pp. 233–264.

[29] Olivier Ledoit and Michael Wolf, *A well-conditioned estimator for large-dimensional covariance matrices*, Journal of Multivariate Analysis, 88 (2004), pp. 365 – 411.

[30] Olivier Ledoit and Michael Wolf, *Nonlinear shrinkage estimation of large-dimensional covariance matrices*, The Annals of Statistics, 40 (2012), pp. 1024–1060.

[31] Olivier Ledoit and Michael Wolf, *Optimal estimation of a large-dimensional covariance matrix under Stein's loss*, ECON - Working Papers 122, Department of Economics - University of Zurich, 2013.

[32] Olivier Ledoit and Michael Wolf, *Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets Goldilocks*, ECON - Working Papers 137, Department of Economics - University of Zurich, Jan. 2014.

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

[34] Harry Markowitz, *Portfolio selection*, The Journal of Finance, 7 (1952), pp. 77–91.

[35] Nicolai Meinshausen and Peter Bühlmann, *High-dimensional graphs and variable selection with the lasso*, The Annals of Statistics, 34 (2006), pp. 1436–1462.

[36] Mohsen Pourahmadi, *Cholesky decompositions and estimation of a covariance matrix: Orthogonality of variancecorrelation parameters*, Biometrika, 94 (2007), pp. 1006–1013.

[37] Adam J. Rothman, Peter J. Bickel, Elizaveta Levina, and Ji Zhu, *Sparse permutation invariant covariance estimation*, Electron. J. Statist., 2 (2008), pp. 494–515.

[38] C. Stein, Estimation of a covariance matrix, Rietz lecture, 39th Annual Meeting IMS. Atlanta, Georgia., 1975.

[39] J. Stock and M. Watson, *Implications of dynamic factor models for var analysis.*, NBER Working Papers, (2005).

[40] Minjing Tao, Yazhen Wang, Qiwei Yao, and Jian Zou, *Large volatility matrix inference via combining low-frequency and high-frequency approaches*, Journal of the American Statistical Association, 106 (2011), pp. 1025–1040.

[41] Yazhen Wang and Jian Zou, *Vast volatility matrix estimation for high-frequency financial data*, Ann. Statist., 38 (2010), pp. 943–978.

[42] Lan Zhang, *Estimating covariation: Epps effect, microstructure noise*, Journal of Econometrics, 160 (2011), pp. 33 – 47.

[43] Xinghua Zheng and Yingying Li, *On the estimation of integrated covariance matrices of high dimensional diffusion processes*, Ann. Statist., 39 (2011), pp. 3121–3151.