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

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

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

Declaration 2

Acknowledgements 3

Abstract 4

1 Introduction 7
   1.1 Motivation for the first piece of work .......................... 7
   1.2 Motivation for the second piece of work ........................ 8

2 Autocorrelation-based factor analysis 9
   2.1 Relevant methods in the literature .............................. 9
   2.2 Models and estimation ........................................ 10
      2.2.1 The Models and assumptions .......................... 10
      2.2.2 Estimation for the factor loading space ................ 12
   2.3 Theoretical properties ..................................... 12
   2.4 Simulations .................................................. 14
   2.5 Real data example .......................................... 18
   2.6 Summary of this chapter ...................................... 19
   2.7 Proofs for this chapter ....................................... 20
   2.8 Discussion to “Large covariance estimation by thresholding principal ortho-
      onal complements” ............................................ 30

3 Nonlinear shrinkage of large integrated covariance matrix 33
   3.1 An overview of relevant estimation methods in the literature .. 33
      3.1.1 The problem of covariance matrix estimation ............ 33
      3.1.2 Large dimension integrated covariance matrix estimation .. 35
## Contents

3.2 Introduction ................................................. 35

3.3 Framework and Methodology ................................. 37

3.3.1 Nonlinear shrinkage estimator ......................... 39

3.4 Asymptotic Theory and Practical Implementation ......... 40

3.4.1 Practical Implementation ............................... 43

3.5 Empirical Results ........................................... 44

3.5.1 A market trading exercise ............................. 45

3.5.2 Portfolio allocation on NYSE data ..................... 48

3.6 Summary of this chapter .................................... 51

3.7 Proofs for this chapter ..................................... 51

Bibliography ..................................................... 59
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

$$y_t = Ax_t + \epsilon_t,$$

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 $A$ denotes the unknown constant factor loading matrix of size $p \times r$. Here we assume the number of factors $r$ is much smaller than $p$. Finally, $\epsilon_t$ is a $p \times 1$ white noise vector with mean zero and some covariance matrix $\Sigma_\epsilon$.

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

noise linear combination of \( x_t \) are absorbed into \( \epsilon_t \), and the rank of \( A \) 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 \( A = 0 \), \( y_t \) 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 \( A \). 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 \( \frac{p_j}{p} \). Then the model can be written as

\[
y_t = A x_t + \epsilon_t = A_s x_{ts} + A_w x_{tw} + \epsilon_t, \tag{2.2.2}
\]

where \( A_s' A_w = 0 \). In addition, for \( A_w \), we assume it adopts a known factor structure group:

\[
A_w = \begin{pmatrix}
A_{w_1} & 0 & \cdots & 0 \\
0 & A_{w_2} & 0 & \vdots \\
\vdots & 0 & \ddots & 0 \\
0 & \cdots & 0 & A_{w_r}
\end{pmatrix}, \quad A_{w_i} \in \mathbb{R}^{p_i}. \tag{2.2.3}
\]

Here is some more notations. Define \( \Sigma_s(k) = \text{Cov}(x_{ts}, x_{(t-k)s}) \), \( \Sigma_{w_j}(k) = \text{Cov}(x_{t,w_j}, x_{(t-k),w_j}) \) and \( \Sigma_x(k) = \text{Cov}(x_{t+k}, \epsilon_t) \). For \( k = 0, 1, 2, \ldots, k_0 \). Also put

\[
\Sigma_y(k) = \text{Cov}(y_t, y_{t-k}), \quad \Sigma_x(k) = \text{Cov}(x_t, x_{t-k}),
\]

\[
\Sigma_{x\epsilon}(k) = \text{Cov}(x_t, \epsilon_{t-k}).
\]

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, \( A'A = I_r \).

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

(C3) The observable series \( y_t \) is strictly stationary and \( \Psi \)-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 \( \| \Sigma_s(k) \| \asymp p \asymp \| \Sigma_s(k) \|_{\text{min}} \), and for a local factor \( w_j \), \( \| \Sigma_{w_j}(k) \| \asymp p_j \asymp \| \Sigma_{w_j}(k) \|_{\text{min}} \). Here, \( p_j \) denotes the number of non-zero elements in a specific local factor \( w_j \).

(C5) For \( k = 0, 1, \cdots, k_0 \), \( \| \Sigma_{x\epsilon}(k) \| = o(p_j) \).
2.2.2 Estimation for the factor loading space

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

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
\[ \Sigma_y(k) = A\Sigma_x(k)A' + A\Sigma_{xe}(k), \quad k \geq 1. \]

Define the $p \times p$ nonnegative definite matrix $M$ by
\[ M = \sum_{k=1}^{k_0} \Sigma_y(k)\Sigma_y(k)', \quad \text{where } k_0 \text{ is a prescribed integer.} \]

This matrix is constructed to accumulate the information from different time lags. Since
\[ \Sigma_y(k) = A\Sigma_x(k)A' + A\Sigma_{xe}(k), \quad k \geq 1, \]
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 $M$ corresponding to its nonzero eigenvalues. We take the $r$ orthonormal eigenvectors corresponding to non-zero eigenvalues of $M$ as the columns of $A$.

Finally, $\hat{A}$ is found by performing an eigenanalysis on the sample version $\hat{M}$:
\[ \hat{M} = \sum_{k=1}^{k_0} \hat{\Sigma}_y(k)\hat{\Sigma}_y(k)', \]
where $\hat{\Sigma}_y(k)$ 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 $G$, $\|G\|$ is the $L_2$ norm, $\|G\|_F$ is the Frobenius norm and $\|G\|_{\min}$ is the square root of the smallest non-zero eigenvalue of $GG'$. 
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 \( \| \Sigma_e \| = O(p^\gamma), \gamma \in (0, 1) \), and denote the PCA estimator for fact loading matrix by \( \hat{A}_{PCA} \), it holds for PCA that:

\[
\| \hat{A}_{PCA} - A \| = O_p\left( \left( \frac{p}{\min_j p_j} \right)^{1/2} n^{-\frac{1}{2}} \| B' \Sigma_s B \|^{\frac{1}{2}} \left( 1 + \left( \frac{p}{\min_j p_j} \right)^{1/2} p^{\gamma/2} \right) + \min_j p_j^{-1} \| B' \Sigma_s A \| \right),
\]

and for AFA,

\[
\| \hat{A} - A \| = O_p\left( \left( \frac{p}{\min_j p_j} \right)^{1/2} n^{-\frac{1}{2}} \| B' \Sigma_s B \|^{\frac{1}{2}} \left( 1 + \left( \frac{p}{\min_j p_j} \right)^{1/2} p^{\gamma/2} \right) \right).
\]

From theorem 1, we can see the difference of those two rates is that PCA has an extra term \( \min_j p_j^{-1} \| B' \Sigma_s A \| \) 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, \( \min_j p_j^{-1} \| B' \Sigma_s A \| \) 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 \( \hat{A} \equiv (\hat{A}_s, \hat{A}_w) \) for the factor loading matrix \( A = (A_s, A_w) \), then the effects of pervasive factors can be removed from the data using

\[
y_t^* = y_t - \hat{A}_s \hat{A}_s' y_t.
\]

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

\[
\tilde{A} = (\hat{A}_s, \tilde{A}_w).
\]

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

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

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

$$\|\tilde{A}_{PCA} - A\| = O_P\left(\frac{1}{2}p^2 p_M p_m^{-1} n^{-\frac{1}{2}} \right),$$

and the counterpart rate for AFA is

$$\|\tilde{A} - A\| = O_P\left(\frac{1}{2}p^2 p_M p_m^{-1} n^{-\frac{1}{2}} \right).$$

If we know the local factor structure group, and denote $\Sigma_e^{(i)}$ as the covariance for $i$th group of noise, then for local factors $i$th group $A_w^{(i)},$

$$\|\tilde{A}_{PCA}^{(i)} - A_w^{(i)}\| = O_P\left(n^{-\frac{1}{2}} + p_i^{-1}\|B^t \Sigma_e^{(i)} A_w^{(i)}\|\right),$$

and

$$\|\tilde{A}^{(i)} - A_w^{(i)}\| = O_P\left(n^{-\frac{1}{2}}\right).$$

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 $100 \times 11$ factor loading matrix has the following structure: (hence the factor structure group is known as we assumed.)

$$A = \begin{pmatrix} A_{s1} & A_{w1} & 0 & \cdots & 0 \\ A_{s2} & 0 & A_{w2} & 0 & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ A_{s10} & 0 & \cdots & 0 & A_{w10} \end{pmatrix}, A_{si} \in \mathbb{R}^{10}$$

(2.4.1)

Every non-zero entry in $A$ 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 $\eta_t$ are independent $N(0, 1)$ random variables. The performance of estimator is measured by

$$d(\hat{A}) = \| (I - AA') \hat{A} \|,$$

(2.4.2)
2.4. Simulations

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 $\epsilon_t$.

**Scenario I**

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

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

and $H$ is the Hurst parameter which takes value in $[0.5, 1]$. Larger $H_p$ means stronger cross-
correlations for the components in $\epsilon_t$ and larger $H_n$ accounts for stronger autocorrelations
for $\epsilon_t$’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>
<thead>
<tr>
<th>$H_p$</th>
<th>$H_n$</th>
<th>Errors/s.d.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>AFA1</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>26(22)</td>
</tr>
<tr>
<td>0.7</td>
<td>0.5</td>
<td>26(22)</td>
</tr>
<tr>
<td>0.9</td>
<td>0.5</td>
<td>26(36)</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>42(28)</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>43(37)</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>46(61)</td>
</tr>
<tr>
<td>0.5</td>
<td>0.9</td>
<td>89(91)</td>
</tr>
<tr>
<td>0.7</td>
<td>0.9</td>
<td>88(89)</td>
</tr>
<tr>
<td>0.9</td>
<td>0.9</td>
<td>81(115)</td>
</tr>
</tbody>
</table>

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 $\Sigma_2 = \begin{pmatrix} H & 0 & \cdots & 0 \\ 0 & H & 0 & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & H \end{pmatrix}$, and each $H$ is a $10 \times 10$ diagonal matrix. In this case, the white noise $\epsilon_t$ adopts a heterogeneous and diagonal covariance structure.

To obtain $\epsilon_t$ in scenario II, we first generate $p \times n$ matrix $Z$ with entries being independently identically distributed standard normal random variable. And let $H = \text{diag}\{\sigma_1^2, \sigma_1^2, \sigma_2^2, \sigma_2^2, \sigma_3^2, \sigma_3^2, \sigma_3^2, \sigma_3^2\}$. Therefore $\Sigma_2^{1/2}Z$ 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., $\Sigma_2$ has all diagonal entries equal one.

<table>
<thead>
<tr>
<th>n</th>
<th>AFA1</th>
<th>PCA1</th>
<th>AFA2</th>
<th>PCA2</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>91(79)</td>
<td>74(92)</td>
<td>52(115)</td>
<td>42(95)</td>
</tr>
<tr>
<td>500</td>
<td>48(64)</td>
<td>36(30)</td>
<td>35(78)</td>
<td>28(64)</td>
</tr>
<tr>
<td>1000</td>
<td>26(21)</td>
<td>21(17)</td>
<td>26(70)</td>
<td>20(58)</td>
</tr>
</tbody>
</table>

Table 2.2: Scenario II Homogeneous noise

*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 heterogeneous noise.
Table 2.3: Scenario II Heteroscedastic noise

<table>
<thead>
<tr>
<th>( \sigma_1/\sigma_2/\sigma_3 )</th>
<th>Error/s.d.</th>
<th>AFA1</th>
<th>PCA1</th>
<th>AFA2</th>
<th>PCA2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/1/1</td>
<td>26(22)</td>
<td>22(16)</td>
<td>26(68)</td>
<td>19(52)</td>
<td></td>
</tr>
<tr>
<td>1/1.2/1.3</td>
<td>32(32)</td>
<td>30(23)</td>
<td>25(59)</td>
<td>20(43)</td>
<td></td>
</tr>
<tr>
<td>1/1.3/1.5</td>
<td>35(29)</td>
<td>37(29)</td>
<td>26(75)</td>
<td>23(46)</td>
<td></td>
</tr>
<tr>
<td>1/1.4/1.8</td>
<td>43(50)</td>
<td>53(53)</td>
<td>28(70)</td>
<td>29(42)</td>
<td></td>
</tr>
<tr>
<td>1/1.5/2</td>
<td>47(19)</td>
<td>62(60)</td>
<td>27(61)</td>
<td>34(35)</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>n</th>
<th>Error</th>
<th>AFA1</th>
<th>PCA1</th>
<th>AFA2</th>
<th>PCA2</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>99</td>
<td>99</td>
<td>77</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>99</td>
<td>99</td>
<td>55</td>
<td>87</td>
<td></td>
</tr>
<tr>
<td>300</td>
<td>99</td>
<td>99</td>
<td>48</td>
<td>81</td>
<td></td>
</tr>
<tr>
<td>400</td>
<td>99</td>
<td>99</td>
<td>40</td>
<td>67</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>98</td>
<td>99</td>
<td>37</td>
<td>66</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>60</td>
<td>99</td>
<td>28</td>
<td>52</td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>28</td>
<td>98</td>
<td>20</td>
<td>49</td>
<td></td>
</tr>
<tr>
<td>5000</td>
<td>16</td>
<td>99</td>
<td>13</td>
<td>37</td>
<td></td>
</tr>
</tbody>
</table>

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 $x_{i+490}^{(1)}$, so as to obtain a one-step ahead forecast $\hat{y}_{i+490}^{(1)} = \hat{A}x_{i+490}$ for $y_{i+490}$. For comparison, we calculated the forecast error for the $(i + 490)$th month for each method, defined by

$$\text{Forecast error} = p^{-\frac{1}{2}}\|\hat{y}_{i+490}^{(1)} - y_{i+490}\|.$$  \hfill (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
local factors for categories with more than 4 variables. The cumulative forecast errors for estimation with known factor structure group are 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.
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:

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

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

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

Figure 2.2: cumulative forecast error for estimation without assuming a known structure, blue line for AFA, black line for PCA
such that the columns of $\hat{Q}_1 = (Q_1 + Q_2P)(I + P'P)^{-1/2}$ define an orthonormal basis for a subspace that is invariant for $A + E$.

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

$$\Sigma_y = A\Sigma_xA' + \Sigma_e,$$

Its corresponding sample version is

$$\hat{\Sigma}_y = n^{-1}\Sigma_{t=1}^n y_t y_t'$$

$$= A\hat{\Sigma}_x A + A\hat{\Sigma}_e\xi + \hat{\Sigma}_e A' + \hat{\Sigma}_e.$$

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

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

Hence,

$$\text{sep}(D, 0) = \lambda_{\min}(D)$$
\[ \lambda_{\min}( \Sigma_x ) \]
\[ \approx p_j. \]

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,

\[
Q' EQ = \begin{pmatrix} A' \\ B' \end{pmatrix} (\hat{\Sigma}_y - A \Sigma_x A') (A B).
\]

So we have \( E_{21} = B' \hat{\Sigma}_{ex} + B' \hat{\Sigma}_e A \).

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

Therefore,

\[
\| E_{21} \| \leq \| B' \hat{\Sigma}_{ex} \| + \| B' \Sigma_e A \| + \| B' (n^{-1} \epsilon' - \Sigma_e) A \|.
\]

This implies,

\[
\| B' \hat{\Sigma}_{ex} \| \leq \| B' \Sigma_e^{\frac{1}{2}} \| \cdot \| \hat{\Sigma}_{uz} \| \cdot \| \Sigma_e^{\frac{1}{2}} \|.
\]

Using \( \| \hat{\Sigma}_{uz} \| = O_P( p^{\frac{1}{2}} n^{\frac{1}{2}} ) \), and \( \| \Sigma_e^{\frac{1}{2}} \| = O_P( p_j^{\frac{1}{2}} ) \), we have

\[
\| B' \hat{\Sigma}_{ex} \| = O_P( \| B' \Sigma_e A \|^{\frac{1}{2}} \cdot p_j^{\frac{1}{2}} \cdot p^{\frac{1}{2}} n^{\frac{1}{2}} )
\]

Then for the last term,

\[
B' (n^{-1} \epsilon' - \Sigma_e) A = B' (N^{\frac{1}{2}} n^{-1} uu' \Sigma_e^{\frac{1}{2}} - \Sigma_e^{\frac{1}{2}} \Sigma_e^{\frac{1}{2}}) A
\]
\[= B' \Sigma_e^{\frac{1}{2}} (n^{-1} uu' - I) \Sigma_e^{\frac{1}{2}} A.
\]

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

Put everything together, we have

\[
\| B' (n^{-1} \epsilon' - \Sigma_e) A \| \leq \| B' \Sigma_e B \|^{\frac{1}{2}} O_P( p^{n^{-\frac{1}{2}}} ) O_P( p^{\gamma/2} ).
\]
2.7. Proofs for this chapter

\[ = O_P(\|B^T \Sigma B\|^{\frac{1}{2}} p n^{\frac{1}{2}} p^{\gamma/2}). \]

This leads to

\[ \|E_{21}\| = O_P(\|B^T \Sigma B\|^{\frac{1}{2}} p_{jj}^{\frac{1}{2}} n^{\frac{1}{2}} + \|B^T \Sigma A\|) \]

Finally,

\[ \|\hat{A}_{\text{PCA}} - A\| = \frac{\|E_{21}\|}{\text{sep}(D, 0)} \]

\[ = O_P((pp_j)^{\frac{1}{2}} n^{\frac{1}{2}} (\frac{p_j}{p_j})^{\frac{1}{2}} p^{\gamma/2} + 1 + \|B^T \Sigma A\|)/p_j \]

This proves the PCA rate in Theorem 1.

Now, for AFA, define

\[ L = \sum_{k=1}^{k_0} \Sigma_y(k) \Sigma_y(k)' \]

\[ = A \left( \sum_{k=1}^{k_0} (\Sigma_x(k)A' + \Sigma_{xe}(k)) (\Sigma_x(k)A' + \Sigma_{xe}(k))' \right) A'. \]

and

\[ \hat{L} = \sum_{k=1}^{k_0} \hat{\Sigma}_y(k) \hat{\Sigma}_y(k)' \]

\[ = \sum_{k=1}^{k_0} (A \hat{\Sigma}_x(k)A' + A \hat{\Sigma}_{xe}(k) + \hat{\Sigma}_{ex}(k)A' + \hat{\Sigma}_e(k)) \]

\[ (A \hat{\Sigma}_x(k)A' + A \hat{\Sigma}_{xe}(k) + \hat{\Sigma}_{ex}(k)A' + \hat{\Sigma}_e(k))'. \]

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

Since \( B' L = 0 \) and noting that \( B' A = 0 \), we have

\[ E_{21} = B' (\hat{L} - L) A \]

\[ = B' \hat{L} A. \]

Then,

\[ B' \sum_{k=1}^{k_0} (A \hat{\Sigma}_x(k)A' + A \hat{\Sigma}_{xe}(k) + \hat{\Sigma}_{ex}(k)A' + \hat{\Sigma}_e(k)) \]
(A\hat{\Sigma}_x(k)A' + A\hat{\Sigma}_{xe}(k) + \hat{\Sigma}_{xx}(k)A' + \hat{\Sigma}_e(k))'A

= \sum_{k=1}^{k_0}(B'\hat{\Sigma}_{xx}(k)A' + B'\hat{\Sigma}_e(k))(A\hat{\Sigma}_x(k)A' + A\hat{\Sigma}_{xe}(k) + \hat{\Sigma}_{xx}(k)A' + \hat{\Sigma}_e(k))'A.

We can write,

A\hat{\Sigma}_x(k)A' + A\hat{\Sigma}_{xe}(k) + \hat{\Sigma}_{xx}(k)A' + \hat{\Sigma}_e(k) = \hat{\Sigma}_y(k)

= \hat{\Sigma}_y(k) - \Sigma_y(k) + \Sigma_y(k)

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

\|\hat{\Sigma}_y(k) - \Sigma_y(k)\| = O_p((pp_j)^{\frac{1}{2}}n^{-\frac{1}{2}})

\|\Sigma_y(k)\| = O(p_j).

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

\|A\hat{\Sigma}_x(k)A' + A\hat{\Sigma}_{xe}(k) + \hat{\Sigma}_{xx}(k)A' + \hat{\Sigma}_e(k)\| \leq \|\hat{\Sigma}_y(k) - \Sigma_y(k)\| + \|\Sigma_y(k)\|

= O_p((\frac{p}{p_j})^{\frac{1}{2}}n^{-\frac{1}{2}} + p_j)

= O_p(p_j).

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

What is left is to find the order of \(\|B'\hat{\Sigma}_{xx}(k)A' + B'\hat{\Sigma}_e(k)\|\).

Recall we have defined \(\epsilon_t = \Sigma_2^\frac{1}{2}u_t\) and \(x_t = \Sigma_2^\frac{1}{2}z_t\), so we have

\[B'\hat{\Sigma}_{xx}(k)A' = B'\frac{1}{n-k}\Sigma_2^\frac{1}{2}u_tz_{t-k}'\Sigma_2^\frac{1}{2}A^T = B'\Sigma_2^\frac{1}{2}\Sigma_{uz}(k)\Sigma_2^\frac{1}{2}A^T.\]

This implies,

\[\|B'\hat{\Sigma}_{xx}(k)A'\| \leq \|B'\Sigma_2B\|^\frac{1}{2} \cdot \|\Sigma_{uz}(k)\| \cdot \|\Sigma_2^\frac{1}{2}\| \]

\[= O_p(\|B'\Sigma_2B\|^\frac{1}{2}p^\frac{1}{2}n^{-\frac{1}{2}}p_j^\frac{1}{2}).\]
= O_P(∥B'Σ_cB∥^{1/2}(pp_j)^{1/2} n^{-1/2}).

And (recall that Σ_c(k) = 0),

\[ \|B'\hat{Σ}_c(k)\| = \|B'(Σ_{1,2}^{1/2}(n-k)^{-1} u_{\ell} u_{\ell-k} Σ_{1,2}^{1/2})\| \leq \|B'Σ_cB\|^ {1/2} \|\hat{Σ}_c(k)\| \cdot \|Σ_{1,2}^{1/2}\| = O_P(∥B'Σ_cB∥^{1/2} pn^{-1/2} p^{2}). \]

Therefore,

\[ \|B'\hat{Σ}_{cs}(k)A' + B'\hat{Σ}_c(k)\| \leq \|B'\hat{Σ}_{cs}(k)A'\| + \|B'\hat{Σ}_c(k)\| = O_P((pp_j)^{1/2} n^{-1/2} \|B'Σ_cB\|^ {1/2} (1 + (p/p_j)^{1/2} p^{2})). \]

Finally, we have

\[ \|\hat{A} - A\| = \frac{O_P((pp_j)^{1/2} n^{-1/2} \|B'Σ_cB\|^ {1/2} (1 + (p/p_j)^{1/2} p^{2}) \cdot p_j)}{\text{sep}(D, 0)} = \frac{O_P(p^{1/2} p_j^{3/2} n^{-1/2} \|B'Σ_cB\|^ {1/2} (1 + (p/p_j)^{1/2} p^{2}))}{p_j^2} = O_P((p/p_j)^{1/2} n^{1/2} \|B'Σ_cB\|^ {1/2} (1 + (p/p_j)^{1/2} p^{2})). \]

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

\[ \hat{Σ}_y = n^{-1} \sum_{t=1}^{n} y_t y_t' = A_s \hat{Σ}_s A_s' + A_w \hat{Σ}_w A_w' + A_e \hat{Σ}_e \]

+ A_w \hat{Σ}w_s A_s' + A_w \hat{Σ}_w A_w' + A_e \hat{Σ}_w e' + \hat{Σ}_e.

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

\[ Σ_y = A_s Σ_s A_s' + A_w Σ_w A_w' + Σ_e. \]

We write \( \hat{L} = \hat{Σ}_y = A_s Σ_s A_s' + A_w Σ_w A_w' + (\hat{Σ}_y - A_s Σ_s A_s' - A_w Σ_w A_w'). \)

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, \( A_s \) and \( A_w \) contain \( r_1 + r_2 \) eigenvectors of \( L \). Let \( B \) be the orthogonal complement of \((A_s \ A_w)\), that is \( B'B = I_{p-r_1-r_2} \), and \( B'A_s = B'A_w = 0 \). Then

\[
\left( \begin{array}{c} A'_s \\ A'_w \\ B' \end{array} \right) \ L(A_s \ A_w \ B) = \left( \begin{array}{cc} D_s & 0 \\ 0 & D_w \\ 0 & 0 \end{array} \right),
\]

where \( A'_s LA_s = D_s \). To find the order of \( \| \hat{A}_s - A_s \| \), we need to know \( \text{sep}(D_s, \left( \begin{array}{cc} D_w & 0 \\ 0 & 0 \end{array} \right)) \).

\[ \text{sep}(D_s, \left( \begin{array}{cc} D_w & 0 \\ 0 & 0 \end{array} \right)) = \lambda_{\min}(D_s) - \lambda_{\max}(D_w) = \lambda_{\min}(\Sigma_s) - \lambda_{\max}(\Sigma_w) \]

\[ \approx p - \max_{1 \leq j \leq r_2} p_j \]

\[ \approx p. \]

\[
Q'EQ = \left( \begin{array}{c} A'_s \\ A'_w \\ B' \end{array} \right) \ (\Sigma_y - A_s \Sigma_s A'_s - A_w \Sigma_w A'_w) (A_s \ A_w \ B) = \left( \begin{array}{c} E_{21} \end{array} \right).
\]

In our case, \( Q = [Q_1 \ Q_2] \), where we take \( Q_1 = A_s \) and \( Q_2 = [A_w \ A_s] \),

\[
E_{21} = \left( \begin{array}{c} A' \\ B' \end{array} \right) \ (\Sigma_y - A_s \Sigma_s A'_s - A_w \Sigma_w A'_w) A_s = \left( \begin{array}{c} A'_w \ (\Sigma_y A_s) \end{array} \right).
\]

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

\[
\| A'_w \hat{\Sigma}_{\epsilon} A_s \| = \| \hat{\Sigma}_{ws} + \hat{\Sigma}_{we} A_s + A'_w \hat{\Sigma}_{\epsilon} A_s + A'_w \hat{\Sigma}_{\epsilon} A_s \|
\]

\[
\leq \| \Sigma^{\frac{1}{2}}_w (n^{-1} z_w' z_s') \Sigma^{\frac{1}{2}}_x \| + \| \Sigma^{\frac{1}{2}}_w (n^{-1} z_w' u') \Sigma^{\frac{1}{2}}_{\epsilon} A_s \|
\]

\[
+ \| A'_w \Sigma^{\frac{1}{2}}_{\epsilon} (n^{-1} u z'_s) \Sigma^{\frac{1}{2}}_x \| + \| A'_w \Sigma^{\frac{1}{2}}_{\epsilon} (n^{-1} u u') \Sigma^{\frac{1}{2}}_x \Sigma^{\frac{1}{2}} \|
\]

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

\[
\| \Sigma^{\frac{1}{2}}_w (n^{-1} z_w' z_s') \Sigma^{\frac{1}{2}}_x \| + \| \Sigma^{\frac{1}{2}}_w (n^{-1} z_w' z_s') \Sigma^{\frac{1}{2}}_x \|
\]

\[
= O_P(\max_j p_j^{\frac{1}{2}} \cdot p^{\frac{1}{2}} \cdot n^{-\frac{1}{2}}).
\]
\[
\| \Sigma_{w}^{\frac{1}{2}} \Sigma_{z_{w}} \Sigma_{w}^{\frac{1}{2}} A_s \| + \| \Sigma_{w}^{\frac{1}{2}} (n^{-1} z_{w} u' - \Sigma_{z_{w}}) \Sigma_{w}^{\frac{1}{2}} A_s \| = O_P(\max_j p_j^{\frac{1}{2}} \cdot p^{\frac{1}{2}} \cdot p^{\frac{1}{2}} n^{-\frac{1}{2}}),
\]
\[
\| A'_{w} \Sigma_{z_{w}}^{\frac{1}{2}} \Sigma_{z_{w}}^{\frac{1}{2}} \| + \| A'_{w} \Sigma_{z_{w}}^{\frac{1}{2}} (n^{-1} u z_{s} - \Sigma_{z_{w}}) \Sigma_{z_{w}}^{\frac{1}{2}} \| = O_P(p^{\frac{1}{2}} \cdot p^{\frac{1}{2}} \cdot p^{\frac{1}{2}} n^{-\frac{1}{2}}),
\]
\[
\| A'_{w} \Sigma_{w} A_s \| + \| A'_{w} \Sigma_{z_{w}}^{\frac{1}{2}} (n^{-1} u w' - I) \Sigma_{z_{w}}^{\frac{1}{2}} A_s \| = O_P(p^{\gamma} + p n^{-\frac{1}{2}}).
\]

This gives
\[
\| A'_{w} \hat{\Sigma}_{y} A_s \| = O_P(p^{\frac{1}{2}} p n^{-\frac{1}{2}} + p^{\gamma} + p n^{-\frac{1}{2}}).
\]

\[
\| B' \hat{\Sigma}_{y} A_s \| = \| B' \hat{\Sigma}_{e} + B' \hat{\Sigma}_{c} A_s \|
\]
\[
\leq \| B' \hat{\Sigma}_{e} \| + \| B' \hat{\Sigma}_{c} A_s \|
\]
\[
= O_P(\| B' \Sigma_{e} B \|^\frac{1}{2} p n^{-\frac{1}{2}} + \| B' \Sigma_{e} B \|^\frac{1}{2} p^{\frac{1}{2}} p n^{-\frac{1}{2}} + \| B' \Sigma_{c} A \|)
\]
\[
= O_P(\| \Sigma_{e} \|^\frac{1}{2} p n^{-\frac{1}{2}} + \| \Sigma_{e} \|^\frac{1}{2} p^{\frac{1}{2}} p n^{-\frac{1}{2}} + \| \Sigma_{c} \|)
\]
\[
= O_P(p^{\gamma} + p^{\gamma} p^{\frac{1}{2}} p^{\frac{1}{2}} n^{-\frac{1}{2}} + p^{\gamma}),
\]
\[
\| B' \hat{\Sigma}_{y} A_s \| = O_P(p^{\frac{1}{2}} p n^{-\frac{1}{2}} + p^{\gamma} + p^{\gamma} p n^{-\frac{1}{2}}).
\]

Therefore, if \( p^{\gamma} \geq 1 \),
\[
\| E_{21} \| = O_P(p^{\gamma} + p^{\gamma} p n^{-\frac{1}{2}}).
\]

Finally,
\[
\| \hat{A}_s - A_s \| = \frac{\| E_{21} \|}{\text{sep} \left( D_{s}, \begin{pmatrix} D_{w} & 0 \\ 0 & 0 \end{pmatrix} \right)} = O_P(p^{\gamma} + p^{\gamma} p^{\frac{1}{2}} n^{-\frac{1}{2}}) = O_P(p^{\gamma - 1} + p^{\gamma} n^{-\frac{1}{2}}).
\]

The overall rate is determined by the local factor convergence speed. If we remove the effect of pervasive factors, then we have
\[
\Sigma_{g} = A_{w} \Sigma_{w} A_{w}^{\prime} + A_{s} \Sigma_{s} A_{s}^{\prime} + A_{\hat{\Sigma}_{e}}, H_{s} + H_{s} \Sigma_{s} H_{s}, \text{ where } H_{s} = I - A_{s} A_{s}^{\prime}.
\]
\[
\hat{\Sigma}_{g} = \hat{H}_{s}, \Sigma_{g} \hat{H}_{s}, \text{ where } \hat{H}_{s} = I - \hat{A}_{s} \hat{A}_{s}^{\prime}.
\]

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

Hence,

\[
\left( \begin{array}{c}
A_w' \\
B'
\end{array} \right)(A_w \Sigma_w A_w')(A_w \Sigma_w A_w') = \left( \begin{array}{c}
D_w^* \\
0
\end{array} \right).
\]

After some algebra, we have

\[
\| E_{21}^* \| = O_P(\| B' \tilde{\Sigma}_y A_w \|) = O_P(\| B' \tilde{H}_s \tilde{\Sigma}_w A_w' \tilde{H}_s A_w \| + \| B' \tilde{H}_s \tilde{\Sigma}_w A_w \|) = O_P(r_2 \frac{1}{2} (p \cdot \max p_j)^{\frac{1}{2}} n^{-\frac{1}{2}} + \| B' \Sigma w A_w \|),
\]

and finally, the two step estimation rate for PCA is

\[
\| \tilde{A}_w - A_w \| = O_P(\frac{\| E_{21}^* \|}{\text{sep}(D_w,0)}) = O_P(r_2 \frac{1}{2} (p \cdot \max p_j)^{\frac{1}{2}} n^{-\frac{1}{2}} + \| B' \Sigma w A_w \|).
\]

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 \( i \)th local factor,

\[
\| \tilde{A}_{w_i} - A_{w_i} \| = O_P(\frac{\| B' E_i A_{w_i} \|}{\text{sep}(\Sigma_{w_i}^{(i)},0)}),
\]
where $B_i$ is the orthogonal complement of $A_{w_i}$, and $\Sigma_w^{(i)}$ is the covariance matrix for the $i$th group local factors. For $i$th group of local factors, we data $y_{i}^{*}$, that is the corresponding rows of $y_i^{*}$. Then we have

$$E_i = \hat{\Sigma}_{y_i^{*}} - A_{w_i} \Sigma_w^{(i)} A_{w_i}'.$$

Therefore,

$$B_i' E_i A_{w_i} = B_i' \hat{H}_s^{(i)} A_s \hat{\Sigma}_s A_s' \hat{H}_s^{(i)} A_w + B_i' \hat{H}_s^{(i)} A_s \hat{\Sigma}_w A_w' \hat{H}_s^{(i)} A_w + B_i' \hat{H}_s^{(i)} A_s \hat{\Sigma}_w A_w' \hat{H}_s^{(i)} A_w$$

$$+ B_i' \hat{H}_s^{(i)} A_w \hat{\Sigma}_w A_w' \hat{H}_s^{(i)} A_w + B_i' \hat{H}_s^{(i)} A_w \hat{\Sigma}_w A_w' \hat{H}_s^{(i)} A_w + B_i' \hat{H}_s^{(i)} A_w \hat{\Sigma}_w A_w' \hat{H}_s^{(i)} A_w$$

$$= \sum_{j=1}^{9} I_j.$$

Also we have,

$$\|\hat{\Sigma}_s^{(i)}\| = O_P(p),$$

$$\|\hat{\Sigma}_w^{(i)}\| = O_P(p_i),$$

$$\|\hat{\Sigma}_w^{(i)}\| = \|\hat{\Sigma}_w^{(i)}\| = O_P\left(p_i^2 p_n^{-\frac{1}{2}}\right),$$

$$\|\hat{\Sigma}_w^{(i)}\| = \|\hat{\Sigma}_w^{(i)}\| = O_P\left(p_i^2 n^{-\frac{1}{2}}\right),$$

$$\|\hat{\Sigma}_w^{(i)} - \Sigma_w^{(i)}\| = O_P\left(p_i n^{-\frac{1}{2}}\right).$$

Hence the dominating terms are $\|I_s\| + \|I_0\|$, therefore,

$$B_i' E_i A_{w_i} = O_P\left(p_i n^{-\frac{1}{2}} + \|B_i' \Sigma_w^{(i)} A_{w_i}'\|\right).$$

Finally, if we know the local factor structure group, for $i$th group of local factors,

$$\|\hat{A}_w^{(i)} - A_{w_i}\| = O_P\left(p_i n^{-\frac{1}{2}} + \|B_i' \Sigma_w^{(i)} A_{w_i}'\|\right)$$

$$= O_P\left(n^{-\frac{1}{2}} + p_i^{-1} \|B_i' \Sigma_w^{(i)} A_{w_i}'\|\right).$$

We can use a similar argument to prove the rate for AFA. \qed
We congratulate the authors for this insightful paper. 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.](image)

---

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

$$y_t = X_t \beta + \epsilon_t, \beta = (-0.5, 0.5, 0.3, -0.6)'$$ \hspace{1cm} (2.8.1)

with $x_{it}$ being independent AR(1) processes and $\epsilon_t$ the standardized macroeconomic data in [6] plus independent $N(0, 0.2)$ noise. Following Example 5 of the paper, we estimate $\Sigma^{-1}_\epsilon$ using different methods and plot the sum of absolute bias for estimating $\beta$ 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](image)

Figure 2.5: Sum of absolute bias (averaged over 100 simulations) for estimating $\beta$ 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.
2.8. Discussion to “Large covariance estimation by thresholding principal orthogonal complements”

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 $Y$ to be the $p \times n$ data matrix consists of $n$ independent and identically distributed sample that we can observe,\footnote{Without loss of generality we assume that the mean of each row vector of $Y$ is zero.} sample covariance matrix which is defined by

$$S_n = \frac{1}{n} YY', \quad (3.1.1)$$

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 $p^2 \left(\frac{1}{2}p(p + 1)\right)$. When the dimensionality $p$ grows, the number of parameters to be estimated increases proportional to $p^2$, but the
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:

\[
\min_D \| PDP' - \Sigma \|_F. \tag{3.1.2}
\]

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

\[
d_i = p_i \Sigma p_i, i = 1, \cdots, p, p_i \text{isthe eigenvector}. \tag{3.1.3}
\]

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 (NER-COME) 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., $p/n \rightarrow c > 0$ 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 $p/n \rightarrow c > 0$. 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 $X_t = (X_t^{(1)}, \cdots, X_t^{(p)})^T$ be a $p$-dimensional log-price process which is modeled by the diffusion process

$$dX_t = \mu_t dt + \Theta_t dW_t, \quad t \in [0, 1],$$

(3.3.1)

where $\mu_t$ is the drift, $\Theta_t$ is a $p \times p$ matrix called the (instantaneous) covolatility process, and $W_t = (W_t^{(1)}, \cdots, W_t^{(p)})^T$ is a $p$-dimensional standard Brownian motion. We want to estimate the integrated covariance matrix

$$\Sigma_p = \int_0^1 \Theta_t \Theta_t^T dt.$$  

(3.3.2)

It is well-known that the log-price process $X_t$ 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 $X_t$ at synchronous
time points $\tau_{n,\ell}$, $\ell = 0, 1, \ldots , n$. The realized covariance matrix is then defined as

$$\Sigma_p^{RCV} = \sum_{\ell=1}^{n} \Delta X_\ell \Delta X_\ell^T, \quad \text{where } \Delta X_\ell := X_{\tau_{n,\ell}} - X_{\tau_{n,\ell-1}}.$$  

(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 $\Theta_t$ and $W_t$ 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 $Z_t = \left(Z_t^{(1)}, \ldots , Z_t^{(p)}\right)^T$, defined by

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

is constant and non-zero on $(0,1]$ for each $j,k$. Furthermore, the correlation matrix process of $X_t$, defined by

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

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 $X_t$ 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 $\Lambda$ satisfying $\text{tr}(\Lambda \Lambda^T) = p$ such that

$$\Theta_t = \gamma_t \Lambda.$$  

(3.3.4)

The nonlinear shrinkage estimator described in the next section is based on this property.
3.3. Framework and Methodology

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 $\Delta X_\ell$’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 $\Sigma_p^i = \int_0^1 \gamma_t^2 dt \cdot \Lambda \Lambda^T$. [43] proposed a so-called Time-variation adjusted realized covariance matrix, defined as

$$\tilde{\Sigma}_p^i := \frac{\text{tr}(\Sigma_p^{RCV})}{p} \Phi,$$

where $\Phi := \frac{p}{n} \sum_{\ell=1}^n \frac{\Delta X_\ell \Delta X_\ell^T}{\|\Delta X_\ell\|^2},$ (3.3.5)

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

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

$$\Delta X_\ell = \int_{\tau_{n,\ell-1}}^{\tau_{n,\ell}} \gamma_t \Delta W_t \overset{d}{=} \left( \int_{\tau_{n,\ell-1}}^{\tau_{n,\ell}} \gamma_t^2 dt \right)^{1/2} \Phi^{1/2} Z_\ell,$$

where $\overset{d}=$ stands for equal in distribution, and the $Z_\ell$’s are independent random vectors each with $Z_\ell \sim N(0, I_p)$. Then

$$\Phi \overset{d}{=} \frac{1}{n} \sum_{\ell=1}^n \frac{\Phi^{1/2} Z_\ell \Phi^{1/2}}{Z_\ell \Phi Z_\ell / p} = \Phi^{1/2} \left( \frac{1}{n} \sum_{\ell=1}^n \frac{Z_\ell Z_\ell^T}{Z_\ell \Phi Z_\ell / p} \right) \Phi^{1/2}.$$
3.4 Asymptotic Theory and Practical Implementation

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

Following [23], since the $\Delta X_\ell$'s are independent following model (3.3.1), we split the data $\Delta X = (\Delta X_1, \ldots, \Delta X_n)$ into two independent parts, say $\Delta X = (\Delta X^1, \Delta X^2)$, with $\Delta X^i$ having size $p \times n_i$ for $i = 1, 2$, such that $n = n_1 + n_2$. Define

$$\tilde{\Phi}_i = \frac{p}{n_i} \sum_{\ell \in I_i} \frac{\Delta X_\ell \Delta X^T_\ell}{\|\Delta X_\ell\|^2}, \quad (3.3.6)$$

where $I_i = \{\ell : \Delta X_\ell \in \Delta X^i\}$. Carrying out an eigen-analysis on $\tilde{\Phi}_1$ defined in (3.3.6) above, suppose $\tilde{\Phi}_1 = P_1 D_1 P^T_1$. Then we introduce our estimator as

$$\hat{\Sigma}_p := \frac{\text{tr}(\Sigma^{RCV}_p)}{p} \bar{\Phi}, \quad \text{where} \quad \bar{\Phi} := P_1 \text{diag}(P^T_1 \tilde{\Phi}_2 P_1) P^T_1, \quad (3.3.7)$$

with $\text{diag}(\cdot)$ setting all non-diagonal elements of a matrix to 0. The estimator $\hat{\Phi}$ above belongs to a class of rotation equivariant estimator $\Phi(D) = P_1 D P^T_1$, where $D$ is a diagonal matrix, and $P_1$ is the matrix containing all the eigenvectors of $\tilde{\Phi}_1$. The choice of $D = \text{diag}(P^T_1 \tilde{\Phi}_2 P_1)$ comes from solving

$$\min_D \left\| P_1 D P^T_1 - \tilde{\Phi}_2 \right\|_F,$$

where $\|A\|_F = \text{tr}^{1/2}(AA^T)$ is the Frobenius norm of a matrix. Similar to [23], regularization of the eigenvalues in $D = \text{diag}(P^T_1 \tilde{\Phi}_2 P_1)$ comes from the independence between $P_1$ and $\tilde{\Phi}_2$, since $\Delta X^1$ is independent of $\Delta X^2$.

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 $\mu_t = 0$ for $t \in [0, 1]$. All eigenvalues of $\Theta_t \Theta^T_t$ 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 $X_t$, and there exists a constant $C > 0$ such that for all positive integer $n$,

$$\max_{1 \leq \ell \leq n} n (\tau_{n,\ell} - \tau_{n,\ell-1}) \leq C.$$
We set \( \mu_t = 0 \) in Assumption 3.4.1 for the ease of proofs and presentation. If \( \mu_t \) 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 \( \Theta_t \Theta_t^T \) 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, \( \| \Sigma_p \| = O(1) \) uniformly as a result, which are all needed for our results to hold. These assumptions essentially treat \( \gamma_t \) as non-random. Extension to \( \gamma_t \) 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 \( X_t \) in (3.3.1). Then for the estimator \( \hat{\Phi} \) in (3.3.7), writing \( P_1 = (p_{11}, \ldots, p_{1p}) \), if \( p/n \to c > 0 \) and \( \sum_{n_2 \geq 1} pn_2^{-5} < \infty \), we have

\[
\max_{1 \leq i \leq p} \left| \frac{p_{i1}^T \tilde{\Phi}_2 p_{i1}^T - p_{i1}^T \Phi p_{i1}}{p_{i1}^T \Phi p_{i1}} \right| \xrightarrow{a.s.} 0,
\]

where \( \xrightarrow{a.s.} \) represents almost sure convergence.

Since the eigenvalues of \( \tilde{\Phi} \) are the \( p_{i1}^T \tilde{\Phi}_2 p_{i1} \)'s, the above Lemma shows that they are regularized to \( p_{i1}^T \Phi p_{i1} \) asymptotically almost surely, which has values bounded by \( \lambda_{\min}(\Phi) \) and \( \lambda_{\max}(\Phi) \), the minimum and maximum eigenvalues of \( \Phi \) respectively. Assumption 3.4.1 ensures that these eigenvalues are uniformly bounded away from 0 and infinity, and hence \( \tilde{\Phi} \) 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 \( p/n \to c > 0 \), \( \hat{\Sigma}_p \) defined in (3.3.7) is almost surely positive definite.

This is an important result since \( \Sigma_p \) 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 \( \hat{\Sigma}_p \) 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.
3.4. Asymptotic Theory and Practical Implementation

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

\[ \Sigma_{\text{ideal}} = \int_0^1 \gamma_t^2 dt \cdot \text{Pdiag}(\text{P}^T \Phi \text{P}) \text{P}^T. \]

This is similar to the proposed estimator defined in (3.3.7), except that the estimator \( \text{tr}(\Sigma_{\text{RCV}}^p)/p \) is replaced by the population counterpart \( \int_0^1 \gamma_t^2 dt \), while \( \Phi_2 \) is replaced by the population counterpart \( \Phi \). Also, \( \Phi_1 \) is replaced by \( \Phi \), which is the matrix containing all orthonormal eigenvectors for the covariance-type matrix \( \Phi \) 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 \( \Phi \), and assumes the knowledge of \( \Phi \) and \( \int_0^1 \gamma_t^2 dt \).

With this, we define the efficiency loss of any estimator \( \hat{\Sigma} \) as

\[ EL(\Sigma_p, \hat{\Sigma}) := 1 - \frac{L(\Sigma_p, \Sigma_{\text{Ideal}})}{L(\Sigma_p, \hat{\Sigma})}, \]

where \( L(\Sigma_p, \hat{\Sigma}) \) is a loss function for estimating \( \Sigma_p \) by \( \hat{\Sigma} \). We consider the Frobenius loss

\[ L(\Sigma_p, \hat{\Sigma}) = \| \hat{\Sigma} - \Sigma_p \|_F^2, \]  

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

\[ L(\Sigma_p, \hat{\Sigma}) = \text{tr}(\Sigma_p \hat{\Sigma}^{-1}) - \log \text{det}(\Sigma_p \hat{\Sigma}^{-1}) - p. \]  

The class of rotation-equivariant estimator \( \Sigma(D) = \text{PDP}^T \) minimizes the Frobenius norm exactly at \( \Sigma_{\text{Ideal}} \), while similar to Proposition 2 in [23], \( \Sigma_{\text{Ideal}} \) also minimizes the inverse Stein’s loss within such a class of estimator. Hence it is intuitive that our estimator \( \hat{\Sigma}_p \) will be relatively less efficient in the sense that \( EL(\Sigma_p, \hat{\Sigma}_p) > 0 \). It turns out that asymptotically, \( \hat{\Sigma}_p \) is doing as good as \( \Sigma_{\text{Ideal}} \), 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 \( \Phi \). Define \( H_n(v) = p^{-1} \sum_{i=1}^p 1_{\{v_{n,i} \leq v\}} \) 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 Φ.

**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 \( E[L(\Sigma_p, \Sigma_\text{ideal})] \leq 0 \) almost surely with respect to both the Frobenius and the inverse Stein’s loss functions, as long as \( p^{-1}L(\Sigma_p, \Sigma_\text{ideal}) \) does not tend to 0 almost surely.

The requirement \( p^{-1}L(\Sigma_p, \Sigma_\text{ideal}) \) not going to 0 almost surely eliminates the case \( \Sigma_p = \int_0^1 \gamma_t^2 dt \cdot I_p \), when both the loss functions will attain 0 for the ideal estimator. Our estimator will still do a good job in such a case since \( \text{tr}(\Sigma_p RCV_p) \) will still be a good estimator for \( \int_0^1 \gamma_t^2 dt \) by the proof of Theorem 3, while \( \hat{\Phi} \) 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, \( \Delta X_\ell \Delta X_\ell^T / \| \Delta X_\ell \|^2 \) is independent of \( \gamma_t \) and is similar to a data point in constructing a sample covariance matrix, which is independent of each others for different \( \ell \); see Remark 3.3.2 in Section 3.3.1. This observation permits us to permute the data beforehand, say at the \( j \)th permutation, we form a data matrix \( \Delta X^{(j)} = (\Delta X_1^{(j)}, \Delta X_2^{(j)}) \), with \( \Delta X_i^{(j)} \) having size \( p \times n_i \) for \( i = 1, 2 \), such that \( n = n_1 + n_2 \). Then we construct

\[
\tilde{\Phi}_i^{(j)} = \frac{p}{n_i} \sum_{\ell \in I_i^{(j)}} \Delta X_\ell \Delta X_\ell^T / \| \Delta X_\ell \|^2,
\]

where \( I_i^{(j)} = \{ \ell : \Delta X_\ell \in \Delta X_i^{(j)} \} \), and perform eigen-analysis on \( \tilde{\Phi}_i^{(j)} \), say \( \tilde{\Phi}_i^{(j)} = P_i^{(j)} D_i^{(j)} P_i^{(j)\top} \).

The we can form the \( j \)th estimator as

\[
\hat{\Sigma}_p^{(j)} := \frac{\text{tr}(\Sigma_p RCV_p)}{p} \tilde{\Phi}_i^{(j)}, \quad \text{where} \quad \tilde{\Phi}_i^{(j)} := P_i^{(j)} \text{diag}(P_1^{(j)} \tilde{\Phi}_2^{(j)} P_2^{(j)} P_1^{(j)}) P_i^{(j)\top}.
\]

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

\[
\hat{\Sigma}_{p,M} := \frac{1}{M} \sum_{j=1}^M \hat{\Sigma}_p^{(j)}.
\]

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:

\[
g(m) = \left\| \frac{1}{M} \sum_{j=1}^M (\tilde{\Phi}_p^{(j)} - \tilde{\Phi}_2^{(j)}) \right\|_F^2.
\]
where \( \tilde{\Phi}_2^{(j)} \) is defined in (3.4.3) and \( \hat{\Phi}^{(j)} \) 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 \( EL(\Sigma_p, \hat{\Sigma}_{p,M}) \leq 0 \) almost surely with respect to both the Frobenius and the inverse Stein’s loss functions, as long as \( p^{-1}L(\Sigma_p, \hat{\Sigma}_{\text{ideal}}) \) does not tend to 0 almost surely.

In practice, the estimator \( \hat{\Sigma}_{p,M} \), 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 \( \Sigma_p = \int_0^1 \gamma_t^2 dt I_p \) 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 \( \sum_{n_2 \geq 1} pn_2^{-5} < \infty \) 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 \( \{X_t\} \) defined in (3.3.1), with \( \mu_t = 0 \) and \( \Theta_t = \gamma_t \Lambda \) as in (3.3.4). One has \( \gamma_t \) being piecewise constant, the other has \( \gamma_t \) being continuous, detailed as follows:
3.5. Empirical Results

Design I: Piecewise constants. We take $\gamma_t$ 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 $\gamma_t$ to be

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

We assume $\Lambda = I_p$ and the observation times are taken to be equidistant, where $\tau_{n, \ell} = \ell/n, \ell = 1, \ldots, n$. We generate $\{X_t\}$ 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 $\Sigma_p$, the minimum variance portfolio solves

$$
\min_{w: w^T 1_p = 1} w^T \Sigma_p w,
$$

where $1_p$ is a vector of $p$ ones. The solution to the above is given by

$$
w_{\text{opt}} = \frac{\Sigma_p^{-1} 1_p}{1_p^T \Sigma_p^{-1} 1_p}.
$$

(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 $\hat{\Sigma}_{p,M}$. 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 $A$ to be $\text{Cond}(A) = \lambda_{\max}(A)/\lambda_{\min}(A)$.

**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,

$$
p^{1/2} \| \hat{w}_{\text{opt}} \|_{\max} \leq \text{Cond}(\Phi), \quad pR(\hat{w}_{\text{opt}}) \leq \text{Cond}^2(\Phi) \lambda_{\max}(\Sigma_p),
$$

where...
### 3.5. Empirical Results

#### 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.

<table>
<thead>
<tr>
<th>Design</th>
<th>Proposed</th>
<th>Time variation-adjusted</th>
<th>Realized covariance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Frobenius loss</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$p = 100$</td>
<td>.13(02)</td>
<td>2.8(04)</td>
</tr>
<tr>
<td></td>
<td>$p = 200$</td>
<td>.55(17)</td>
<td>69(31)</td>
</tr>
<tr>
<td></td>
<td>Inverse Stein’s loss</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$p = 100$</td>
<td>.17(014)</td>
<td>5.63(058)</td>
</tr>
<tr>
<td></td>
<td>$p = 200$</td>
<td>88(15)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Frobenius loss</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$p = 100$</td>
<td>.29(03)</td>
<td>6.32(09)</td>
</tr>
<tr>
<td></td>
<td>$p = 200$</td>
<td>.38(03)</td>
<td>13(10)</td>
</tr>
<tr>
<td></td>
<td>Inverse Stein’s loss</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$p = 100$</td>
<td>.54(1)</td>
<td>693(31)</td>
</tr>
<tr>
<td></td>
<td>$p = 200$</td>
<td>88(16)</td>
<td>-</td>
</tr>
</tbody>
</table>
where \( \hat{w}_{opt} \) is the weight in (3.5.1) with \( \Sigma_p \) substituted by \( \hat{\Sigma}_{p,M} \) defined in (3.4.5). The function \( R(\hat{w}_{opt}) = \hat{w}_{opt}^T \Sigma_p \hat{w}_{opt} \) represents the actual risk when investing using \( \hat{w}_{opt} \) 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 \( \sigma_t^{(i)} = \exp(\phi_t^{(i)}) \) follows the independent Ornstein-Uhlenbeck process

\[
d\phi_t^{(i)} = \alpha^{(i)} (\beta_0^{(i)} - \phi_t^{(i)}) dt + \beta_1^{(i)} dU_t^{(i)},
\]

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 \( \phi_0^{(i)} = 0 \).

We simulate the trading times independently from the price data assuming the transaction times for each stock follow independent Poisson processes with rates \( \lambda_1, \ldots, \lambda_{100} \) 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
3.5. Empirical Results

<table>
<thead>
<tr>
<th>Theoretical risk</th>
<th>( \text{Proposed Time variation-adjusted Realized covariance} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual risk</td>
<td>.922 3.918 4.115</td>
</tr>
<tr>
<td>Perceived risk</td>
<td>.753 3.869 4.034</td>
</tr>
</tbody>
</table>

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

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 \( R(\hat{w}_{\text{opt}}) = w_{\text{opt}}^{T} \Sigma_p w_{\text{opt}} \), where \( w_{\text{opt}} \) 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 \( \Sigma_p \) is the true integrated covariance matrix over the 5-day investment period. The second one is the actual risk \( R(\hat{w}_{\text{opt}}) = \hat{w}_{\text{opt}}^{T} \Sigma_p \hat{w}_{\text{opt}} \), where \( \hat{w}_{\text{opt}} \) is calculated using different integrated covariance matrix estimators. Finally the perceived risk is defined by \( \hat{R}(\hat{w}_{\text{opt}}) = \hat{w}_{\text{opt}}^{T} \hat{\Sigma}_p \hat{w}_{\text{opt}} \).

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
3.5. Empirical Results

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 $\hat{\sigma}$, together with the annualized portfolio return $\hat{\mu}$ and the Sharpe ratio $\hat{\mu}/\hat{\sigma}$ to gauge the performance of each method. For 20-day training windows and 5 day re-evaluation period, $\hat{\mu}$ and $\hat{\sigma}$ are defined by

$$\hat{\mu} = 52 \times \frac{1}{45} \sum_{i=5}^{49} w_i^T r_i, \quad \hat{\sigma} = \left( 52 \times \frac{1}{45} \sum_{i=5}^{49} (w_i^T r_i - \hat{\mu})^2 \right)^{1/2}. $$

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, $\hat{\mu}$ and $\hat{\sigma}$ are defined by

$$\hat{\mu} = 252 \times \frac{1}{240} \sum_{i=6}^{245} w_i^T r_i, \quad \hat{\sigma} = \left( 252 \times \frac{1}{240} \sum_{i=6}^{245} (w_i^T r_i - \hat{\mu})^2 \right)^{1/2}. $$

On top of all the above, we also report the mean maximum exposure $\|\hat{W}_{\text{opt}}\|_{\max}$ 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
### 3.5. Empirical Results

#### 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.

<table>
<thead>
<tr>
<th></th>
<th>20-day training window</th>
<th>5-day training window</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Proposed (%)</td>
<td>Time variation adjusted(%)</td>
</tr>
<tr>
<td><strong>Re-evaluate every 5 days</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Annualized return</td>
<td>5.50</td>
<td>5.87</td>
</tr>
<tr>
<td>Annualized out-of-sample SD</td>
<td>11.07</td>
<td>11.16</td>
</tr>
<tr>
<td>Sharpe ratio</td>
<td>49.66</td>
<td>52.62</td>
</tr>
<tr>
<td>Mean maximum exposure</td>
<td>15.19 (3.26)</td>
<td>19.82 (1.74)</td>
</tr>
</tbody>
</table>
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 estimation 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 demonstrated 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

$$\Delta X_\ell = \left( \int_{\tau_{n,\ell-1}}^{\tau_{n,\ell}} \gamma_t^2 dt \right)^{1/2} \Phi^{1/2} Z_\ell,$$

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

$$\frac{p_{1i}^T \hat{\Phi} p_{1i} - p_{1i}^T \Phi p_{1i}}{p_{1i}^T \Phi p_{1i}} = I_{i1} + I_{i2},$$

where

$$I_{i1} = \frac{p_{1i}^T \Phi p_{2i} - n_2^{-1} \sum_{\ell \in I_2} (p_{1i}^T \Phi^{1/2} Z_\ell)^2}{p_{1i}^T \Phi p_{1i}}, \quad I_{i2} = \frac{n_2^{-1} \sum_{\ell \in I_2} (p_{1i}^T \Phi^{1/2} Z_\ell)^2 - p_{1i}^T \Phi p_{1i}}{p_{1i}^T \Phi p_{1i}}.$$
First of all, since we can write
\[ n_2^{-1} \sum_{\ell \in I_2} (p_{i}^T \Phi^{1/2} Z_{\ell})^2 = p_{i}^T (n_2^{-1} \sum_{\ell \in I_2} \Phi^{1/2} Z_{\ell} Z_{\ell}^T \Phi^{1/2}) p_{i}, \]
with \( n_2^{-1} \sum_{\ell \in I_2} \Phi^{1/2} Z_{\ell} Z_{\ell}^T \Phi^{1/2} \) a proper sample covariance matrix for estimating \( \Phi \), we can use Lemma 1 of [23] to conclude that
\[ \max_{1 \leq i \leq p} |I_{i}| \overset{a.s.}{\rightarrow} 0. \]
Hence it remains to show that \( \max_{1 \leq i \leq p} |I_{i}| \overset{a.s.}{\rightarrow} 0 \). To this end, consider
\[
\max_{1 \leq i \leq p} |I_{i}| = \max_{1 \leq i \leq p} \left| \frac{n_2^{-1} \sum_{\ell \in I_2} \left( \frac{p_{i}^T \Phi^{1/2} Z_{\ell}}{Z_{\ell}^T \Phi Z_{\ell}/p} \right)^2 - p_{i}^T \left( n_2^{-1} \sum_{\ell \in I_2} \Phi^{1/2} Z_{\ell} Z_{\ell}^T \Phi^{1/2} \right) p_{i}}{p_{i}^T \Phi p_{i}} \right| \\
\leq \max_{\ell \in I_2} \left| \frac{1}{Z_{\ell}^T \Phi Z_{\ell}/p} - 1 \right| \cdot \left( 1 + \max_{1 \leq i \leq p} |I_{i}| \right) \overset{a.s.}{\rightarrow} 0,
\]
if we can show further that \( \max_{\ell \in I_2} \left| \frac{1}{Z_{\ell}^T \Phi Z_{\ell}/p} - 1 \right| \overset{a.s.}{\rightarrow} 0. \)

To show this, using Lemma 2.7 of [8], we have
\[
E(\Phi^{1/2} Z_{\ell} - \text{tr}(\Phi)) \leq K_6 \left( E^{3} |z_{\ell,1}|^4 \text{tr}(\Phi^2) + E |z_{\ell,1}|^{12} \text{tr}(\Phi^6) \right),
\]
where \( K_6 \) is a constant independent of \( \ell, n \) and \( p \). This implies that, since \( \text{tr}(\Phi) = p \),
\[
E \left( \max_{\ell \in I_2} \left| \frac{Z_{\ell}^T \Phi Z_{\ell}}{p} - 1 \right|^{6} \right) \leq n_2 \cdot K_6 \left( E^{3} |z_{\ell,1}|^4 \frac{\text{tr}(\Phi^2)}{p^6} + E |z_{\ell,1}|^{12} \frac{\text{tr}(\Phi^6)}{p^6} \right) = O(n_2 p^{-3}).
\]
The rate in the last line comes from Assumption 3.4.1 that \( \Theta_i \Theta_i^T = \gamma_i^2 \Phi \) has all its eigenvalues uniformly bounded away from 0 and infinity, so that \( \text{tr}(\Phi^2) = O(p^3) \) and \( \text{tr}(\Phi^6) = O(p) \), 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 \( O(n_2 p^{-3}) = O(n^{-5/2}) \). Since \( \sum_{n \geq 1} n^{-5/2} < \infty \), through the Borel-Cantelli lemma, we have proved that \( \max_{\ell \in I_2} \left| \frac{Z_{\ell}^T \Phi Z_{\ell}}{p} - 1 \right| \overset{a.s.}{\rightarrow} 0 \), meaning that \( \max_{\ell \in I_2} \left| \frac{1}{Z_{\ell}^T \Phi Z_{\ell}/p} - 1 \right| \overset{a.s.}{\rightarrow} 0 \) as well. This completes the proof of the lemma. □

**Proof of Theorem 3.** Firstly, by Lemma 2, \( \hat{\Phi} \) defined in (3.3.7) is almost surely positive definite since all its eigenvalues are almost surely \( p_{i}^T \Phi p_{i} \) 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 
\[
\text{tr}(\Sigma^Rcv)_p / p \xrightarrow{a.s.} 0.
\]

To this end, consider
\[
\left| \frac{\text{tr}(\Sigma^Rcv)_p}{p} - \int_0^1 \gamma_i^2 dt \right| = \left| \frac{\sum_{\ell=1}^n \Delta X^\ell \Delta X}{p} - \int_0^1 \gamma_i^2 dt \right|
\]
\[
= \left| \frac{\sum_{\ell=1}^n \int_{\gamma_{\ell-1}}^{\gamma_{\ell}} \gamma_i^2 dt \cdot Z^\ell \Phi Z^\ell}{p} - \int_0^1 \gamma_i^2 dt \right|
\]
\[
\leq \max_{1 \leq \ell \leq n} \left| \frac{Z^\ell \Phi Z^\ell}{p} - 1 \right| \int_0^1 \gamma_i^2 dt.
\]
From the proof of \( \max_{1 \leq \ell \leq n} \left| \frac{Z^\ell \Phi Z^\ell}{p} - 1 \right| \xrightarrow{a.s.} 0 \) in the last part of the proof of Lemma 2, we can replace \( n_2 \) there by \( n \) and conclude that
\[
E\left( \max_{1 \leq \ell \leq n} \left| \frac{Z^\ell \Phi Z^\ell}{p} - 1 \right| 6 \right) \leq n \cdot K_6 \left( E\left| z_{\ell,1} \right|^4 \frac{\text{tr}(\Phi_2)}{p^6} + E\left| z_{\ell,1} \right|^{12} \frac{\text{tr}(\Phi^6)}{p^6} \right)
\]
\[
= O(np^{-3}) = O(n^{-2}).
\]
Since \( \sum_{n \geq 1} n^{-2} < \infty \), we have proved that \( \max_{1 \leq \ell \leq n} \left| \frac{Z^\ell \Phi Z^\ell}{p} - 1 \right| \xrightarrow{a.s.} 0 \). This shows that 
\[
\text{tr}(\Sigma^Rcv)_p / p \xrightarrow{a.s.} \int_0^1 \gamma_i^2 dt,
\]
which is uniformly larger than 0 by Assumption 3.3.1. This completes the proof of the theorem. \( \square \)

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 \( v^{(1)}_1 \geq \cdots \geq v^{(1)}_p \) the eigenvalues of \( \tilde{\Phi}_1 \) defined in (3.3.6) with corresponding eigenvectors \( p_1, \ldots, p_p \), and \( v_1 \geq \cdots \geq v_p \) the eigenvalues of \( \Phi \) defined in (3.3.5) with corresponding eigenvectors \( p_1, \ldots, p_p \). Then there exist positive functions \( \delta_1(\cdot) = \delta(\cdot) \)

and distribution functions \( F_1 = F \) such that
\[
p^{-1} \sum_{j=1}^p \mathbb{1}_{\{x \geq v^{(1)}_j\}} \xrightarrow{a.s.} F_i(x), \quad p^{-1} \sum_{j=1}^p \mathbb{1}_{\{x \geq v_j\}} \xrightarrow{a.s.} F(x),
\]

\( \delta(\cdot) \) are given here, since they are not important for the proof of any subsequent theorems. For any \( \lambda \in \mathbb{R} \), \( \delta(\lambda) = \begin{cases} \frac{\lambda}{(c-1)\lambda_{\text{min}}}, & \text{if } \lambda > 0; \\
\frac{1}{(c-1)\lambda_{\text{min}}}, & \text{if } \lambda = 0 \text{ and } c > 1; \\
0, & \text{otherwise.} \end{cases} \)

\[
\delta_1(\lambda) = \begin{cases} \frac{\lambda}{(\lambda - c)\lambda_{\text{min}}}^2, & \text{if } \lambda > 0; \\
\frac{1}{(c-1)\lambda_{\text{min}}}, & \text{if } \lambda = 0 \text{ and } c > 1; \\
0, & \text{otherwise.} \end{cases}
\]
Proof of Lemma 3.7.1. Write $\Delta X_\ell = (f_{n,\ell}^\tau \gamma_i^2 dt)^{1/2} \Phi^{1/2}Z_\ell$ as in Remark 3.3.2. Define for $i = 1, 2,$

$$\tilde{\Phi}_{i,\text{sam}} = n_i^{-1} \sum_{\ell \in I_i} \Phi^{1/2}Z_i \Phi^{1/2}, \quad \Phi_{\text{sam}} = n^{-1} \sum_{\ell = 1}^n \Phi^{1/2}Z_\ell \Phi^{1/2},$$

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

\begin{align}
& p^{-1}\text{tr}((\Phi_{1,\text{sam}} - zI_p)^{-1}) - p^{-1}\text{tr}((\Phi_{1} - zI_p)^{-1}) \xrightarrow{a.s.} 0, \\
& p^{-1}\text{tr}((\Phi_{\text{sam}} - zI_p)^{-1}) - p^{-1}\text{tr}((\Phi - zI_p)^{-1}) \xrightarrow{a.s.} 0, \quad (3.7.1) \\
& p^{-1}\text{tr}((\Phi_{1,\text{sam}} - zI_p)^{-1}) - p^{-1}\text{tr}((\Phi_{1} - zI_p)^{-1}) \xrightarrow{a.s.} 0, \\
& p^{-1}\text{tr}((\Phi_{\text{sam}} - zI_p)^{-1}) - p^{-1}\text{tr}((\Phi - zI_p)^{-1}) \xrightarrow{a.s.} 0.
\end{align}

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{align}
& p^{-1}\sum_{j=1}^p 1_{\{x \geq v^{(i)}_{j,\text{sam}}\}} - p^{-1}\sum_{j=1}^p 1_{\{x \geq v^{(i)}_{j}\}} \xrightarrow{a.s.} 0, \\
& p^{-1}\sum_{j=1}^p 1_{\{x \geq v_{j,\text{sam}}\}} - p^{-1}\sum_{j=1}^p 1_{\{x \geq v_{j}\}} \xrightarrow{a.s.} 0, \quad (3.7.2) \\
& p^{-1}\sum_{j=1}^p p_{1,j,\text{sam}}^T \Phi p_{1,j,\text{sam}} 1_{\{x \geq v^{(i)}_{j,\text{sam}}\}} - p^{-1}\sum_{j=1}^p p_{1,j}^T \Phi p_{1,j} 1_{\{x \geq v^{(i)}_{j}\}} \xrightarrow{a.s.} 0, \\
& p^{-1}\sum_{j=1}^p p_{j,\text{sam}}^T \Phi p_{j,\text{sam}} 1_{\{x \geq v_{j,\text{sam}}\}} - p^{-1}\sum_{j=1}^p p_{j}^T \Phi p_{j} 1_{\{x \geq v_{j}\}} \xrightarrow{a.s.} 0.
\end{align}

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{align*}
& p^{-1}\sum_{j=1}^p 1_{\{\lambda \geq v^{(i)}_{j,\text{sam}}\}} \xrightarrow{a.s.} F_1(\lambda), \\
& p^{-1}\sum_{j=1}^p 1_{\{\lambda \geq v_{j,\text{sam}}\}} \xrightarrow{a.s.} F(\lambda),
\end{align*}
\[ p^{-1} \sum_{j=1}^{p} p_{j, \text{sam}}^T \Phi p_{1, \text{sam}} 1_{\{ x \geq v_{j, \text{sam}}^{(1)} \}} \overset{a.s.}{\rightarrow} \int_{-\infty}^{x} \delta_1(\lambda) dF(\lambda), \]
\[ p^{-1} \sum_{j=1}^{p} p_{j, \text{sam}} \Phi p_{j, \text{sam}} 1_{\{ x \geq v_{j, \text{sam}} \}} \overset{a.s.}{\rightarrow} \int_{-\infty}^{x} \delta(\lambda) dF(\lambda). \]

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 \( k = 0, 1, \)
\[ p^{-1} \text{tr}((\tilde{\Phi}_1 - z I_p)^{-1} \Phi^k) = p^{-1} \text{tr}((I_p - (\tilde{\Phi}_{1, \text{sam}} - z I_p)^{-1}(\tilde{\Phi}_{1, \text{sam}} - \tilde{\Phi}_1))^{-1}(\tilde{\Phi}_{1, \text{sam}} - z I_p)^{-1} \Phi^k) \]
\[ = p^{-1} \text{tr}((\tilde{\Phi}_{1, \text{sam}} - z I_p)^{-1} \Phi^k) + \text{R, with} \]
\[ \text{R} = \sum_{j \geq 1} p^{-1} \text{tr}([(\tilde{\Phi}_{1, \text{sam}} - z I_p)^{-1}(\tilde{\Phi}_{1, \text{sam}} - \tilde{\Phi}_1)]^j(\tilde{\Phi}_{1, \text{sam}} - z I_p)^{-1} \Phi^k), \]
where \( \text{R} \) comes from a Neumann series expansion. Such an expansion is valid since
\[ r := \|(\tilde{\Phi}_{1, \text{sam}} - z I_p)^{-1}(\tilde{\Phi}_{1, \text{sam}} - \tilde{\Phi}_1)\| \leq \|(\tilde{\Phi}_{1, \text{sam}} - z I_p)^{-1}\| \cdot \|\tilde{\Phi}_{1, \text{sam}} - \tilde{\Phi}_1\| \]
\[ \leq \frac{1}{|\text{Im}(z)|} \cdot \max_{\ell \in I_1} \left| \frac{1}{Z_{\ell}^{1/2} \Phi Z_{\ell}/p} - 1 \right| \|\tilde{\Phi}_{1, \text{sam}}\| \]
\[ \leq \frac{\|\Phi^{1/2}\|^2}{|\text{Im}(z)|} \cdot \max_{\ell \in I_1} \left| \frac{1}{Z_{\ell}^{1/2} \Phi Z_{\ell}/p} - 1 \right| \left| n_1^{-1} \sum_{\ell \in I_1} Z_{\ell} Z_{\ell}^{*} \right| \]
\[ \leq \frac{\|\Phi^{1/2}\|^2 (1 + \sqrt{c})^2}{|\text{Im}(z)|} \cdot \max_{\ell \in I_1} \left| \frac{1}{Z_{\ell}^{1/2} \Phi Z_{\ell}/p} - 1 \right| \overset{a.s.}{\rightarrow} 0, \]
where we used Lemma S.2 of [24] to conclude \( \|\tilde{\Phi}_{1, \text{sam}} - z I_p\|^{-1} \leq 1/|\text{Im}(z)| \), and we used Theorem 5.11 of [7] to conclude that \( \|n_1^{-1} \sum_{\ell \in I_1} Z_{\ell} Z_{\ell}^{*}\| \leq (1 + \sqrt{c})^2 \) almost surely, since \( p/n_1 \), like \( p/n \), goes to \( c > 0 \) also. Finally, the term \( \max_{\ell \in I_1} \left| \frac{1}{Z_{\ell}^{1/2} \Phi Z_{\ell}/p} - 1 \right| \overset{a.s.}{\rightarrow} 0 \) by the last part of the proof of Lemma 2. With this,
\[ |\text{R}| \leq \sum_{j \geq 1} r^j \|\Phi^k\| \cdot \|(\tilde{\Phi}_{1, \text{sam}} - z I_p)^{-1}\| \leq \frac{r \|\Phi^k\|}{(1 - r)|\text{Im}(z)|} \overset{a.s.}{\rightarrow} 0, \]
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 \( \tilde{\Phi}_{1, \text{sam}} \) by \( \tilde{\Phi}_{\text{sam}} \) and \( \tilde{\Phi}_1 \) by \( \tilde{\Phi} \). This completes the proof of the lemma. □
Proof of Theorem 4. Define \( \theta = \int_0^1 \gamma_t^2 dt \), \( \hat{\theta} = \text{tr}(\Sigma_p^{\text{RCV}})/p \) and \( \Phi_{\text{ideal}} = P \text{diag}(P^T \Phi P)P^T \).

We can easily see that

\[
EL(\Sigma_p, \hat{\Sigma}_p) \leq 1 - \left( \frac{p^{-1/2}\| (\hat{\theta} - \theta) \hat{\Phi} \|_F}{p^{-1/2}\| \Phi_{\text{ideal}} - \Phi \|_F} + \frac{p^{-1/2}\| \hat{\Phi} - \Phi \|_F}{p^{-1/2}\| \Phi_{\text{ideal}} - \Phi \|_F} \right)^{-2},
\]

where \( p^{-1/2}\| (\hat{\theta} - \theta) \hat{\Phi} \|_F \leq |\hat{\theta} - \theta| \max_{1 \leq i \leq p} p_i^T \tilde{\Phi}_i p_i \xrightarrow{a.s.} 0 \) by Lemma 2 and the proof of Theorem 3. Since we are equivalently assuming that \( p^{-1/2}\| \Sigma_{\text{ideal}} - \Phi \|_F \) is not going to 0, it remains to show that the rightmost term above is going to 1. To this end, observe that

\[
p^{-1}\| \hat{\Phi} - \Phi \|_F^2 = \frac{p^{-1}\sum_{i=1}^p (p_i^T \tilde{\Phi}_i p_i - p_i^T \Phi p_i)^2}{p^{-1}\| \Phi_{\text{ideal}} - \Phi \|_F^2} + \frac{p^{-1}\| P \text{diag}(P^T \Phi P)P^T - \Phi \|_F^2}{p^{-1}\| \Phi_{\text{ideal}} - \Phi \|_F^2}.
\]

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

\[
p^{-1}\| P \text{diag}(P^T \Phi P)P^T - \Phi \|_F^2 = p^{-1}\text{tr}(\Phi^2) - p^{-1}\sum_{i=1}^p (p_i^T \Phi p_i)^2 \xrightarrow{a.s.} \int \tau^2 dH(\tau) - \int \delta^2(\lambda) dF_1(\lambda) = \int \tau^2 dH(\tau) - \int \delta^2(\lambda) dF_3(\lambda),
\]

which is non-zero if \( \Phi \neq \mathbf{I}_p \). This is also the almost sure limit of \( p^{-1}\| \Phi_{\text{ideal}} - \Phi \|_F^2 \), and hence the rightmost term in (3.7.3) is indeed going to 1 if we can also show that \( p^{-1}\sum_{i=1}^p (p_i^T \tilde{\Phi}_i p_i - p_i^T \Phi p_i)^2 \xrightarrow{a.s.} 0 \). By Lemma 2,

\[
p^{-1}\sum_{i=1}^p (p_i^T \tilde{\Phi}_i p_i - p_i^T \Phi p_i)^2 \leq \max_{1 \leq i \leq p} \left| \frac{p_i^T \tilde{\Phi}_i p_i - p_i^T \Phi p_i}{p_i^T \Phi p_i} \right|, \quad \max_{1 \leq i \leq p} p_i^T \Phi p_i \xrightarrow{a.s.} 0.
\]

This completes the proof for the Frobenius loss.

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

\[
p^{-1}L(\Sigma_p, \Sigma_{\text{ideal}}) = p^{-1}\sum_{i=1}^p \log(p_i^T \Phi p_i) - p^{-1}\sum_{i=1}^p \log(v_{n,i}) \xrightarrow{a.s.} \int \log(\delta(\lambda)) dF(\lambda) - \int \log(\tau) dH(\tau),
\]

where \( v_{n,i} \) is the \( i \)-th largest eigenvalue of \( \Phi \). Now consider the decomposition

\[
p^{-1}L(\Sigma_p, \hat{\Sigma}_p) = I_1 + I_2 + I_3 + I_4 + I_5, \quad \text{where}
\]

\[
I_1 = \log(\hat{\theta}/\theta),
\]

\[
I_2 = \left( \frac{\theta}{\hat{\theta}} - 1 \right) p^{-1}\sum_{i=1}^p \left( \frac{p_i^T \Phi p_i}{p_i^T \tilde{\Phi}_i p_i} \right),
\]

\[
I_3 = \left( \frac{\theta}{\hat{\theta}} - 1 \right) p^{-1}\sum_{i=1}^p \left( \frac{p_i^T \Phi p_i}{p_i^T \tilde{\Phi}_i p_i} \right),
\]

\[
I_4 = \left( \frac{\theta}{\hat{\theta}} - 1 \right) p^{-1}\sum_{i=1}^p \left( \frac{p_i^T \Phi p_i}{p_i^T \tilde{\Phi}_i p_i} \right),
\]

\[
I_5 = \left( \frac{\theta}{\hat{\theta}} - 1 \right) p^{-1}\sum_{i=1}^p \left( \frac{p_i^T \Phi p_i}{p_i^T \tilde{\Phi}_i p_i} \right).
\]
\[ I_3 = p^{-1} \sum_{i=1}^{p} \left( \frac{p_i^T \Phi p_i}{p_i^T \Phi_2 p_i} \right) - 1, \]
\[ I_4 = p^{-1} \sum_{i=1}^{p} \log \left( \frac{p_i^T \Phi p_i}{p_i^T \Phi_2 p_i} \right), \]
\[ I_5 = p^{-1} \sum_{i=1}^{p} \log(p_i^T \Phi p_i) - p^{-1} \sum_{i=1}^{p} \log(v_{n,i}). \]

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
\[ I_5 \xrightarrow{a.s.} \int \log(\delta(\lambda))dF(\lambda) - \int \log(\tau)dH(\tau), \]
so that \( p^{-1}L(\Sigma, \Sigma_{\text{ideal}})/p^{-1}L(\Sigma, \hat{\Sigma}_p) \xrightarrow{a.s.} 1 \), showing \( EL(\Sigma, \hat{\Sigma}_p) \xrightarrow{a.s.} 0 \). This completes the proof of the theorem. \( \square \)

**Proof of Theorem 5.** For the Frobenius loss,
\[ \| \hat{\Sigma}_{p,M} - \Sigma_p \|_F^2 = \| \frac{1}{M} \sum_{i=1}^{M} (\hat{\Sigma}_p^{(i)} - \Sigma_p) \|_F^2 \leq \left( \frac{1}{M} \sum_{i=1}^{M} \| \hat{\Sigma}_p^{(i)} - \Sigma_p \|_F \right)^2 \]
\[ \leq \frac{1}{M} \sum_{i=1}^{M} \| \hat{\Sigma}_p^{(i)} - \Sigma_p \|_F^2, \]
so that
\[ EL(\Sigma_p, \hat{\Sigma}_{p,M}) \leq 1 - \frac{\| \hat{\Sigma}_{\text{ideal}} - \Sigma_p \|_F^2}{\| \Sigma_{\text{ideal}} - \Sigma_p \|_F^2} = 1 - \frac{1}{\frac{1}{M} \sum_{i=1}^{M} \| \Sigma_p^{(i)} - \Sigma_p \|_F^2} \xrightarrow{a.s.} 0, \]

since \( EL(\Sigma_p, \hat{\Sigma}_p^{(i)}) \xrightarrow{a.s.} 0 \) 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. \( \square \)

**Proof of Theorem 6.** We have
\[ p^{1/2} \| w_{\text{opt}} \|_{\infty} \leq p^{1/2} \| \Sigma_p^{-1} \|_{1} \leq p^{1/2} \| \Phi^{-1} \|_{1} \leq p^{1/2} \cdot p^{1/2} \lambda_{\min}^{-1}(\hat{\Phi}) \]
\[ = \lambda_{\max}(\hat{\Phi}) \leq \max_{1 \leq i \leq p} p_i \Phi_{2i} \xrightarrow{a.s.} \max_{1 \leq i \leq p} p_i \Phi p_i \]
\[ = \lambda_{\max}(\Phi) \leq \lambda_{\min}(\Phi) \leq \lambda_{\max}(\Phi) = \text{Cond}(\Phi), \]
where we used the results of Lemma 2 for the almost sure convergence. For the actual risk bound,

$$pR(\hat{\Sigma}_p) = \frac{p1_p^T\hat{\Sigma}_p^{-1}\Sigma_p\hat{\Sigma}_p^{-1}1_p}{(1_p^T\hat{\Sigma}_p^{-1}1_p)^2} \leq \frac{p \cdot p \lambda^2_{\text{max}}(\hat{\Sigma}_p) \lambda_{\text{max}}(\Sigma_p)}{p^2 \lambda^2_{\text{min}}(\hat{\Sigma}_p)} = \left(\frac{\lambda_{\text{max}}(\hat{\Phi})}{\lambda_{\text{min}}(\hat{\Phi})}\right)^2 \lambda_{\text{max}}(\Sigma_p)$$

$$\left(\frac{\max_{1 \leq i \leq p} p_i^T\hat{\Phi}_2 p_i}{\min_{1 \leq i \leq p} p_i^T\hat{\Phi}_2 p_i}\right)^2 \lambda_{\text{max}}(\Sigma_p) \xrightarrow{a.s.} \left(\frac{\max_{1 \leq i \leq p} p_i^T\Phi_1 p_i}{\min_{1 \leq i \leq p} p_i^T\Phi_1 p_i}\right)^2 \lambda_{\text{max}}(\Sigma_p)$$

$$\leq \left(\frac{\lambda_{\text{max}}(\Phi)}{\lambda_{\text{min}}(\Phi)}\right)^2 \lambda_{\text{max}}(\Sigma_p) = \text{Cond}^2(\Phi) \lambda_{\text{max}}(\Sigma_p).$$

This completes the proof of the theorem. □
Bibliography


