Large Covariance Matrix Estimation Based on POET with Application to Risk Estimation in Finance

A PROJECT
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA DULUTH
BY

Ying Liu

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

Advisors: Kang James, Barry James

June 2015

Acknowledgements

First of all, I would love to thank Dr. Kang James and Dr. Barry James for their expert advice and extraordinary support throughout this challenging project. I would also thank Dr. Xuan Li for his encouragement during this process and serving as my committee member. I have gained so much from the mathematical and statistical courses offered by Department of Mathematic and Statistics. This project would have been impossible without these courses and the professors who have instructed me.

Dedication

This project is dedicated to my husband, Yi Xiao, who always supports me.

Abstract

This project discusses statistical methods for estimating complex correlation structures of high-dimensional datasets in financial research. We study the Principal Orthogonal complEment Thresholding (POET) method, and using Monte Carlo simulation we compare the performance of the POET estimator with the usual sample covariance matrix. The POET estimator performs better when the dimension of the data is large relative to the sample size. This method is applied to the Shanghai and Shenzhen Stock Exchange markets.

Table of Contents

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Portfolio risk can be described as the potential loss of a portfolio. Estimating the risk of a portfolio is very important in financial research, which is the foundation of risk management and asset pricing. The variance of return measures how far the return is spread out. In this project, we use the variance of return to measure the risk of large portfolios.

The variance matrix of the return of stocks in a portfolio is crucial. Based on the covariance matrix of the stock returns in the portfolio and the asset allocation, the variance of the portfolio can be obtained. The sample covariance matrix is employed to estimate the covariance matrix in most cases, and in some circumstances it works well. But as the number of stocks in the portfolio grows, it becomes challenging to estimate the covariance matrix accurately with the sample covariance estimation.

We assess the portfolio risk based on the Principal Orthogonal ComplEment Thresholding (POET) method, which is proposed by Fan et al.[4]. In this project, we assume that the number of factors influencing the risk is known. The POET method is based on the factor model and sparse matrix estimation through thresholding procedures. We also use Monte Carlo simulation based on Fan, Liao and Shi [7] to compare the performance of the sample covariance estimator and the POET estimator. Last, we apply the POET method to assess the risk of portfolios based on historical trading data in China’s stock markets and analyze the results.

Chapter 2

Notation and some definitions

There are two main sources of risk in financial markets: systematic risk, which is the risk inherent in the entire stock market such as interest rate, exchange rate, depression, etc.; and idiosyncratic risk, which refers to the risk that is embedded in certain group of stocks, such as the price of the resource for a particular industry, and special regulations. Almost all the stocks in the financial market would be affected by the systematic risk, while certain idiosyncratic risk could only affect one or a few stocks. Unlike systematic risk, idiosyncratic risk can be effectively reduced through diversification.

2.1 Notation in this project

The notations that we use are as follow:

Let p be the number of stocks in the portfolio. Let T be the number of time points we observe, which is also the number of observations in the dataset. Let $\boldsymbol{\Sigma}$ be the covariance matrix of interest. Let $\boldsymbol{S}$ be the sample covariance estimator based on the dataset. We denote by $(\lambda_i, \boldsymbol{\xi_i})$, $i = 1, 2, 3, \dots, p$ the ordered eigenvalue, eigenvector pairs of matrix $\boldsymbol{\Sigma}$. And $(\hat{\lambda}_i, \hat{\boldsymbol{\xi}}_i)$, $i = 1,2, \dots, p$ are the ordered estimated eigenvalue, eigenvector pairs of the matrix $\boldsymbol{S}$. Let $\widehat{\boldsymbol{\Sigma_f}}$ be the estimator for the systematic risk part of the overall ($\boldsymbol{\Sigma}$), $\widehat{\boldsymbol{\Sigma_u}}$ be the idiosyncratic risk part of the overall risk, and $\widehat{\boldsymbol{\Sigma_p}}$ be the estimator for $\boldsymbol{\Sigma}$ using the POET method.

Let K be the number of systematic risk factors, and $\boldsymbol{f_t} = (f_{t1}, f_{t2}, \dots, f_{tK})'$ be the risk vector at time t, with f_{ti} the expected return that the stock would gain by taking one unit of the i^{th} risk factor at time t. Let $\boldsymbol{b_i} = (b_{i1},\ b_{i2},\ \dots,\ b_{iK})$ with b_{ij} representing the units of the j^{th} risk factor undertaken by the i^{th} stock. We call $\boldsymbol{b_i}$ the factor loading of the i^{th} stock. Let $\boldsymbol{\varepsilon_t} = (\varepsilon_{1t}, \varepsilon_{2t}, \dots, \varepsilon_{pt})'$ be the return generated by the idiosyncratic risk at time t. And $\boldsymbol{R_t} = (R_{1t}, R_{2t}, \dots, R_{pt})'$ represents the return of the portfolio at time t.

Finally, let $\boldsymbol{w} = (w_1, w_2, \dots, w_p)$ be the predetermined asset allocation vector of the portfolio, with w_i denoting the proportion of the value of the i^{th} stock to the total value of the whole portfolio such that $\sum_{i=1}^{p} w_i = 1$.

2.2 Some definitions

Result 2.2.1 Given the variance matrix of $\boldsymbol{R_t}$, $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{R_t}) = (\sigma_{ij})$, and weight vector $\boldsymbol{w} = (w_1, w_2, \dots, w_p)'$, the variance of $\boldsymbol{w'R_t}$ can be computed as follow:

$$Var(\boldsymbol{w'R_t}) = \boldsymbol{w'}Cov(\boldsymbol{R_t})\boldsymbol{w} = \boldsymbol{w'}\Sigma\boldsymbol{w}.$$

Definition 2.2.1 (Frobenius Norm)

If $\boldsymbol{\Sigma}$ is an $m \times n$ matrix, the Frobenius norm of $\boldsymbol{\Sigma}$ is defined as $\|\boldsymbol{\Sigma}\|_F = (\sum_{i=1}^{m}\sum_{j=1}^{n}\sigma_{ij}^2)^{1/2}$. Thus, $\|\boldsymbol{\Sigma}\|_F^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\sigma_{ij}^2 = tr(\boldsymbol{\Sigma'\Sigma})$.

Definition 2.2.2 (Spectral decomposition)

Let $\boldsymbol{\Sigma}$ be a $p \times p$ matrix, with $(\lambda_i, \boldsymbol{\xi_i}), i = 1, 2, 3, \dots, p$ the ordered eigenvalue-eigenvector pairs of $\boldsymbol{\Sigma}$. The spectral decomposition of $\boldsymbol{\Sigma}$ is defined as: $\boldsymbol{\Sigma} = \sum_{i=1}^{p} \lambda_i\, \boldsymbol{\xi_i}\boldsymbol{\xi_i'}$.

Definition 2.2.3

If $\boldsymbol{b_i}$ is a column vector, $\|\boldsymbol{b_i}\|$ is defined as: $\|\boldsymbol{b_i}\| = (\boldsymbol{b_i}'\boldsymbol{b_i})^{1/2}$ **.**

Definition 2.2.4 Let $\{f_p\}_{p=1}^{\infty}$ be a non-negative real sequence. We have:

$f = O(p)$ means $\lim_{p\to\infty} \frac{f_p}{p} = C$ for some nonnegative constant C, and

$f = o(p)$ means $\lim_{p\to\infty} \frac{f_p}{p} = 0$.

Chapter 3

Estimation of High-dimensional Matrix

3.1 Motivation

Estimating and assessing the risk of assets is important in financial research. For a portfolio with multiple investments, the risk can be measured by the variance of the portfolio return $Var(\boldsymbol{w}'\boldsymbol{R_t})$, which is further decided by the asset allocation($\boldsymbol{w}$) and covariance matrix of the return of investments $\boldsymbol{\Sigma} = Var(\boldsymbol{R_t})$. Given the covariance matrix $\boldsymbol{\Sigma}$, and the asset allocation vector $\boldsymbol{w}$, which we assume is predetermined, we can estimate the risk of portfolio as $\boldsymbol{w}'\Sigma\boldsymbol{w}$. $\boldsymbol{\Sigma}$ is not easy to estimate, especially when high-dimensional datasets are involved. $\boldsymbol{\Sigma}$ is the key to estimating the risk of investment portfolios.

In practice, investors, especially institutional investors, need to construct investment portfolios consisting of a large number of stocks for various purposes. Thus, a high dimensional covariance matrix needs to be estimated.

Based on Fan, Han, and Liu[2], when the dimension p of the covariance matrix grows, the number of parameters to be estimated becomes larger, which causes challenge to estimate the covariance matrix. For example, suppose the investor constructs a portfolio with 200 stocks from a stock market. There are $1 + 2 + 3 + \cdots + 200 = \frac{200\times201}{2} = 20100$ parameters to estimate in order to construct the covariance matrix.

The trading data of daily traded stocks would only give us a sample size of around 252 per year. Also, in order to reflect the current economic and financial market conditions, data from several years ago is not appropriate to estimate the parameters about current risk. That causes the sample size to be limited, which makes it more difficult to assess the risk associated with large portfolios.

3.2 Sample covariance matrix

The sample covariance matrix $\boldsymbol{S}$ is commonly used to estimate the covariance matrix $\boldsymbol{\Sigma}$. Let $\boldsymbol{R_t}, t = 1,2,3, \dots, T$ be the observed portfolio return over T, and $\overline{\boldsymbol{R}} = \frac{1}{\mathrm{T}}\sum_{t=1}^{T} \boldsymbol{R_t}$ be the mean return over T. The sample covariance estimator is obtained as follow:

$$\boldsymbol{S} = \frac{1}{\mathrm{T}-1}(\boldsymbol{R_t} - \overline{\boldsymbol{R}})(\boldsymbol{R_t} - \overline{\boldsymbol{R}})',$$

This sample covariance estimator is convenient to compute, and it performs well when the dimension of the covariance matrix p is small. According to the research of Geman[6], when the dimension of the covariance matrix p increases, the sample covariance matrix estimator can be biased. More specifically, when the ratio p/T becomes large, this sample covariance estimator is usually biased.

Based on simulation, we observe that when p/T approaches 1, the eigenvalues of the sample covariance estimator $\boldsymbol{S}$ deviate from the true eigenvalues of the covariance matrix. This becomes more problematic when the dimension of the covariance matrices p grows.

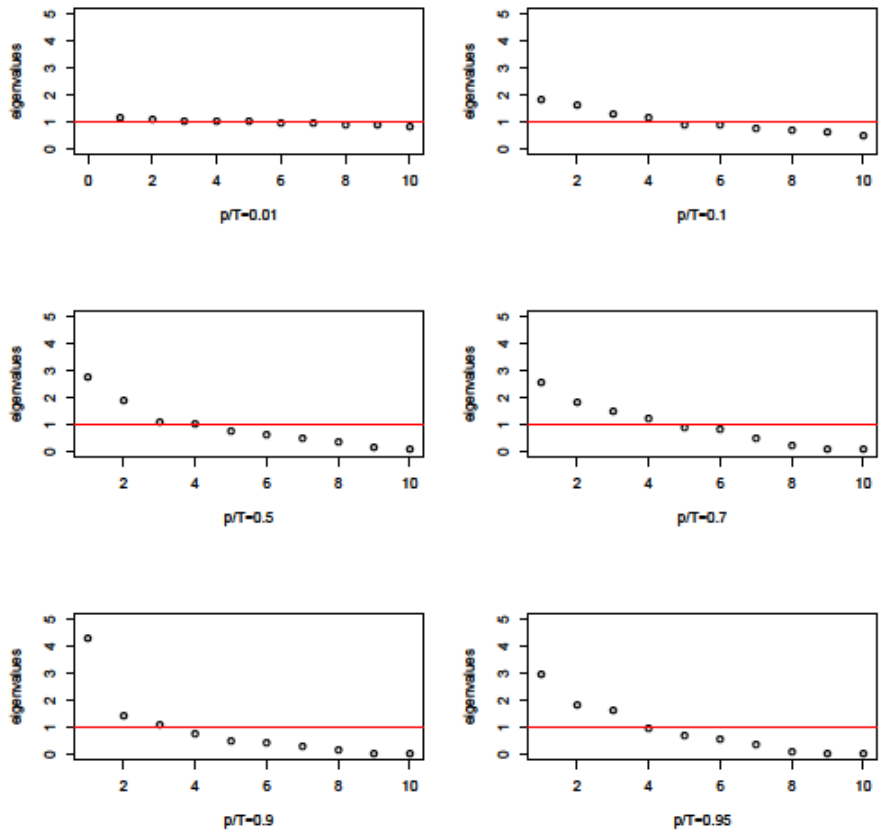

Figure 3.1 The sorted eigenvalues when p=10 and p/T becomes larger

We denote the multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ by $MN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. In the above simulation, we simulate data samples following $MN(\mathbf{0}, \boldsymbol{I_{10}})$ of size T. Letting the sample size T equal to $10, 11, 15, 20, 100, 1000$, we observe that when T is large, i.e. when the ratio p/T is small, the eigenvalues of the sample covariance estimator are close to the theoretical eigenvalues. But when the ratio p/T becomes larger, the eigenvalues of the sample covariance estimator deviate from the true value.

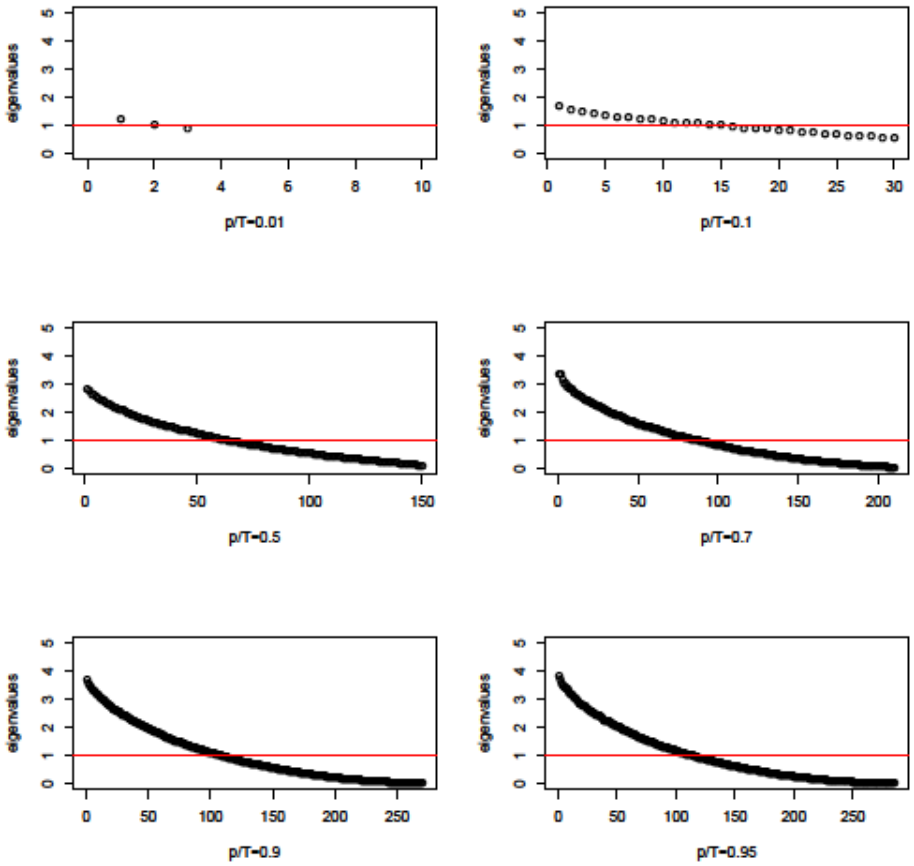

Figure 3.2 Sorted eigenvalues when T=300 and p becomes larger

In the above simulation, we simulate data samples following $MN(\mathbf{0}, \boldsymbol{I_p})$ of size 300. Letting the dimension of data samples take values $3, 30, 150, 210, 270, 285$ respectively, we observe that the eigenvalues of the sample covariance estimator are close to the theoretical eigenvalue when $p = 3$. But when p grows larger, even when p/T is still small, for example, when $p/T = 0.1$, the eigenvalues of the sample covariance estimator still deviate from the true value.

Fan and Liu[4] and Yin, Bai and Krishnaiah[5] illustrate inconsistency of the sample covariance matrices in high dimensional conditions. They explaine this phenomenon by random matrix theory:

When $p/n \to \gamma$ and $\gamma \in (0,1)$, The largest eigenvalue of $\boldsymbol{S}$ converges to $(1+\sqrt{\gamma})^2$ almost surely.

$$\mathrm{P}(\lim_{n\to\infty} \lambda_{max}(\boldsymbol{S}) = (1+\sqrt{\gamma})^2) = 1,$$

where $\lambda_{max}(\boldsymbol{S})$ is the largest eigenvalue of sample covariance matrix. When the dimension of a matrix becomes large, the largest eigenvalue of the sample covariance will not converge to the true value of largest eigenvalue of $\boldsymbol{\Sigma}$, which makes the estimation based on the sample covariance biased.

According to the research of Fan, Liao and Mincheva[4], aside from the inconsistent estimation of covariance, the sample covariance estimator is often singular when the dimension of covariance matrix is large, which makes the further analysis challenging.

3.3 Sparse matrices

To deal with the inconsistent sample covariance estimator for high-dimensional matrices, it is often assumed that the covariance matrices are sparse. Under this assumption, most of the off-diagonal entries in the matrix are 0. According to the research of Fan, Liao and Mincheva[4], this assumption effectively lowers the number of parameters to estimate, and, to some extent, guarantees the consistency of the estimator of the covariance matrix. This sparsity assumption is reasonable in some circumstances. For example, in

biostatistics research, when it comes to genes, it is reasonable and convenient to assume most genes are non-correlated.

In our case, the object is to obtain an estimator $\widehat{\boldsymbol{\Sigma}}$ that best reflects the covariance structure of the stocks based on the sample covariance matrix $\boldsymbol{S}$. Without the sparsity assumption, the covariance matrix is estimated based on the least squares method as below:

$$\widehat{\boldsymbol{\Sigma}} = \arg\min_{(\sigma_{ij})} \|\boldsymbol{S} - \boldsymbol{\Sigma}\|_F^2$$

$$= \arg\min_{(\sigma_{ij})} tr(\boldsymbol{S} - \boldsymbol{\Sigma})'(\boldsymbol{S} - \boldsymbol{\Sigma})$$

$$= \arg\min_{(\sigma_{ij})} \sum_{i=1}^{p} \sum_{j=1}^{p} (s_{ij} - \sigma_{ij})^2.$$

In order to obtain a sparse estimator of the covariance matrix, a thresholding procedure is often used. The thresholding procedure requires a penalty function to be introduced into above least squares function.

The penalized least squares function is as follows:

$$\widehat{\boldsymbol{\Sigma}} = \arg\min_{(\sigma_{ij})} \{\frac{1}{2} \|\boldsymbol{S} - \boldsymbol{\Sigma}\|_F^2 + \sum_{i \neq j} P_\lambda(\sigma_{ij})\}$$

$$= \arg\min_{(\sigma_{ij})} \{\sum_{i=1}^{p} \sum_{j=1}^{p} \frac{1}{2} (s_{ij} - \sigma_{ij})^2 + \sum_{i \neq j} P_\lambda(\sigma_{ij})\}$$

$$= \arg\min_{(\sigma_{ij})} \sum_{i=1}^{p} \sum_{j=1}^{p} [\frac{1}{2} (s_{ij} - \sigma_{ij})^2 + I(i \neq j) P_\lambda(\sigma_{ij})],$$

where $I(i \neq j) = \begin{cases} 0, & when\ i = j \\ 1, & when\ i \neq j \end{cases}$. Based on the penalized least squares function, for each element in the minimization, it only depends on s_{ij} and σ_{ij}. We can simplify the above minimization into the following formula: [1][3]

$$\hat{\sigma}_{ij} = \min_{\sigma_{ij}}[\frac{1}{2}\left(s_{ij} - \sigma_{ij}\right)^2 + P_\lambda(\sigma_{ij})].$$

With thresholding procedures, the off-diagonal entries in the sample covariance matrix would be transformed into 0 if the values are "small enough." The procedure works as follows: Compute the sample covariance matrix first; introduce a penalty threshold operator; perform thresholding on the off-diagonal entries in the sample covariance matrix. This procedure provides a sparse estimator for $\boldsymbol{\Sigma}$. This method can improve the estimating property of estimators based on the sample covariance matrices.

3.4 Soft-thresholding

There are several commonly used penalty functions, such as complexity penalty, hard-thresholding penalty, and soft-thresholding[3]. In this project, we choose to use the soft-thresholding as below:

$$P_\lambda(u) = \lambda|u|.$$

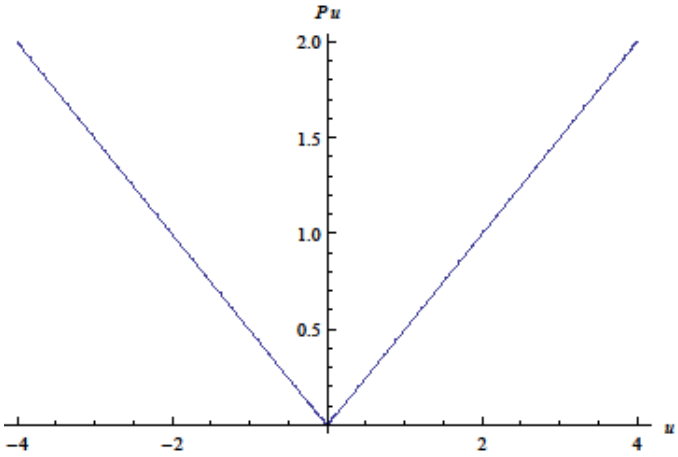

Figure 3.3 The value of Penalty function

After introducing the soft-thresholding penalty function, the least squares function is as follows:

$$\widehat{\boldsymbol{\Sigma}} = \arg\min_{(\sigma_{ij})}\{\frac{1}{2}\|\boldsymbol{S}-\boldsymbol{\Sigma}\|_F^2 + \sum_{i\neq j}\lambda|\sigma_{ij}|\},$$

which is equivalent to:

$$\hat{\sigma}_{ij} = \arg\min_{\sigma_{ij}}[\frac{1}{2}(s_{ij}-\sigma_{ij})^2 + \lambda|\sigma_{ij}|],$$

for each $1 \le i \le p, 1 \le j \le p, and\ i \neq j$.

Let $f(u) = \frac{1}{2}(s_{ij}-u)^2 + \lambda|u|$. Then the problem above would be simplified to find the root to the first derivative of $f(u)$.

$$\frac{df(u)}{du} = -(s_{ij}-u) + sgn(u)\lambda,$$

where, $sgn(u) = \begin{cases} 1 & when\ s_{ij} > 0 \\ 0 & when\ s_{ij} = 0 \\ -1 & when\ s_{ij} < 0 \end{cases}$.

Let $\frac{df(\hat{\sigma}_{ij})}{d\hat{\sigma}_{ij}} = 0$. And since $\hat{\sigma}_{ij}$ and s_{ij} should be of the same sign, we have

$$\hat{\sigma}_{ij} = s_{ij} - sgn(s_{ij})\lambda.$$

Also the thresholding method is not expected to change the correlation relation between the return of two stocks, s_{ij} and $\hat{\sigma}_{ij}$ should be of the same sign. If the absolute value of s_{ij} is smaller than λ, we treat $\hat{\sigma}_{ij}$ as 0.

$$\hat{\sigma}_{ij} = sgn(s_{ij})(|s_{ij}| - \lambda)_+,$$

that is, $\hat{\sigma}_{ij} = \begin{cases} 0 & \text{when } |s_{ij}| \le \lambda \\ sgn(s_{ij})(|s_{ij}| - \lambda) & \text{when } |s_{ij}| > \lambda \end{cases}$.

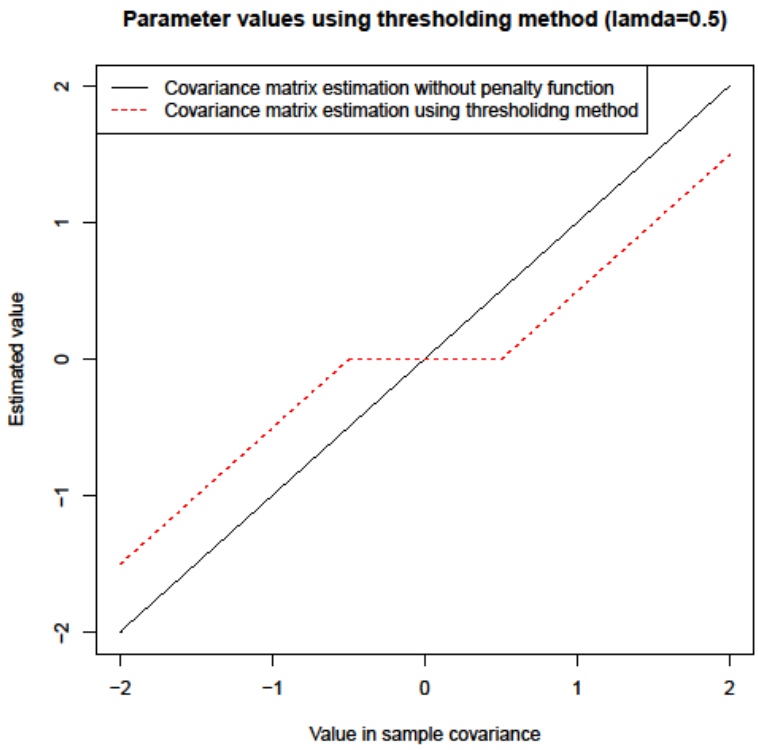

Figure 3.4 $\hat{\sigma}_{ij}$ after thresholding procedure

3.5 Limitation of Sparse Estimators

A sparse estimator may improve consistency of the estimation in many circumstances. However, the sparsity assumption requires that the returns of most investments be uncorrelated, which would not be satisfied in real economic entities. In light of systematic risk factors such as interest rate, inflation, recessions, exchange rates, etc., the assumption of sparsity of covariance matrices is inappropriate in financial contexts.

Also, since the sparsity estimator is solely based on the current dataset, the estimator could reflect the trend and characteristics of the returns of stocks in the distant past rather than the varying feature of risk.

Another problem with using the sparsity estimator is that the estimator is completely data-driven. Thus the estimator has no way to capture changes in return volatility. Volatility of the stocks' return is usually serially correlated and is often influenced by some common factors, which may not be reflected by the sparsity estimator.

Fan, Liao and Mincheva[4] introduced the thresholding Principal Orthogonal complEment (POET) estimator to address the limitation of sparsity estimators.

Chapter 4

POET Method

4.1 Factor model

Factor models provide a solution to the problem that sparsity assumption is not appropriate in financial context. First, factor models reflect risk decompositions for more meaningful analysis. That is of great importance in making investment decisions. Second, factor models can reveal correlation relationships among stocks' returns in the future. Hence, they capture the time varying feature of volatility.

In factor models, we assume the stocks' returns are affected by several common factors:

$$\boldsymbol{R_t} = \boldsymbol{B f_t} + \boldsymbol{u_t},$$

where $\boldsymbol{R_t}$ is a $p \times 1$ vector, denoting the return of p stocks at time point t. Let $\boldsymbol{f_t}$ be a $K \times 1$ vector, denoting the K independent systematic risk factors, with f_{ti} being the return that a stock gains by taking one unit of the i^{th} risk factor at time t. $\boldsymbol{u_t} = (u_{1t}, u_{2t}, \dots, u_{pt})'$ is a $p \times 1$ vector**,** which describes the non-systematic risk, and $Cov(\boldsymbol{u_t}) = \boldsymbol{\Sigma_u}$ is a diagonal matrix.

$\boldsymbol{b_i} = (b_{i1},\ b_{i2},\ \dots,\ b_{iK})$, where b_{ij} represents the units of the j^{th} risk factor undertaken by the ith stock. We call $\boldsymbol{b_i}$ the factor loading of the i^{th} stock. Let $\boldsymbol{B} = (\boldsymbol{b_1}, \boldsymbol{b_2}, \dots, \boldsymbol{b_p})'$ be a $p \times K$ vector, which denotes the risk loading of the portfolio.

We assume that $Cov(\boldsymbol{B}\boldsymbol{f}_t, \boldsymbol{u}_t) = 0$, which means that the systematic risk and idiosyncratic risk are uncorrelated. To further simplify the problem, without losing generality, we also assume that the systematic risk factors are identically distributed and uncorrelated, i.e. $Cov(\boldsymbol{f}_t) = \boldsymbol{I}_K$, so that the factor loading matrix $\boldsymbol{B}$ has orthogonal columns, i.e. $\boldsymbol{b}_i \boldsymbol{b}_j' = 0 \; when \; i \neq j$.

The issue of interest is to estimate the $p \times p$ matrix covariance matrix $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma} = Cov(\boldsymbol{R}_t) = Cov(\boldsymbol{B}\boldsymbol{f}_t + \boldsymbol{u}_t) = \boldsymbol{B}cov(\boldsymbol{f}_t)\boldsymbol{B}' + \boldsymbol{\Sigma}_{\boldsymbol{u}} = \boldsymbol{B}\boldsymbol{B}' + \boldsymbol{\Sigma}_{\boldsymbol{u}}.$$

We assume $\boldsymbol{B}\boldsymbol{B}'$ has spiked eigenvalues, i.e. $\lambda_{i+1} - \lambda_i = O(p)$ for $i \leq K$, such that the remaining entries of the covariance matrix are close to 0 after the systematic risk factor parts are taken out. The rest part, $\boldsymbol{\Sigma}_{\boldsymbol{u}} = \boldsymbol{\Sigma} - \boldsymbol{B}cov(\boldsymbol{f}_t)\boldsymbol{B}'$, is a sparse matrix, which represents the non-systematic risk of stocks. This sparsity assumption is reasonable because given an idiosyncratic risk factor, it only affects one or a few securities in the financial market.

Also, according to the research by Fan, Liao and Mincheva[4], the systematic risk part of the covariance matrix is approximately the same with the first K main principals of the covariance matrix $\boldsymbol{\Sigma}$. This result makes it easier to obtain the estimation for $ov(\boldsymbol{f}_t)\boldsymbol{B}'$.

4.2 Deriving the Estimation for Systematic Risk

Based on the factor model, we already know that the risk covariance matrix can be decomposed into two parts: the systematic risk part $\boldsymbol{B}\boldsymbol{B}'$, and the idiosyncratic risk

part $\boldsymbol{\Sigma_u}$. Given the number of systematic risk factors K, we can derive the estimation for the systematic risk part.

Lemma 1: Let $\boldsymbol{B} = (\boldsymbol{b_1}, \boldsymbol{b_2}, \dots, \boldsymbol{b_K})$ be a $p \times K,\ p > K$ matrix, where the K columns of matrix $\boldsymbol{B}$ are orthogonal, then,

(i) $\frac{\boldsymbol{b_i}}{\|\boldsymbol{b_i}\|}$, where $\|\boldsymbol{b_i}\| = \boldsymbol{b_i}'\boldsymbol{b_i}^{1/2}$ are eigenvectors of the $p \times p$ matrix $\boldsymbol{BB'}$ with the corresponding eigenvalue $\|\boldsymbol{b_i}\|^2$, and the other $p - K$ eigenvalues are 0.

(ii) The K non-vanishing eigenvalues of the $K \times K$ matrix $\boldsymbol{B'B}$ are the same as the K eigenvalues of $p \times p$ matrix $\boldsymbol{BB'}$.

Proof:

$$\boldsymbol{BB'}\frac{\boldsymbol{b_i}}{\|\boldsymbol{b_i}\|} = \boldsymbol{B}\frac{\begin{pmatrix}\boldsymbol{b_1'}\\ \dots \\ \boldsymbol{b_K'}\end{pmatrix}\boldsymbol{b_i}}{\|\boldsymbol{b_i}\|} = \frac{1}{\|\boldsymbol{b_i}\|}(\boldsymbol{b_1}, \boldsymbol{b_2}, \dots, \boldsymbol{b_K})\begin{pmatrix}0\\ \vdots \\ 0\\ \|\boldsymbol{b_i}\|^2\\ 0\\ \vdots \\ 0\end{pmatrix} = \|\boldsymbol{b_i}\|\boldsymbol{b_i} = \|\boldsymbol{b_i}\|^2\frac{\boldsymbol{b_i}}{\|\boldsymbol{b_i}\|}\ .$$

Thus, $(\|\boldsymbol{b_i}\|^2, \frac{\boldsymbol{b_i}}{\|\boldsymbol{b_i}\|})$ is eigenvalue eigenvector pair of matrix of $\boldsymbol{BB'}$. Also, since $Rank(\boldsymbol{B}) = K$, we have $Rank(\boldsymbol{BB'}) = K$, and the remaining $p - K$ eigenvalues are 0.

$$\boldsymbol{B'B} = \begin{pmatrix}\boldsymbol{b_1'}\\ \dots \\ \boldsymbol{b_K'}\end{pmatrix}(\boldsymbol{b_1}, \boldsymbol{b_2}, \dots, \boldsymbol{b_K}) = \begin{bmatrix}\|\boldsymbol{b_1}\|^2 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \|\boldsymbol{b_K}\|^2\end{bmatrix}.$$

Since $\boldsymbol{B'B}$ is diagonal matrix, then we have the eigenvalues are the elements on the diagonal position.

Lemma 2 (Weyl's eigenvalue theorem)[9]: Let $\boldsymbol{\Sigma}$ and $\widehat{\boldsymbol{\Sigma}}$ be two $p \times p$ matrices, $\{\lambda_i, \boldsymbol{\xi_i}\}$, $\{\hat{\lambda}_i, \hat{\boldsymbol{\xi}}_i\}$ be the ordered eigenvalue, eigenvector pairs of $\boldsymbol{\Sigma}$ and $\widehat{\boldsymbol{\Sigma}}$ respectively. We have,

$$\left|\lambda_i - \hat{\lambda}_i\right| \leq \left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_F \text{ for any } i = 1, 2, \dots, p.$$

Lemma 3 ($\sin(\theta)$ theorem of Davis and Kalan) [8]:

Let $\boldsymbol{\Sigma}$ and $\widehat{\boldsymbol{\Sigma}}$ be two $p \times p$ matrices, $\{\lambda_i, \boldsymbol{\xi_i}\}$, $\{\hat{\lambda}_i, \hat{\boldsymbol{\xi}}_i\}$ be the ordered eigenvalue, eigenvector pairs of $\boldsymbol{\Sigma}$ and $\widehat{\boldsymbol{\Sigma}}$ respectively. Then,

$$\left\|\hat{\boldsymbol{\xi}}_i - \boldsymbol{\xi_i}\right\| \leq \frac{\sqrt{2}\left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_F}{\min(|\hat{\lambda}_{i-1} - \lambda_i|, |\lambda_i - \hat{\lambda}_{i+1}|)} \text{ for } i = 2, 3, \dots, p-1.$$

Theorem 1 Let $\{\lambda_i, \boldsymbol{\xi_i}\}, i = 1,2, \dots, p$ be the ordered eigenvalue, eigenvector pairs of $\boldsymbol{\Sigma}$, and K is the number of factors. We have that the result of factor analysis is asymptotically the same as summation of the first K main principals of $\boldsymbol{\Sigma}$.

The spectral decomposition of $\boldsymbol{\Sigma}$ is $\boldsymbol{\Sigma} = \sum_{i=1}^{p} \lambda_i \boldsymbol{\xi_i} \boldsymbol{\xi_i}'$. As the number of factors is known to be K, the result of the main principal analysis is as follow:

$$\boldsymbol{\Sigma} = \sum_{i=1}^{K} \lambda_i \boldsymbol{\xi_i} \boldsymbol{\xi_i}' + \sum_{i=K+1}^{p} \lambda_i \boldsymbol{\xi_i} \boldsymbol{\xi_i}'.$$

On the other hand, based on the result of factor analysis, $\boldsymbol{\Sigma} = \boldsymbol{BB}' + \boldsymbol{\Sigma_u}$, where $\boldsymbol{BB}'$ is the part of covariance that can be explained by systematic risk factors.

According to Theorem 1, $\sum_{i=1}^{K} \lambda_i \boldsymbol{\xi_i} \boldsymbol{\xi_i}'$ and $\boldsymbol{BB}'$ are asymptotically the same.

Proof:

By lemma 1, $\{\|\boldsymbol{b_i}\|^2, \frac{\boldsymbol{b_i}}{\|\boldsymbol{b_i}\|}\}$, $i = 1,2, \dots, K$ are the first K ordered eigenvalue, eigenvector pairs of $\boldsymbol{BB}'$. And the other eigenvalues of $\boldsymbol{BB}'$ are 0.

For $j \leq K$, $\left|\lambda_j - \left\|\boldsymbol{b_j}\right\|^2\right| \leq \|\boldsymbol{\Sigma} - \boldsymbol{BB}'\|_F = \|\boldsymbol{\Sigma_u}\|_F$.

For $j > K$, $\left|\lambda_j - 0\right| \leq \|\boldsymbol{\Sigma} - \boldsymbol{BB'}\|_F = \|\boldsymbol{\Sigma_u}\|_F$.

Since we already assume that $\boldsymbol{\Sigma_u}$ is sparse, i.e. $\lim_{p\to\infty} \frac{\|\boldsymbol{\Sigma_u}\|_F}{p} = 0$. When p goes to infinity, the eigenvalues of $\boldsymbol{\Sigma}$ and $\boldsymbol{BB'}$ are asymptotically the same.

Also, based on the $\sin(\theta)$ theorem of Davis and Kalan, we have when $i \leq K$,

$$\left\|\boldsymbol{\xi_i} - \frac{\boldsymbol{b_i}}{\|\boldsymbol{b_i}\|}\right\| \leq \frac{\sqrt{2}\|\boldsymbol{\Sigma - BB'}\|_F}{\min(|\|\boldsymbol{b_{i-1}}\|^2 - \lambda_i|, |\lambda_i - \|\boldsymbol{b_{i+1}}\|^2|)} = \frac{\sqrt{2}\|\boldsymbol{\Sigma_u}\|_F}{\min(|\|\boldsymbol{b_{i-1}}\|^2 - \lambda_i|, |\lambda_i - \|\boldsymbol{b_{i+1}}\|^2|)}.$$

We note that,

$$|\|\boldsymbol{b_{i-1}}\|^2 - \lambda_i| - |\lambda_i - \lambda_{i-1}| \leq |\|\boldsymbol{b_{i-1}}\|^2 - \lambda_i + \lambda_i - \lambda_{i-1}|$$

$$= |\|\boldsymbol{b_{i-1}}\|^2 - \lambda_{i-1}| \leq \|\boldsymbol{\Sigma_u}\|_F = o(p),$$

and $|\|\boldsymbol{b_{i-1}}\|^2 - \lambda_i|$ must be $O(p)$ since $|\lambda_i - \lambda_{i-1}| = O(p)$.

Similarly,

$$|\|\boldsymbol{b_{i+1}}\|^2 - \lambda_i| - |\lambda_i - \lambda_{i+1}| \leq |\|\boldsymbol{b_{i+1}}\|^2 - \lambda_i + \lambda_i - \lambda_{i+1}|$$

$$= |\|\boldsymbol{b_{i+1}}\|^2 - \lambda_{i+1}| \leq \|\boldsymbol{\Sigma_u}\|_F = o(p),$$

and $|\|\boldsymbol{b_{i+1}}\|^2 - \lambda_i|$ must be $O(p)$ since $|\lambda_i - \lambda_{i+1}| = O(p)$.

Therefore, $\min(|\|\boldsymbol{b_{i-1}}\|^2 - \lambda_i|, |\lambda_i - \|\boldsymbol{b_{i+1}}\|^2|) = O(p)$, and

$$\frac{\sqrt{2}\|\boldsymbol{\Sigma_u}\|_F}{\min(|\|\boldsymbol{b_{i-1}}\|^2 - \lambda_i|, |\lambda_i - \|\boldsymbol{b_{i+1}}\|^2|)} = \frac{\sqrt{2}\|\boldsymbol{\Sigma_u}\|_F}{O(p)} \xrightarrow[p\to\infty]{} 0,$$

which means the eigenvectors of $\boldsymbol{\Sigma}$ and $\boldsymbol{BB'}$ are asymptotically the same.

Overall, we can use the result of the principal analysis to estimate the systematic risk part in the covariance matrix.

4.3 POET estimator

As stated above, under high-dimensional models, the normalized columns of the factor loading matrix $\boldsymbol{B}$ are approximately the same as the K largest eigenvetors. This means that the covariance matrix of p stocks can be decomposed as:

$$\boldsymbol{\Sigma} = \sum_{i=1}^{p} \lambda_i \boldsymbol{\xi_i} \boldsymbol{\xi_i}' = \sum_{i=1}^{K} \lambda_i \boldsymbol{\xi_i} \boldsymbol{\xi_i}' + \sum_{i=K+1}^{p} \lambda_i \boldsymbol{\xi_i} \boldsymbol{\xi_i}' = \boldsymbol{BB}' + \boldsymbol{\Sigma_u},$$

$$\boldsymbol{S} = \sum_{i=1}^{p} \hat{\lambda}_i \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i{}' = \sum_{i=1}^{K} \hat{\lambda}_i \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i{}'{}' + \sum_{i=K+1}^{p} \hat{\lambda}_i \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i{}' = \widehat{\boldsymbol{BB}}' + \sum_{i=K+1}^{p} \hat{\lambda}_i \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i{}',$$

where $\hat{\lambda}_i$ denotes the i^{th} largest eigenvalue of sample covariance matrix $\boldsymbol{S}$ and $\hat{\boldsymbol{\xi}}_i$ denotes the eigenvector corresponds to it. As proved above, we can treat $\sum_{i=1}^{K} \hat{\lambda}_i \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i{}'$ as the estimator for $\boldsymbol{BB}'$.

As we assume, $\boldsymbol{\Sigma_u}$ is a sparse matrix. We can introduce a thresholding procedure to gain a sparse estimator $\widehat{\boldsymbol{\Sigma_u}}$ for the idiosyncratic risk part of the covariance matrix based on the remaining part of the spectral decomposition of $\boldsymbol{S}$.

By adding the estimation for the two parts of the covariance matrix, the POET estimator is:

$$\widehat{\Sigma_P} = \sum_{i=1}^{K} \hat{\lambda}_i \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i{}' + \widehat{\boldsymbol{\Sigma_u}}.$$

Finally, given the dataset of stocks' returns, the procedure to obtain the POET estimator for a covariance matrix can be summarized as below:

(1) Calculate the sample covariance matrix $\boldsymbol{S}$, and find the eigenvalues and corresponding eigenvectors for $\boldsymbol{S}$.

(2) Based on the eigenvalues combined with the meaning of risk factors, decide the number of systematic risk factors K.

(3) Calculate the POET estimator: $\widehat{\boldsymbol{\Sigma_P}} = \sum_{i=1}^{K} \hat{\lambda}_i \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i{}' + \widehat{\boldsymbol{\Sigma_u}}$.

Chapter 5

Risk estimation

Based on the covariance estimator above, the estimation of portfolio risk can be obtained. Let the weights of the stocks in a portfolio be $\boldsymbol{w}$, which is predetermined and known.

5.1 Sample covariance based risk estimator

Based on the sample covariance estimator, the portfolio risk can be estimated as: $R(\boldsymbol{w}) = \sqrt{\boldsymbol{w}'\boldsymbol{S}\boldsymbol{w}}$.

5.2 POET based risk estimator

Under the POET method, the covariance matrix of stocks' returns is decomposed into two parts: The factor dominated part that corresponds to the K largest eigenvalues and reflects the systematic risk part; and the sparse part that is obtained through performing thresholding to the remaining part of the spectral decomposition of the sample. Under this method, the risk estimator is:

$$R(\boldsymbol{w}) = \sqrt{\boldsymbol{w}'\widehat{\boldsymbol{\Sigma}_{\boldsymbol{p}}}\boldsymbol{w}}, \text{ where } \widehat{\boldsymbol{\Sigma}_{\boldsymbol{p}}} = \sum_{i=1}^{K} \hat{\lambda}_i \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i' + \widehat{\boldsymbol{\Sigma}_{\boldsymbol{u}}}.$$

Chapter 6

Validation using Monte Carlo Simulation

Based on research of Fan, Liao and Shi[7], we can use Monte Carlo simulation to evaluate the performance of risk estimators.

6.1 Generating data

(1) Generate $b_{ik}, k = 1,2,3,\ i = 1, 2, 3, \dots p$ independently from $b_{i1} \sim N(0.5, 1)$, $b_{i2} \sim N(1, 1)$, $b_{i3} \sim N(0.8, 1)$. Construct $\boldsymbol{B} = (b_{ik})$, where $\boldsymbol{B}$ is a $p \times 3$ matrix. p is the number of stocks in the portfolio. In our simulation, p ranges from 50 to 200.

(2) Generate a $p \times 1$ vector $\boldsymbol{u_t}$ independently from $N(0, I_p)$, $t = 1, 2, 3, \dots 200$.

(3) Generate a 3×1 vector $\boldsymbol{f_t}$ from a vector auto regression model $VAR(1)$ $\boldsymbol{f_t} = \boldsymbol{\mu} + \boldsymbol{\phi} \boldsymbol{f_{t-1}} + \boldsymbol{\varepsilon_t}$, where $\boldsymbol{\phi}$ is a 3×3 matrix, and $\boldsymbol{\varepsilon_t} \sim N(0, \boldsymbol{I_3})$.

Based on the research of Fan, Liao and Shi[7], we use the following parameters to generate $\boldsymbol{f_t}$:

$\boldsymbol{\mu}$	$\boldsymbol{\phi}$			$\boldsymbol{\Sigma_f}$		
0.0260	-0.1006	0.2803	-0.0365	3.2351	0.1783	0.7783
0.0211	-0.0191	-0.0944	0.0186	0.1783	0.5069	0.0102
-0.0043	0.0116	-0.0272	0.0272	0.7783	0.0102	0.6586

Table 6.1 The parameters used to generate risk factor values

(4) $\boldsymbol{R_t} = \boldsymbol{B f_t} + \boldsymbol{u_t}$, for $t = 1, 2, 3, \dots 200$.

(5) Calculate the estimator for the covariance matrix with the sample covariance estimator and the POET estimator.

(6) Generate weights for allocation of stocks in the portfolio $\boldsymbol{w}$. Under the constraints $\sum_{i=1}^{p} w_i = 1$, $\sum_{i=1}^{p} |w_i| = c$, where c is the total risk exposure of certain portfolio. Vector $\boldsymbol{w}$ can be generated as follow:

i. Generate the number of the long position in the portfolio, $L \sim BIN(p, \frac{c+1}{2c})$;

ii. Generate $\eta_i, i = 1, 2, \dots p$ from an exponential distribution, the proportion of assets allocated to each long position would be $w_i = (c+1)\eta_i / (2\sum_{i=1}^{L} \eta_i)$ for $i = 1, 2, \dots L$. The weights of assets allocated to each short position would be $w_i = (1-c)\eta_i / (2\sum_{i=L+1}^{p} \eta_i)$, $i = L+1, L+2, \dots, p$;

iii. $\boldsymbol{w}$ is a $p \times 1$ vector, where the entries are random permutations of the weights generated with the above method, which means we will repeat constructing the portfolio with the same total exposure 50 times.

6.2 Covariance estimation

(1) Compute the true risk with $R(\boldsymbol{w}) = \sqrt{\boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w}}$, where $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{\Sigma}_{\boldsymbol{f}}\boldsymbol{B}' + \boldsymbol{\Sigma}_{\boldsymbol{u}}$, and $\boldsymbol{\Sigma}_{\boldsymbol{f}}$ is decided by $VAR(1)$ model $\boldsymbol{\Sigma}_{\boldsymbol{f}} = \boldsymbol{\phi}\boldsymbol{\Sigma}_{\boldsymbol{f}}\boldsymbol{\phi}' + \boldsymbol{I}_{\boldsymbol{3}}$.

(2) Compute risk estimation base on the sample covariance estimator: $R(\boldsymbol{w}) = \sqrt{\boldsymbol{w}'\boldsymbol{S}\boldsymbol{w}}$.

(3) Compute risk estimation base on the POET method: $R(\boldsymbol{w}) = \sqrt{\boldsymbol{w}'\widehat{\boldsymbol{\Sigma}_{\boldsymbol{P}}}\boldsymbol{w}}$.

6.3 Risk estimation

We evaluate the performance of risk estimators according to their deviance from the true risk, defined as below:

$$\Delta_S = |\boldsymbol{w}'(\boldsymbol{\Sigma} - \boldsymbol{S})\boldsymbol{w}|,$$

$$\Delta_P = |\boldsymbol{w}'(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}_P})\boldsymbol{w}|.$$

With the simulated data, estimators using the POET estimators perform better as the dimension of the covariance matrices grows.

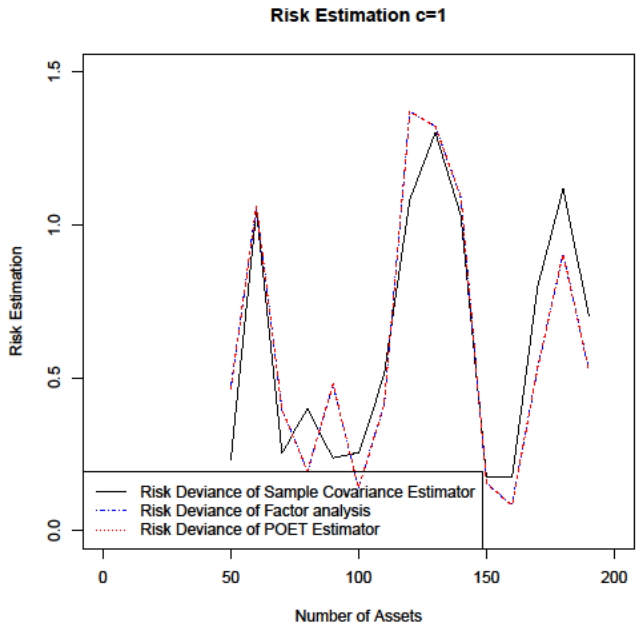

Figure 6.1 Risk Estimation Comparison when c=1

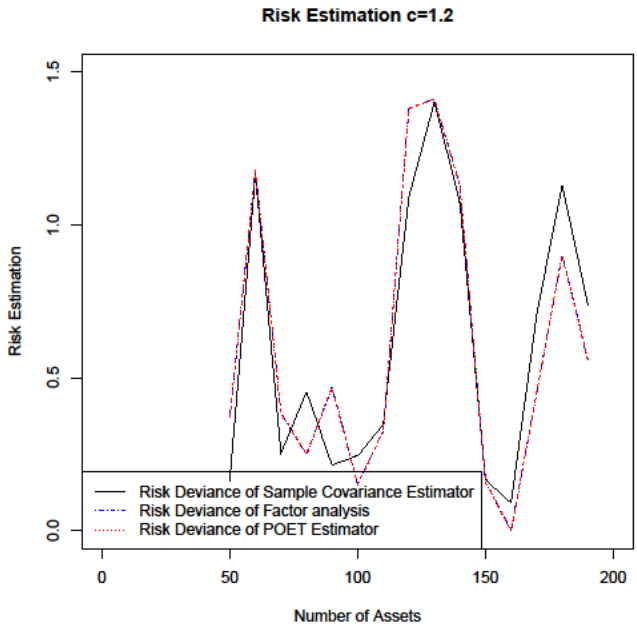

Figure 6.2 Risk Estimation Comparison when c=1.2

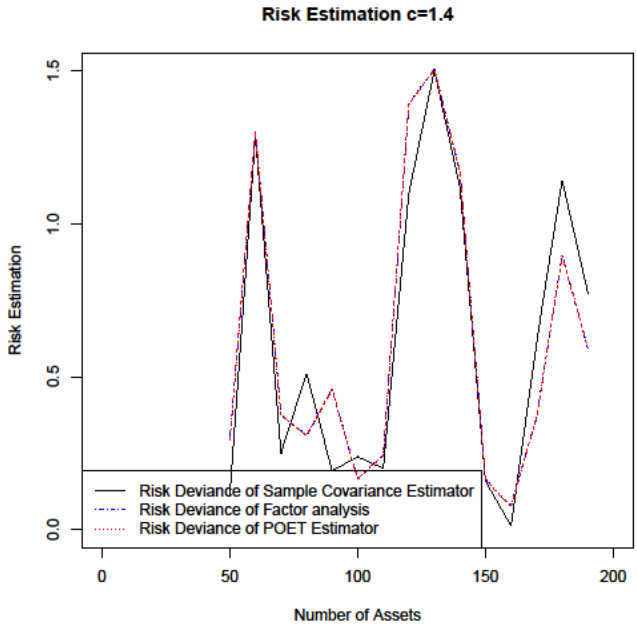

Figure 6.3 Risk Estimation Comparison when c=1.4

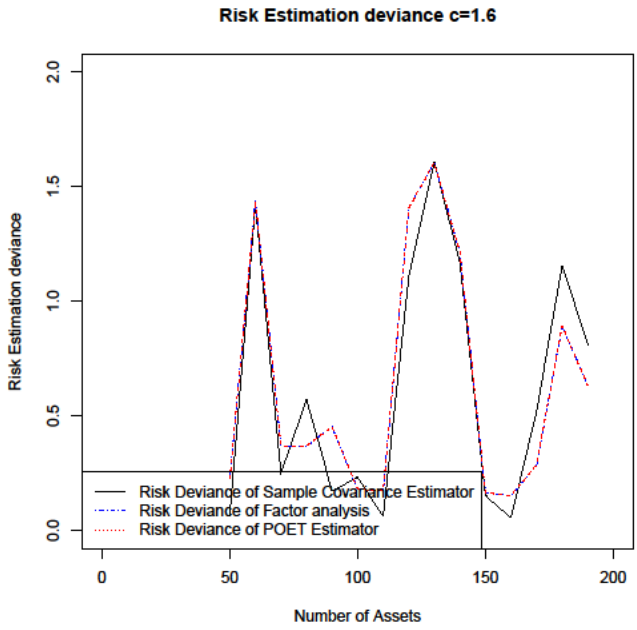

Figure 6.4 Risk Estimation Comparison when c=1.6

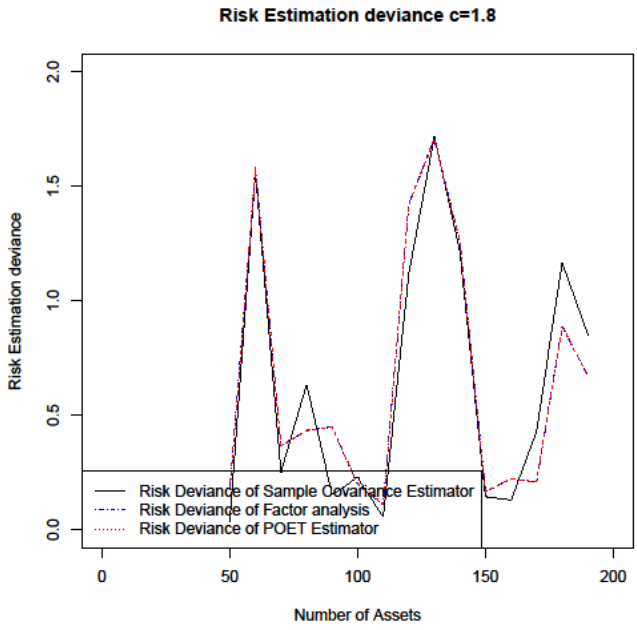

Figure 6.5 Risk Estimation Comparison when c=1.8

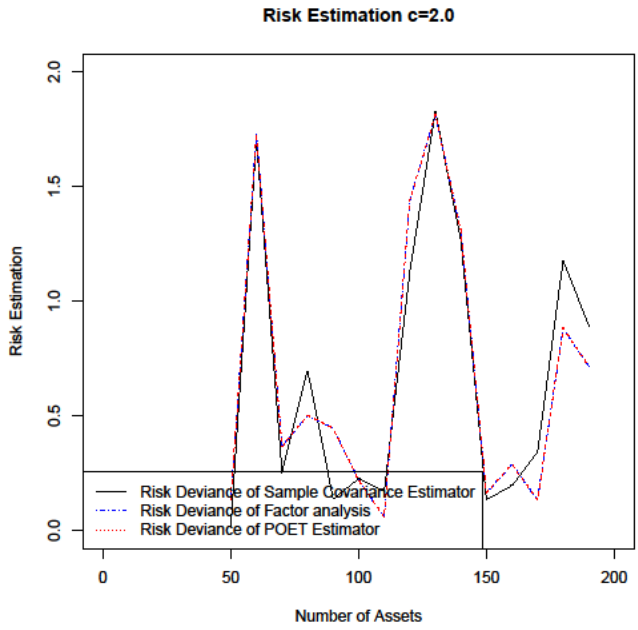

Figure 6.6 Risk Estimation Comparison when c equals to 2.0

Taking T=180 for example, we have the deviation of risk estimation under risk exposure takes value 1, 1.2, 1.4, 1.6, 1.8, 2.0.

c	**1**	**1.2**	**1.4**	**1.6**	**1.8**	**2**
Δ_S	1.1333	1.1523	1.1710	1.1896	1.2079	1.2260
Δ_P	0.9224	0.9226	0.9232	0.9242	0.9256	0.9274

Table 6.2 Risk estimation deviance of estimators

The above table shows that when the number of stocks in a certain portfolio is large enough compared to the number of observations in the dataset, the POET estimator performs better than the sample covariance estimator. And the POET estimator based on factor model can give meaningful insights into where these risks come from.

Chapter 7

Application in portfolio risk estimation

We will present some risk estimation results with the POET method. Consider a portfolio consisting of 254 most frequently traded stocks chosen from Shanghai Stock Exchange and Shenzhen Stock Exchange in China. We choose the risk exposure ranging from 1 to 2 and randomly simulate the allocation weights for securities.

Based on the closing prices of these stocks from January 4th , 2010 to December 1st 2014, we calculate the sample covariance estimator and the POET estimator for the monthly return correlations. Based on these estimators, we further obtain the risk estimation for the portfolios in terms of variances of monthly return:

Exposure (c)	Sample Covariance	POET Estimator
1.0	0.2775	0.4302
1.1	0.3866	0.6941
1.2	0.3585	0.6623
1.3	0.4320	0.8110
1.4	0.5800	1.2187
1.5	0.3066	0.4865
1.6	0.4757	0.9086

1.7	0.4300	0.7778
1.8	0.2624	0.4658
1.9	0.3608	0.5924
2.0	0.6127	1.2351

Table 7.1 Comparison of risk estimation based on different methods

We observe that the risk estimator based on the POET method is consistently higher than the risk estimator based on the sample covariance estimator. Given the POET estimator performs better than the covariance estimator in high-dimensional cases, this indicates that the risk could be underestimated under the sample covariance method.

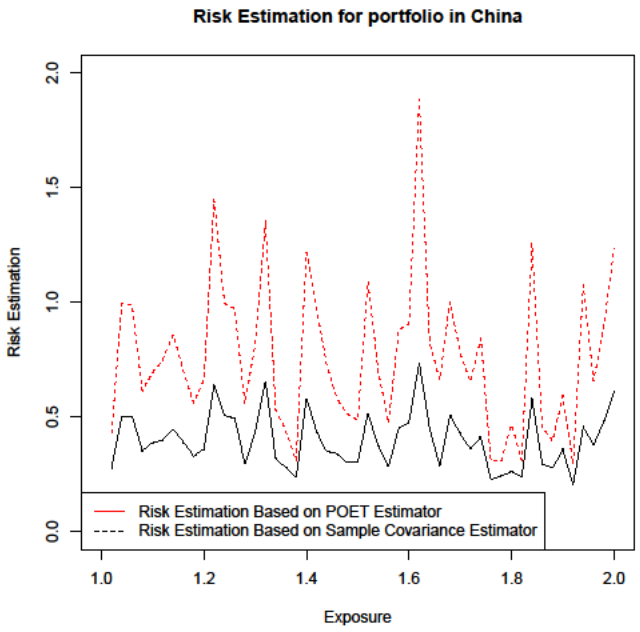

Figure 7.1 Risk estimation comparison of portfolios in China's stock market

Chapter 8

Discussion and future research directions

The massive data in financial research combined with quantitative methods reveals the risks imbedded in portfolios. In this project, we study the POET method which is based on principal component analysis and thresholding process. The POET method can yield better estimation for high-dimensional covariance matrices given the time varying property of risk in financial contexts.

When the POET method is applied in assessing the risk of portfolios in the China's stock market, the POET method gives a risk estimation higher than the risk estimation based on sample covariance estimation. That means the risk taken by investors is consistently underestimated.

In future research, we will focus on large inverse covariance matrix estimation, high-dimensional variable selection, and their applications in financial decisions.

Reference

[1] Adam J Rothman, Elizaveta Levina, and Ji Zhu. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177-186, 2009.

[2] Jianqing Fan, Fang Han, and Han Liu. Challenges of big data analysis. *National science review*, 1(2):293-314, 2014.

[3] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83-85, 2005.

[4] Jianqing Fan, Yuan Liao, and Martina Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):603-680, 2013.

[5] YQ Yin, ZD Bai, and PR Krishnaiah. On the limit of the largest eigenvalue of the large dimensional sample covariance matrix. *Probability theory and related fields*, 78(4):509-521, 1988.

[6] Stuart Geman. A limit theorem for the norm of random matrices. *The Annals of Probability*, pages 252-261, 1980.

[7] Jianqing Fan, Yuan Liao, and Xiaofeng Shi. Risks of large portfolios. *Available at SSRN 2211869*, 2013.

[8] Froil_an M Dopico. A note on sin θ theorems for singular subspace variations. *BIT Numerical Mathematics*, 40(2):395-403, 2000.

[9] Slavi_sa Djordjevic, In Ho Jeon, and Eungil Ko. Weyl's theorem through local spectral theory. *Glasgow Mathematical Journal*, 44(02):323-327, 2002.

Appendix

R code

```
# Code for estimating the bias of sample covariance matrix

setwd("E:/4 project/DATA/finalcode")

# set n=600 and p changes

dev.off()

pdf(file="a")

par(mfrow=c(3,2))

n=600

p=6

np=n*p

set.seed=101

data=rnorm(np,mean=0,sd=1)

mat<-matrix(data,ncol=p,nrow=n)

cov<-cov(mat)

b<-eigen(cov)

value<-b$values

plot(value,xlim=c(0,3),ylim=c(0,5),pch=1,xlab="p/n=0.01", ylab="eigenvalues")

abline(h=1,lty=1,col="red")
```

```
# code for Monte Carlo simulation
c=1.2
a=50
T=200

#risk factor matrix f 3 by 200
epsilon=mat.or.vec(3,T)
for (i in 1:T){
  temp3=i
  set.seed(temp3)
  epsilon[,i]=as.vector(rnorm(3))
}
mu=matrix(c(0.0260,0.0211,-0.0043),nr=3,nc=1)
phi=matrix(c(-0.1006,-0.0191,0.0116,0.2803,-0.0944,-0.0272,-
0.0365,0.0186,0.0272),nr=3,nc=3)

f=mat.or.vec(3,T)
f[,1]=mu
for (i in 2:T){
  j=i-1
```

```
  f[,i]=mu+ phi%*%f[,j]+epsilon[,i]
}

# Variance of risk factors
varf<-matrix(c(3.2351,0.1783,0.7783,0.1783,
               0.5069,0.0102,0.7783,0.0102,0.6586)
             ,byrow=T,ncol=3)

riskerrS<-mat.or.vec(15,10)
riskerrF<-mat.or.vec(15,10)
riskerrP<-mat.or.vec(15,10)


for (k in 1:15){
  p=a+k*10

  for (m in 1:10){
    seed<-m*5+1
    set.seed(seed)

    ## weight of stocks allocation
    K=rbinom(1, p, 0.70)
```

```
Temp=K+1
w=mat.or.vec(p,50)
for (t in 1:50){
  temp1=t
  set.seed(temp1)
  lamda=rexp(p, rate = 1)
  for (i in 1:K){
    w[i,t]=(1+c)*lamda[i]/(2*sum(lamda[1:K]))
  }
  for (j in Temp:p){
    w[j,t]=(1-c)*lamda[j]/(2*sum(lamda[Temp:p]))
  }
}

Mu<-c(0.5,1,0.8)
Sig<-diag(c(1,1,1))
B<-mvrnorm(n =p, Mu, Sig)
MuE<-rep(0,p)
CovE<-diag(0.01*c(1:p))
E<-t(mvrnorm(n=T,MuE,CovE))

# Matrix of Return
```

```
    R=B%*%f+E

    # theoratical covariance matrix
    sigma<-B%*%varf%*%t(B)+CovE

    ### estimate the covariance matrix
    covS=cov(t(R))
    poet<-POET(covS,3,0.9,thres="soft",matrix="cor")
    covF<-poet$SigmaY
    covP<-poet$SigmaY+poet$SigmaU

    riskerrS[k,m]<-mean(diag(t(w)%*%(covS-sigma)%*%w))
    riskerrF[k,m]<-mean(diag(t(w)%*%(covF-sigma)%*%w))
    riskerrP[k,m]<-mean(diag(t(w)%*%(covP-sigma)%*%w))
  }
}
setwd("E:/4 project/DATA/finalcode")
errS2<-abs(rowMeans(riskerrS))
errF2<-abs(rowMeans(riskerrF))
errP2<-abs(rowMeans(riskerrP))

write.table(errS2, "errS2.txt", sep="\t")
```

```
write.table(errF2, "errF2.txt", sep="\t")
write.table(errP2, "errP2.txt", sep="\t")



dev.off()
pdf(file="new2.pdf")
par(mfrow=c(1,1))
x<-seq(50,190,10)
plot(1, type="n", main="Risk Estimation c=1.2",xlab="Number of Assets",ylab="Risk Estimation", xlim=c(0, 200), ylim=c(0, 1.5))
lines(x,errS2,lty=1,col="Black")
lines(x,errF2,lty=4,col="blue")
lines(x,errP2,lty=2,col="Red")
legend("bottomleft", c("Risk Deviance of Sample Covariance Estimator","Risk Deviance of Factor analysis",
                       "Risk Deviance of POET Estimator"), lty=c(1,4,3), col=c("Black","blue","Red"))
dev.off()
```