



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

M.S. THESIS

High-dimension stock price modeling and  
Evaluation of the model-based portfolio  
optimization in the Korea stock market

한국주식시장에서의 고차원 주가 모형 추정과 이를  
이용한 최적 포트폴리오 생성 및 성능평가

BY

KIM DONG-WUK

JANUARY 2021

DEPARTMENT OF STATISTICS  
SEOUL NATIONAL UNIVERSITY

M.S. THESIS

High-dimension stock price modeling and  
Evaluation of the model-based portfolio  
optimization in the Korea stock market

한국주식시장에서의 고차원 주가 모형 추정과 이를  
이용한 최적 포트폴리오 생성 및 성능평가

BY

KIM DONG-WUK

JANUARY 2021

DEPARTMENT OF STATISTICS  
SEOUL NATIONAL UNIVERSITY

# High-dimension stock price modeling and Evaluation of the model-based portfolio optimization in the Korea stock market

한국주식시장에서의 고차원 주가 모형 추정과 이를  
이용한 최적 포트폴리오 생성 및 성능평가

지도교수 이상열  
이 논문을 이학석사 학위논문으로 제출함

2020년 12월




서울대학교 대학원

통계학과

김동욱

김동욱의 이학석사 학위 논문을 인준함

2021년 1월

위원장:	이재용	
부위원장:	이상열	
위원:	원중호	

# Abstract

In this thesis, we propose a model-driven statistical arbitrage method, with application to the Korean stock market from January, 2009 to December, 2020. Specifically, we first estimate high-dimensional systematic risks with principle component analysis. Subsequently, with the estimated systematic risks, we then employ the mean-reverting and volatility clustering strategies, which are representative characteristics frequently observable in various finance data. Unlike previous researches which attempted to model the idiosyncratic risk via stochastic process models, we instead consider a systematic risk-based autoregressive model. Moreover, based on our proposed model, we construct a conditional mean-variance optimized portfolio by building upon Markowitz's mean-variance optimized portfolio method. Our results show that our optimized portfolio outperforms other signal-based strategies throughout the analysis period.

**keywords:** Statistical arbitrage, autoregressive model, principal component analysis, sparse PCA, mean-variance portfolio optimization, asset pricing model, Korean securities

**student number:** 2019-24588

# Contents

<b>Abstract</b>	<b>i</b>
<b>Contents</b>	<b>ii</b>
<b>List of Tables</b>	<b>iv</b>
<b>List of Figures</b>	<b>v</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Exploratory Data Analysis</b>	<b>4</b>
2.1 Data Description . . . . .	4
2.2 Stock returns . . . . .	5
2.3 Covariance of stock returns . . . . .	8
<b>3 Quantative method for risk factors</b>	<b>10</b>
3.1 The PCA approach . . . . .	10
3.2 Covariance vs correlation matrix . . . . .	11
3.3 Sparse PCA . . . . .	13
3.4 Industry based PCA . . . . .	13
<b>4 Portfolio strategy</b>	<b>15</b>
4.1 Mean reversion and momentum . . . . .	15
4.2 Korea stock market; reversion vs. momentum . . . . .	16

4.3	Volatility clustering . . . . .	21
<b>5</b>	<b>Portfolio optimization</b>	<b>23</b>
5.1	Mean-variance portfolio optimization . . . . .	23
<b>6</b>	<b>Model description</b>	<b>25</b>
6.1	Model description . . . . .	25
6.2	Model selection . . . . .	26
<b>7</b>	<b>Results</b>	<b>29</b>
<b>8</b>	<b>Concluding Remark</b>	<b>33</b>
	<b>Abstract (In Korean)</b>	<b>37</b>

# List of Tables

4.1	Summary of significant AR(1) models. . . . .	18
4.2	Summary of the performance of portfolios . . . . .	20
4.3	Basic study with various PCA methodologies. . . . .	21
6.1	Summary of the portfolio performance . . . . .	27
7.1	Basic study with various PCA methodology . . . . .	30
7.2	Summary of conditional optimized portfolio's performance in the training set. . . . .	32
7.3	Summary of conditional optimized portfolio's performance in the test set. . . . .	32



# List of Figures

2.1	The number of listed (orange) and delisted (green) stocks. . . . .	5
2.2	QQ plot of returns in Korea equity market . . . . .	6
2.3	QQ plot with returns before (left) and After June, 2015 (right) . . . . .	6
2.4	Histogram of diagonal components of the sample covariance matrix (left) and JS-based sample covariance (right). . . . .	8
2.5	Difference between covariances estimated at initial time $t_0$ and time $t$ . . . . .	9
3.1	Explained variance ratio of PCA with the sample covariance (left) and the correlation matrix (right). . . . .	11
3.2	QQ plot of the pure returns (left) and the residuals obtained from the principal components (right). . . . .	12
3.3	Difference between the principal components ordered by eigenvalues (left) and similarities (right) within $k \in \{1, \dots, K\}$ . . . . .	12
3.4	QQ plot of the residuals explained by PCA with covariance (left) and correlation (right). . . . .	13
4.1	Profit and Loss(%) of portfolios. . . . .	17

4.2	Histogram and its estimated density function of the AR(1) coefficients using the kernel density estimation scheme. Left above plot is from coefficients estimated from $R_t$ , right above is from $\hat{R}_t$ , bottom left is from $r_t$ , and bottom right is the comparison of the estimated densities of $R_t, \hat{R}_t, \hat{\chi}_t$ . . . . .	19
4.3	Histogram of GARCH coefficients and long run variance based on fitting GARCH(1,1) model. . . . .	22
7.1	Profit and loss(%) of portfolios in training dataset from January, 2009 to May, 2017. . . . .	31
7.2	Profit and loss(%) of portfolios in Test dataset from June, 2017 to December, 2020. . . . .	31

# Chapter 1

## Introduction

Modeling stock prices has been one of the primary interests in various research areas, including economics and finance. Particularly, in economics, the capital asset pricing model (CAPM), which describes the relationship between the stock market and the expected return of individual stocks, was introduced and studied by various researchers, including Treynor (1961), Sharpe (1964), Lintner (1965), and Mossin (1966). Later, Pama and French (1993) suggested the three-factor model, which includes more systematic risk factors upon explaining the individual risks of the stock. However, these asset pricing models (APM) cannot well capture the return on individual shares in the short-term period, because APMs are highly correlated with long-term variables and they primarily focus on the correlation between risk factors and individual shares. Moreover, they fail to recognize the time-dependent characteristics of the short-term stock price models such as the mean-reversion behavior of stock price, see Poterba (1988) and Mukherji (2010). For this reason, models such as APM may not be sufficient enough when inspecting the structural behavior of the stock market.

To alleviate the drawbacks, several statistical models present in the literature can be utilized to capture the structure of sequentially observed time series datasets. To illustrate, the autoregressive moving average (ARMA) coined by Whittle (1951) is utilized for modeling the conditional mean in a given time series, and is further espoused

by the model selection scheme and the estimation method by Box and Jenkins (1954). Furthermore, the volatility clustering phenomenon, initially discovered by Mandelbrot (1963), can be modeled with the autoregressive conditional heteroscedasticity (ARCH; Engle, 1982) and the generalized ARCH (GARCH; Bollerslev, 1986) models. As such, Li et al. (2002) suggested the ARMA-GARCH model for simultaneously modeling the conditional mean and variance.

Some techniques for building a portfolio based on mathematical models have also been extensively reviewed. To elaborate, the mean-variance portfolio theory, first established by Markowitz (1952), provided a way to construct the mean-variance optimized portfolio called the ‘efficient frontier’. Also, Black and Litterman (1992) considered the mean-variance optimized portfolio that models the expected behavior of market participants. However, the mean-variance optimized portfolio model suffer from a critical drawback: the covariance matrix, which is necessary when constructing the optimized portfolio, can easily face rank deficiency problems, ultimately leading to inconsistent estimates, see Choi et al. (2019). To avoid this problem, a portfolio is conventionally constructed based on some key factors, such as business sectors systematic factors; see Fama and French (2015).

Recently, many researches have developed strategies to construct portfolios which takes into account both systematic risks across multi-dimensional time series of stock prices and the sequential characteristics of individual returns. Avellaneda (2010) proposed a factor model with mean-reverting residuals, then applied it to the U.S. equity market. Guijarro-Ordóñez (2019) derived the closed-form optimal strategies via PCA-based systematic factors estimation. However, they applied the statistical methodologies to analyze only idiosyncratic risks, but not systematic risks. Thus, this thesis aims to propose a model which considers systematic risks upon constructing the portfolio with AR models based on systematic factors. Moreover, we improve upon Markowitz’s optimized portfolio to construct our conditional mean-variance optimized portfolio. Our analysis in subsequent chapters reveals that considering systematic risks is more

important than idiosyncratic risks in forecasting stock returns.

The rest of the thesis is structured as follows. Chapter 2 provides comprehensive properties regarding the Korean stock market. Chapters 3 and 4 introduce quantitative methods to model risk factors and portfolio strategies. Chapter 5 explains mean-variance optimized portfolio method in detail. The proposed model and its conditional mean-variance optimized portfolio is described in Chapter 6. Chapter 7 depicts the performance of our proposed portfolio, compared against other strategies. We provide discussion regarding the results and the potential merits of our research in Chapter 8.

## **Chapter 2**

### **Exploratory Data Analysis**

In this chapter, we provide some preliminary information regarding the Korean stock market data, which we utilize in subsequent chapters to assess the performance of portfolios. Also, we summarize some basic concepts regarding portfolio construction.

#### **2.1 Data Description**

We utilize the Korean stock market dataset throughout the thesis, which is obtained from the DataGuide database provided by FNGuide. The dataset includes all currently listed and delisted stocks (3,168 stocks) in Korea from January, 2009 to December, 2020, and is observed on a daily based (3,131 business days). We include all delisted stocks when evaluating the performance, which is crucial when preventing survival bias. Figure 2.1 shows the number of listed and delisted stocks from January, 2009 to December, 2020.

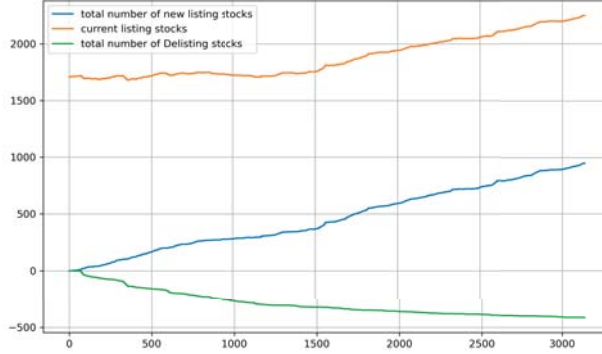


Figure 2.1: The number of listed (orange) and delisted (green) stocks.

## 2.2 Stock returns

Stock prices are often log-normally distributed, and its log-normality can be confirmed from the stochastic differential equation as follows:

$$dS_t = \mu S_t dt + \sigma S_t dW_t, \quad (2.1)$$

where  $W_t$  is a Brownian motion,  $\mu$  is a drift term, and  $\sigma^2 < \infty, \sigma^2 > 0$  is a constant volatility. If the drift term is nonzero, we call this process the geometric Brownian motion (GBM). Therefore, the more drift term becomes larger, the more the return of stocks has a fat-tailed distribution. For this reason, we will use the logarithmic return:

$$r_t = \log\left(\frac{P_t}{P_{t-1}}\right), \quad (2.2)$$

where  $P_t$  is the price at time  $t > 0$ . Figure 2.2 portrays the QQ plot with all stock price returns.

From Figure 2.2, some abnormal patterns can be found at  $|r_t| = 15$  and  $|r_t| = 30$ . This is because the restriction of the daily price range of the Korean market changed from 15 percent to 30 percent in June, 2015. Hence, Figure 2.3 plot separate QQ plots prior and posterior June, 2015.

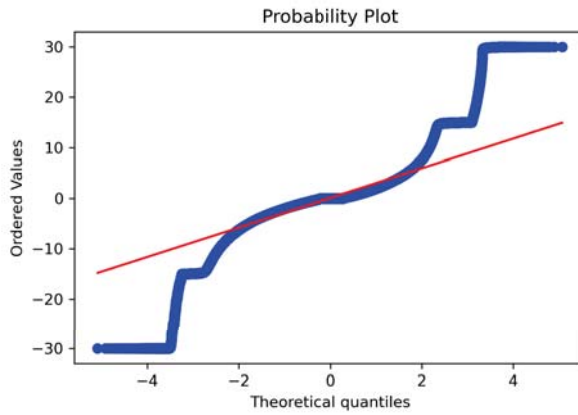


Figure 2.2: QQ plot of returns in Korea equity market

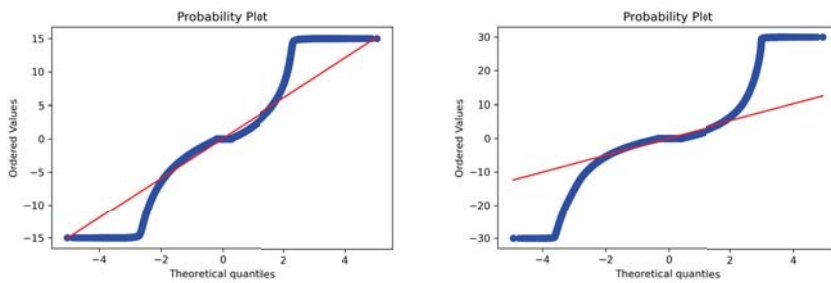


Figure 2.3: QQ plot with returns before (left) and After June, 2015 (right)



Figure 2.3 implies that there exist some drift terms among stock returns. To properly analyze high-dimensional dataset, we use James-Stein (JS) estimator, proposed by Stein (1956). It is known that the JS estimator dominates the least square (LS) estimator when the dimension is large:  $p \geq 3$ . If the variance  $\sigma^2$  is known,  $\mathbf{x} \sim N_p(\theta, \sigma^2 I)$ ,

$$\hat{\boldsymbol{\theta}}_{JS} = \left(1 - \frac{(p-2)\sigma^2}{\|\mathbf{x}\|^2}\right) \mathbf{x}. \quad (2.3)$$

However, when  $\sigma^2$  is unknown, we have to estimate it with  $\hat{\sigma}^2$ . Let  $X_{n \times p} = [x_1, \dots, x_n]^T$ ,  $p \geq 3, n \geq 1$ ,

$$\hat{\sigma}^2 = \frac{1}{p(n-1)} \text{tr}(X^T(I-J)X). \quad (2.4)$$

The difference between LS estimator and JS estimator is calculated as follows:

$$\frac{\|\hat{\boldsymbol{\mu}}_{LS} - \hat{\boldsymbol{\mu}}_{JS}\|}{\|\hat{\boldsymbol{\mu}}_{LS}\|} = 0.375, \quad \frac{\|\hat{\boldsymbol{\mu}}_{JS}\|}{\|\hat{\boldsymbol{\mu}}_{LS}\|} = 0.625. \quad (2.5)$$

## 2.3 Covariance of stock returns

We often face the rank deficiency problem when we obtain a covariance matrix with a high dimensional dataset. To remedy the problem, various regularized estimation method have been proposed, refer to Choi et al. (2019). Before conducting a further study regarding covariance of returns, we first look at some characteristic of covariance of stock returns. Let the demeaned variant of  $\mathbf{x}$  be denoted as  $\mathbf{x}'_{JS} = \mathbf{x} - \hat{\mu}_{JS}$  where  $\mathbf{x} \in \mathbf{R}^p$ . Then, using the JS estimator, the covariance matrix can be obtained as follows:

$$\hat{\Sigma}^{JS} = \frac{1}{n-1} \mathbf{X}_{JS}^T \mathbf{X}_{JS} \quad (2.6)$$

Figure 2.4 plots a histogram of diagonal components of the sample covariance matrix and JS-based sample covariance.

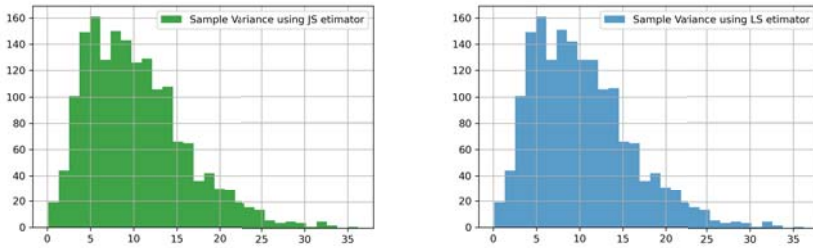


Figure 2.4: Histogram of diagonal components of the sample covariance matrix (left) and JS-based sample covariance (right).

As seen in Figure 2.4, the difference between the sample covariance and JS-based covariance is small, namely  $\|\hat{\Sigma}^{JS} - \hat{\Sigma}\|_F / \|\hat{\Sigma}\|_F = 0.0016 \simeq 0$ . The sample covariance was calculated over a specific horizon window to observe a change of covariance over time. Specifically, the sample covariance at time  $t$ , denoted as  $\hat{\Sigma}_t$ , is obtained with  $X_{w \times p} = [x_t, \dots, x_{t+w}]^T$ , where  $w$  denotes the prescribed window size. Moreover, the

difference between two covariance matrices are defined as follows:

$$\Delta\Sigma_t = \frac{\|\hat{\Sigma}_t - \hat{\Sigma}_0\|_F}{\|\hat{\Sigma}_0\|_F} \quad (2.7)$$

We can notice that the structure among returns are changed along time. As depicted in Figure 2.5, if we have a small window size to estimate the covariance, then it changes more rapidly with as time difference becomes large.

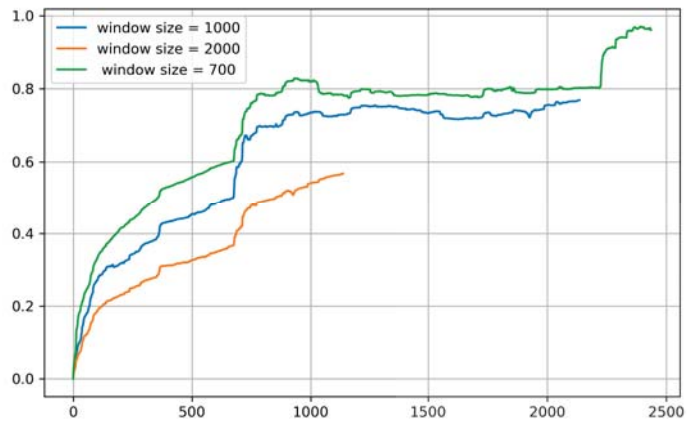


Figure 2.5: Difference between covariances estimated at initial time  $t_0$  and time  $t$ .

## Chapter 3

### Quantative method for risk factors

#### 3.1 The PCA approach

In the asset pricing model (APM), the returns of each stock  $i \in \{1, \dots, p\}$  can be described as

$$r_{i,t} = \sum_{k=1}^K \beta_{k,i} F_{k,t} + \mu + \chi_{i,t}, \quad (3.1)$$

where  $F_{k,t}$ ,  $k = 1, \dots, K$  is the risk factor at time  $t > 0$ ,  $\chi_{i,t}$  is the idiosyncratic component which is independent of any  $F_{k,t}$ , and  $\mu \in \mathbf{R}^p$  is the risk-free return. We assume that  $F_{k,t}$  ( $k = 1, \dots, K$ ) is independent to other systematic risks  $F_{j,t}$ ,  $j \neq k$  and  $j \in \{1, \dots, K\}$ . By assumption, the covariance matrix

$$Cov(r) = \beta\psi\beta^T + \phi \quad (3.2)$$

where  $\beta = [\beta_{ij}]$ ,  $i = 1, \dots, p$ ,  $j = 1, \dots, k$ .  $\psi = diag(var(F_i))$ ,  $\phi = diag(var(\chi_i))$ .

We can use the principal component analysis (Jolliffe, 1986) to estimate the candidates of systematic risk factors. With PCA, we can rewrite the covariance matrix as

$$Cov(r) - \phi = \beta\psi\beta^T = U\Lambda U^T, \quad (3.3)$$

where  $U$  is orthonormal eigenvectors and  $\Lambda = \text{diag}([\lambda_1 \cdots, \lambda_p])$ . We can estimate  $\hat{\psi}, \hat{U}$  by iterative methods.

Figure 3.1 shows the explained variance ratio, which is defined as follows:

$$\text{explained variance ratio}(n) = \sum_{i=1}^n \frac{\lambda_i}{\sum_{j=1}^p \lambda_j} \quad (3.4)$$

Principal components are estimated with using data from January, 2009 to June, 2017.

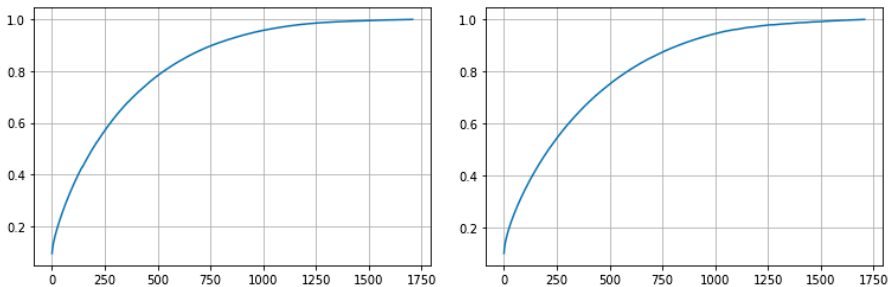


Figure 3.1: Explained variance ratio of PCA with the sample covariance (left) and the correlation matrix (right).

Residual obtained from the principal components are defined as

$$\text{residual}(\chi_{i,t}) = r_{i,t} - \sum_{k=1}^K (u_k^T (R_t - \mu)) \cdot u_k - \mu, \quad (3.5)$$

where  $R_t = [r_{1,t}, \cdots, r_{p,t}]^T$ . Figure 3.2 shows that residuals obtained after performing PCA has lower divergence compared to that of the pure return, and is distributed much closer to the normal distribution.

## 3.2 Covariance vs correlation matrix

There are two options to estimate systematic risk factors: estimating the covariance matrix or the correlation matrix. In this section, we investigate the difference between the two matrices in terms of the principal components. Here, we use the same dataset which was utilized in Section 3.1.

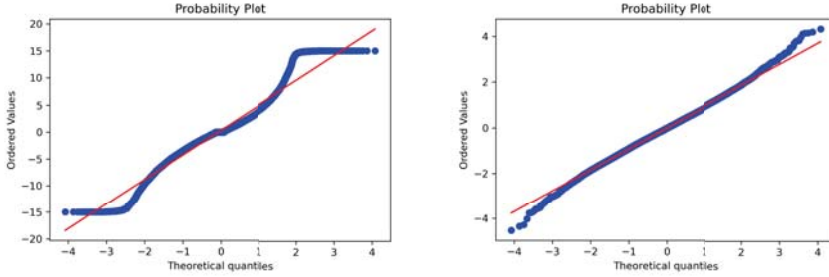


Figure 3.2: QQ plot of the pure returns (left) and the residuals obtained from the principal components (right).

Figure 3.3 illustrates the difference of the principal components between the two matrices. The difference is measured by eigenvalue orders and similarities within the principal components with in  $k \in \{1, \dots, K\}$ .

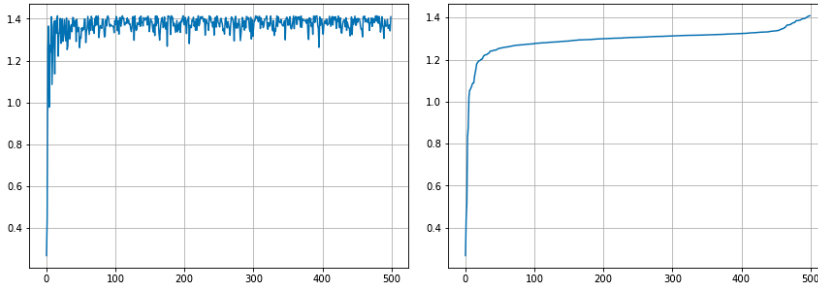


Figure 3.3: Difference between the principal components ordered by eigenvalues (left) and similarities (right) within  $k \in \{1, \dots, K\}$ .

Moreover, the difference between principal components  $v_i$  and  $v_j$  is defined as

$$\text{diff}(v_i, v_j) = \frac{\min(\|v_i - v_j\|, \|v_i + v_j\|)}{\|v_i\|}. \quad (3.6)$$

Figure 3.4 shows that the residuals obtained from the principal components which is estimated with the covariance matrix is more normally distributed than those estimated with the correlation matrix.

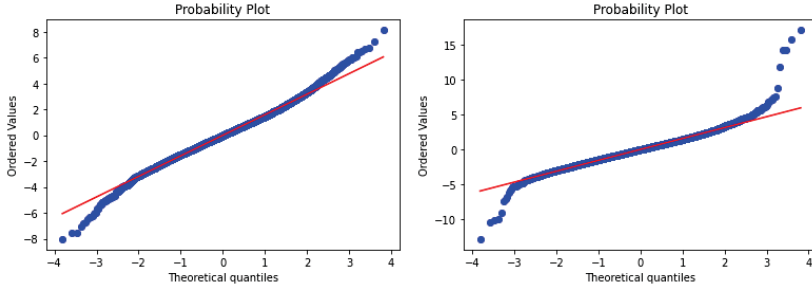


Figure 3.4: QQ plot of the residuals explained by PCA with covariance (left) and correlation (right).

### 3.3 Sparse PCA

The main problem of PCA becomes apparent when the dimension of the dataset gets large. Principal components are the linear combination of  $p$  variables, thus are typically nonzero. To alleviate the problem, Zou et al. (2006) considered the sparse PCA (SPCA), which considers the variable selection upon computing the principal components. To compute the principal components, LASSO (Tibshirani, 1996) can be employed to estimate the modified principal components sparsely. Specifically, the general SPCA algorithm by Zou et al. (2006) aims to minimize Equation 3.7, then solve it via numerical methods.

$$\begin{aligned}
 (\hat{\mathbf{U}}, \hat{\mathbf{L}}) = \arg \min_{\mathbf{L}, \mathbf{U}} \sum_{i=1} \|\mathbf{x}_i - \mathbf{U}\mathbf{L}^T \mathbf{x}_i\|^2 + \lambda \sum_{j=1} \|l_j\|^2 \\
 \text{subject to } \mathbf{U}^T \mathbf{U} = I_{k \times k}
 \end{aligned} \tag{3.7}$$

### 3.4 Industry based PCA

If some domain knowledge is available, it can be utilized upon employing quantitative methods. For instance, FNGuide provides information regarding business sectors, and it can be useful because the stocks within the same sector tends to be highly correlated.

Therefore, we can reformulate the returns as the sector-based APM models:

$$r_{i,t} = \sum_{s=1}^S \sum_{k \in K_s} \beta_{k,i} F_{k,t} I_{(i \in K_s)} + \chi_{i,t}, \quad (3.8)$$

where  $I$  is an indicator function, and  $K_s$  is an index set of stocks in sector  $s$ . Using this model, we have blockwise diagonal covariance matrix of  $r$ . With some additional constraints, we can construct more robust estimation of risk factors, provided that the assumption is true.



## Chapter 4

### Portfolio strategy

#### 4.1 Mean reversion and momentum

The mean-reversion strategy is one of the trends of the financial industries to formulate a portfolio in the securities market. The gist of the concept is as follows: **(1)** some quantities are historically correlated, and **(2)** one expects that the correlation will be restored in the future, if these correlations are disrupted by some unusual market conditions. (Kakushadze, 2014)

To elaborate, let  $P_A(t_1)$  and  $P_B(t_1)$  be the prices of A and B at time  $t_1$ , and let  $P_A(t_2)$  and  $P_B(t_2)$  be the prices of A and B at a later time  $t_2$ . Then, we define  $R_A$  and  $R_B$  as follows:

$$\begin{aligned} r_A &= \log \frac{P_A(t_2)}{P_A(t_1)} \\ r_B &= \log \frac{P_B(t_2)}{P_B(t_1)}. \end{aligned} \tag{4.1}$$

If there are two stocks A and B, and  $r_A > r_B$ , then this information implies that A is relatively expensive, and vice versa. Then, in order to gain profit, we ‘short’ and buy stocks A, B, respectively. That is, we make a bet on A that the price of A will decrease

in the next time period. In equations, this chain of events can be summarized as:

$$\begin{aligned}
\bar{r} &\equiv \frac{1}{2} (r_A + r_B) \\
\tilde{r}_A &\equiv r_A - \bar{r} \\
\tilde{r}_B &\equiv r_B - \bar{r},
\end{aligned} \tag{4.2}$$

where  $\bar{r}$  denotes the mean return. This encapsulates the mean-reversion strategy if there are two participants in the securities market.

One can readily extend the pairwise trading methodology presented above to multiple stocks. Let  $r_i, i = 1, \dots, N$  be

$$\begin{aligned}
r_i &= \ln \left( \frac{P_i(t_2)}{P_i(t_1)} \right) \\
\bar{r} &\equiv \frac{1}{N} \sum_{i=1}^N r_i \\
\tilde{r}_i &\equiv r_i - \bar{r}.
\end{aligned} \tag{4.3}$$

We can decide on the amount of investment for each stock proportional to  $\tilde{r}_i$ .

$$w_i = c\tilde{r}_i. \tag{4.4}$$

We call this strategy the mean-reversion strategy if  $c < 0$ , and the momentum strategy otherwise.

## 4.2 Korea stock market; reversion vs. momentum

In this section, a simple preliminary study on Korean stock market is conducted. Since a portfolio is generated and updated by the daily basis after the market closes, implementation of the portfolio should be delayed one day from its generation. Here, we consider four types of portfolios. First, we consider a strategy where we equally buy then hold all stocks with the same weight. Second, a simple portfolio is generated each day using Equation 4.4, where

$$c = \frac{1}{\sum_{i=1}^p |\tilde{r}_{i,t}|}. \tag{4.5}$$

Third, we formulate the strategy which use  $\hat{R}_t$

$$\hat{R}_t = \sum_{k=1}^K u_k u_k^T (R_t - \mu) + \mu \quad (4.6)$$

instead of  $R_t = [r_{1,t}, \dots, r_{p,t}]^T$ , and generate the portfolio with Equations 4.4 and 4.5. Finally, the fourth portfolio is constructed in the same manner as the second one, but instead, we utilize the projected residuals obtained after PCA, which is denoted as follows:

$$\hat{\chi}_t = R_t - \hat{R}_t, \quad (4.7)$$

where  $\chi_t = [\chi_{1,t}, \dots, \chi_{p,t}]^T$ .

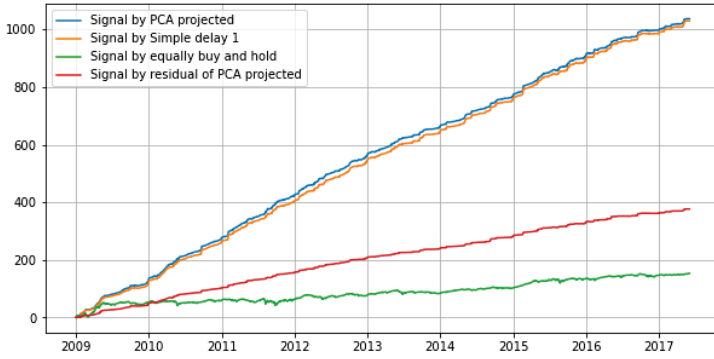


Figure 4.1: Profit and Loss(%) of portfolios.

Figure 4.1 summarizes the result of the four portfolios introduced above. This result shows that there is a strong evidence of relationship between previous return and current returns in Korea stock market. This strongly upholds the validity of employing AR(1) model, which effectively models the autocorrelation between returns. Thus, we consider the following AR(1) model:

$$X_{i,t} = a_i X_{i,t-1} + \epsilon_{i,t}, \quad (4.8)$$

where  $t > 0$  is a time index,  $a_i$  is a coefficient, and  $\epsilon_{i,t}$  is a random variable with mean

zero and variance  $\sigma_i^2$ . Here, we fit AR(1) model with  $R_t, \hat{R}_t$  and  $Z_t \in \mathbf{R}^K$  given as

$$Z_t = [u_1^T R_t, \dots, u_K^T R_t]^T. \quad (4.9)$$

Table 4.1: Summary of significant AR(1) models.

	return	Projected return	PCA scores
DIMENSION	1709	1709	500
NUMBER OF SIGNIFICANT AR(1)	1302	1623	391
RATIO(%)	76.2	95.0	78.2

The number of stocks which have a significant AR(1) coefficient is summarized in Table 4.1. Our results reveal that  $\hat{R}_t$ , the projected returns using PCA, appears to have more autoregressive structures than the time series of unmodified returns  $R_t$ . Using the Gaussian kernel density function, we draw the estimated distribution of AR coefficients of the returns in Figure 4.2. We can observe that the distributions of coefficients are positively skewed, indicating that the projected returns do indeed have more significant autoregressive structure.

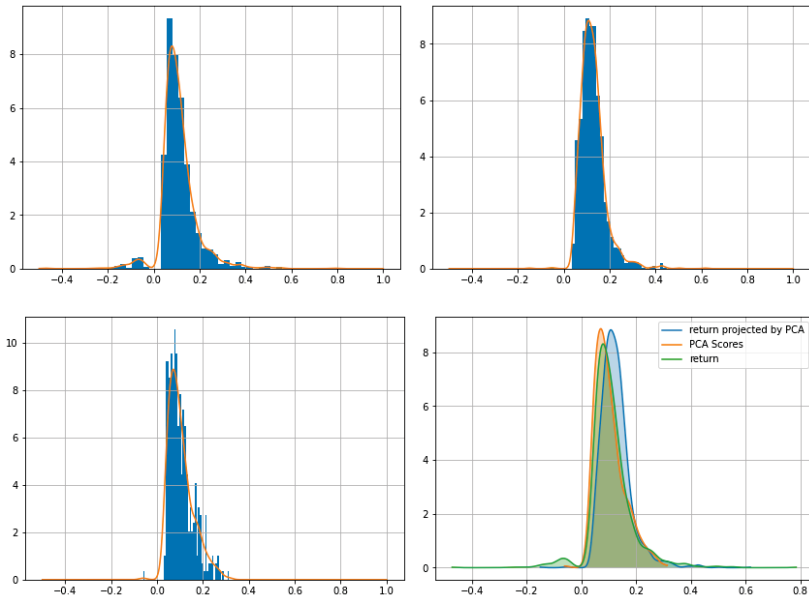


Figure 4.2: Histogram and its estimated density function of the AR(1) coefficients using the kernel density estimation scheme. Left above plot is from coefficients estimated from  $R_t$ , right above is from  $\hat{R}_t$ , bottom left is from  $r_t$ , and bottom right is the comparison of the estimated densities of  $R_t, \hat{R}_t, \hat{\chi}_t$ .

Table 4.2: Summary of the performance of portfolios

Date	equally buy and hold			Simple return( $R_t$ )			hat return( $\hat{R}_t$ )			residual PCA( $\hat{\chi}_t$ )		
	Sharpe ratio	PnL	Turnover	Sharpe ratio	PnL	Turnover	Sharpe ratio	PnL	Turnover	Sharpe ratio	PnL	Turnover
2009	2.47	49.3	0	7.24	116	134	7.46	124	134	-1.48	-5.20	139
2010	0.61	8.13	0	7.98	137	136	8.52	140	133	1.16	4.45	138
2011	0.11	2.25	0	7.35	134	133	7.75	141	132	0.86	3.64	137
2012	1.24	14.5	0	7.11	135	138	7.69	134	132	1.99	9.47	136
2013	0.68	7.07	0	6.25	96.9	139	7.18	95.6	132	2.12	11.1	136
2014	2.14	16.9	0	6.59	107	141	7.29	102	131	3.42	17.4	134
2015	2.38	31.5	0	6.49	132	134	7.05	134	133	0.72	3.15	136
2016	0.89	10.8	0	4.63	89.1	141	4.97	85.7	134	1.53	7.16	136
2017	1.94	14.7	0	4.93	90.1	140	5.54	82.8	130	2.57	14.6	132

### 4.3 Volatility clustering

Volatility clustering, initially coined by Mandelbrot (1963), is one of the most prominent characteristics of financial data. The GARCH model (Bollerslev, 1986) is the representative model to inspect the existence of the volatility clustering phenomenon and its magnitude in time series. The GARCH( $p,q$ ) model is characterized as follows:

$$\begin{aligned} X_t &= \eta_t \sigma_t \\ \sigma_t^2 &= \omega + \sum_{i=1}^p \alpha_i X_{t-i}^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2, \end{aligned} \quad (4.10)$$

where  $\eta_t$  are i.i.d. such that  $E(\eta_t) = 0$ ,  $\text{var}(\eta_t) = 1$  and  $\eta_t \perp \sigma_k$  such that  $\forall k \leq t$ .

We here observe the behavior of the residual  $\hat{\chi}_i$  in Equation 4.7 with GARCH(1,1). let  $\alpha = \alpha_1, \beta = \beta_1$ . the long run variance of GARCH(1,1) becomes

$$\sigma_L^2 = \frac{\omega}{1 - \alpha - \beta}, \quad (4.11)$$

where  $\alpha + \beta < 1$ . Table 4.3 and Figure 4.3 summarize the effect of the volatility clustering on the obtained residuals with GARCH(1,1).

Table 4.3: Basic study with various PCA methodologies.

Type	$\hat{\omega} \neq 0$	$\hat{\alpha} + \hat{\beta} > 0.99$	$\hat{\alpha} + \hat{\beta} > 0.95$	total number of stocks
Number	486	348	846	1709
Ratio(%)	28.4	20.4	49.5	100

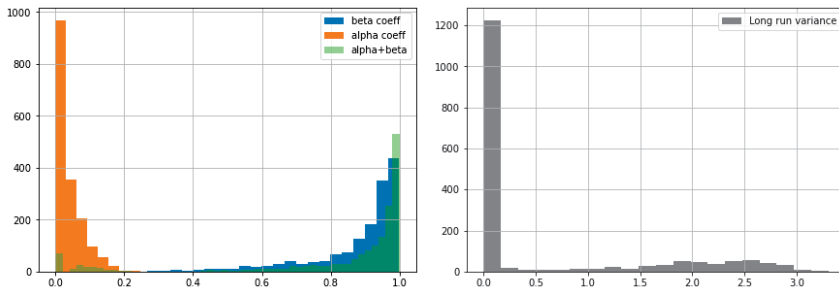


Figure 4.3: Histogram of GARCH coefficients and long run variance based on fitting GARCH(1,1) model.



## Chapter 5

### Portfolio optimization

#### 5.1 Mean-variance portfolio optimization

Mean-variance analysis (Markowitz, 1952) aims to maximize the return for a given risk level. Hence, the mean-variance optimization criterion is suggested as follows:

$$\text{minimize } g(w, \lambda) \equiv \frac{\lambda}{2} w^T \Sigma w - \sum_{i=1}^N R^T w, \quad (5.1)$$

where  $r$  is the  $r \in \mathbf{R}^p$  stock return with mean  $R \in \mathbf{R}^p$  and variance  $\Sigma$ ,  $w \in \mathbf{R}^p$  is the portfolio weight satisfying  $\|w\| = 1$ , and  $\lambda > 0$  can be chosen arbitrarily. Using Cauchy–Schwarz inequality, we have the following closed-form solution for  $w$ :

$$w = \frac{1}{\lambda} \Sigma^{-1} R. \quad (5.2)$$

We can additionally impose another constraint to the weight vector called the dollar neutrality, denoted as  $\mathbf{1}^T w = 0$ . To impose more constraint on the mean-variance criterion then we can rewrite the objective function of Equation 5.1 as

$$\text{minimize } g(w, \mu, \lambda) \equiv \frac{\lambda}{2} w^T \Sigma w - R^T w - w^T C u, \quad (5.3)$$

where  $C \in \mathbf{R}^{p \times m}$  is an  $m$ -homogeneous linear constraint, and  $u \in \mathbf{R}^m$  is a Lagrange multiplier for the constraint  $C$ . Using Sherman and Morrison (1950) and Woodbery

(1950), the optimal solution of the weight vector and the Lagrange multiplier can be obtained as the following:

$$\begin{aligned} w &= \frac{1}{\lambda} \left[ \Sigma^{-1} - \Sigma^{-1} C (C^T \Sigma^{-1} C)^{-1} C^T \Sigma^{-1} \right] R \\ \mu &= - (C^T \Sigma^{-1} C)^{-1} C^T \Sigma^{-1} R. \end{aligned} \tag{5.4}$$

We can consider adding more constraints and bounds on Equation 5.2 to consider turnovers or to impose more limits onto weight values, see Kakushadze (2014).

# Chapter 6

## Model description

### 6.1 Model description

From Section 3.2, we studied the autoregressive characteristics regarding the returns of stocks. Moreover, in Section 4.3, we revealed that the linear heteroscedasticity is not a general characteristic of the PCA-obtained residuals  $\chi_t$ . The significant difference between our proposed model and the ordinary latent variable model formulated with the stochastic process (Guijarro-Ordóñez, 2019 and Avellaneda, 2010) is that we consider the autocorrelation between stock returns. Furthermore, our proposed model constructs autoregressive models based on the systematic risk factors, not the idiosyncratic risk factors, denoted as follows:

$$\begin{aligned} R_t &= \sum_{k=1}^K u_k f_{k,t} + \mu + \chi_t \\ E(R_{t+1} | \mathcal{I}_t) &= \text{diag}(\phi_i) \cdot \left( \sum_{k=1}^K u_k f_{k,t} + \mu \right) \\ \text{cov}(R_t) &= U \Lambda U^T + \Psi, \end{aligned} \tag{6.1}$$

where  $u_k \in \mathbf{R}^p$  is the  $k$ -th systematic risk factor such that  $k \in \{1, \dots, K\}$ ,  $f_{k,t} \in \mathbf{R}$  is the score at time  $t$  of factor  $k$ , and  $\chi_t$  is the white noise process  $WN(0_{p \times 1}, \Psi)$ ,  $\phi_i$  is the autoregressive coefficient of the  $i$ -th equity return,  $\mu \in \mathbf{R}^p$  is the risk-free return,  $\Psi =$

$\text{diag}([\psi_1, \dots, \psi_p])$ ,  $U = [u_1, \dots, u_k]$  where  $U^T U = \mathbf{I}_k$ ,  $\lambda = \text{diag}([\lambda_1, \dots, \lambda_K])$ , and  $\lambda_k$  is the variance of  $k$ -th risk factor. From Chapter 5, we reviewed the fundamental concepts regarding the mean-variance optimized portfolio. Using the proposed model, we can rewrite the analytic mean-variance optimized weight of the portfolio (Equation 5.1) using conditional expectations. The proposed conditional portfolio optimized weight is given by

$$\begin{aligned}
w_{t+1} &= \arg \max_{\|w\|=1, w \in \mathbf{R}^p} E \frac{R_{t+1}^T w}{\sqrt{w^T \Sigma_{t+1} w}} = E (\Sigma_{t+1}^{-1} R_{t+1}) \\
&= E E (\Sigma_{t+1}^{-1} R_{t+1} | \mathcal{I}_t) \\
&= \Psi^{-1} E E [R_{t+1} | \mathcal{I}_t] (\because E(\Sigma_{t+1} | \mathcal{I}_t) = \Psi) \\
&= \Psi^{-1} E \left[ \text{diag}(\phi_i) \cdot \left( \sum_{k=1}^K u_k f_{k,t} + \mu \right) \right] (\because \psi_{t+1} \in \mathcal{I}_t) \\
&= \text{diag} \cdot \left( \frac{\phi_i}{\psi_i} \right) \left( \sum_{k=1}^K u_k u_k^T (r_t - \mu) + \mu \right).
\end{aligned} \tag{6.2}$$

## 6.2 Model selection

In this study, we consider the following parameter space:

$$\mathcal{F}_{p,K} = \{U \in \mathbf{R}^{p \times K}, \text{diag}(\phi_i) \in \mathbf{R}^{p \times p}, \Lambda \in \mathbf{R}^{K \times K}, \Psi \in \mathbf{R}^{K \times K}, \mu \in \mathbf{R}^p\}. \tag{6.3}$$

Hence, note that our proposed model depends on the number of principal components  $K$ , which is a decisive factor when computing the outcome. Kaiser's rule (1960) is the most common method to decide the number of principal components:

$$\lambda_k > \bar{\lambda} = \frac{1}{p} \sum_{i=1}^p \lambda_i. \tag{6.4}$$

Moreover, using the maximum likelihood framework, the probabilistic PCA model (Michael and Bishop, 1999) can be employed. That is,  $\mathbf{x} = \mathbf{Lz} + \mathbf{m} + \mathbf{e}$  where  $\mathbf{e} \sim N(0, v\mathbf{I}_p)$ ,  $\mathbf{z} \sim N(0, \mathbf{I}_K)$ . Following this concept, optimal  $K$  for PCA can be chosen

by using the Bayesian model selection method and Laplace transformation:

$$p(D | k) \approx \left( \prod_{j=1}^k \lambda_j \right)^{-N/2} \hat{v}^{-N(d-k)/2} N^{-(m+k)/2}, \quad (6.5)$$

where  $D = \{x_1, \dots, x_N\}$  is the observed data, see Minka (2000).

In this thesis, we consider two estimation schemes, namely the LS estimator and the JS estimator introduced in Section 2.2, for  $\hat{R}_i$  in equation 4.2. To elaborate, we have the following expressions:

$$\begin{aligned} \hat{R}_t^{LS} &= \sum_{k=1}^K \hat{u}_k \hat{u}_k^T (R_t - \hat{\mu}_{LS}) + \hat{\mu}_{LS} \\ \hat{R}_t^{JS} &= \sum_{k=1}^K \hat{u}_k \hat{u}_k^T (R_t - \hat{\mu}_{JS}) + \hat{\mu}_{JS}. \end{aligned} \quad (6.6)$$

We compare the performance of  $\hat{R}_t^{LS}$  and  $\hat{R}_t^{JS}$  in the same way as in Section 4.2.

Table 6.1: Summary of the portfolio performance

Type	JS estimator ( $\hat{R}_t^{JS}$ )		LS estimator ( $\hat{R}_t^{LS}$ )	
Period	Sharpe ratio	Profit and Loss(%)	Sharp ratio	Profit and Loss(%)
2009	7.32	143.1	7.35	143.7
2010	8.34	152.6	8.38	153.0
2011	6.72	166.5	6.74	166.6
2012	8.02	146.2	8.05	146.4
2013	7.20	110.0	7.23	110.1
2014	7.45	111.5	7.51	111.8
2015	7.58	158.7	7.62	159.1
2016	4.85	96.0	4.85	95.9
2017	5.14	83.5	5.15	83.4

The results are summarized in Table 6.1. It reveals that the portfolio constructed using the LS estimator outperforms that of the JS estimator throughout the whole pe-

riod. We conduct one-sided paired  $t$  tests on daily profit and loss, which is given as below:

$$H_0 : \text{pnl}_{JS} \geq \text{pnl}_{LS} \text{ vs. } H_1 : \text{pnl}_{JS} < \text{pnl}_{LS}, \quad (6.7)$$

where  $\text{pnl}_{JS}$  is the daily profit and loss when employing  $\hat{R}_t^{JS}$ , and  $\text{pnl}_{LS}$  is that of  $\hat{R}_t^{LS}$ . The  $p$ -value of the hypothesis is  $1.86 \times 10^{-3}$ , thus fortifying that the strategy which uses  $\text{pnl}_{LS}$  does indeed have a better performance compared to that of  $\text{pnl}_{JS}$ .

## Chapter 7

### Results

In this chapter, we present the results regarding the performance of the proposed model, when applied to Korean equity market from January, 2009 to December, 2020. First, we divide the data into two parts, the training and the test data. Specifically, the training data is from January, 2009 to May, 2017. Using Kaiser's rule (Kaiser, 1950), the estimated number of principal components is  $\hat{K} = 519$ , and  $\hat{K} = 334$  when estimated via the Minka's method. Principal components estimation problem was computationally solved by 'sklearn' decomposition package, see Minka (2001). Subsequently, using the estimated number of principal components, we compute the principal components and its corresponding scores. The sparse PCA is computed with the 'scikit-learn' package in Python. Moreover, 'statsmodels' package in Python was used to fit AR models. Table 7.1 summarizes the number of principal components and significant AR(1) coefficients, categorized by the type of PCA. We set the significant level to test whether the AR coefficient is statistically significant to  $\alpha = 0.05$ . The results indicate that the sparse PCA detects the largest number of significant AR(1) coefficients.

The profit, loss and the Sharpe ratio were calculated based on a simple portfolio

using hat return( $\hat{R}$ ) presented in Section 4.2.

$$\text{sharpe ratio} = \frac{R_p}{\sigma_p} \quad (7.1)$$

where  $R_p$  is the return of the portfolio and  $\sigma_p$  is a standard deviation of the portfolio's excess return. We then annualized the profit, loss and the Sharpe ratio, and considered the cases of  $K = 334$  and  $K = 519$ , using various types of PCA. The results are all summarized in Table 7.1.

Table 7.1: Basic study with various PCA methodology

Type	Classic PCA		Probabilistic PCA	Sparse PCA	
Number of PCs	519	334	334	519	334
Significant AR(1) number	1623	1681	1680	1707	1705
Profit and loss(%/Yr)	132.4	132.1	132.2	122.5	122.1
Sharpe ratio(/Yr)	6.91	7.12	7.12	6.53	6.50

Conditional optimized portfolio based on each model have different performance levels. The best performing portfolio was the one constructed using the classical PCA with  $K = 519$ . Also, the portfolio constructed using Sparse PCA and the probabilistic PCA underperformed compared to the one utilized their origin signal  $\hat{R}$ . It implies that they failed to estimate  $\Psi$  of the model in Section 6.1. Therefore, we use  $\hat{\Psi}$  estimated from the classical PCA with  $K = 519$  to optimize the conditional portfolio. These models are labeled as the 'mixed probabilistic PCA' and the 'mixed sparse PCA'.

The average portfolio performance of the classical PCA with  $K = 519$  shows 13.9 times higher Sharpe ratio and 16.5 times higher profit compared to KOSPI. A brief description of the portfolios was summarized in Tables 7.2 and 7.3. These tables show that the optimized portfolio which was generated from the systematic factor-based AR(1) model consistently has a outstanding performance compared to the KOSPI.



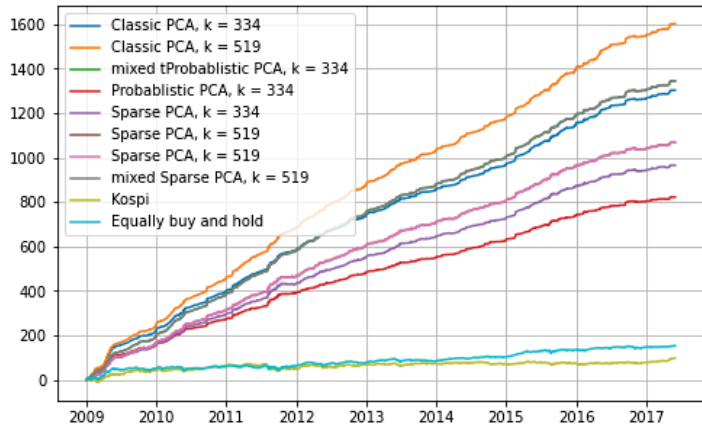


Figure 7.1: Profit and loss(%) of portfolios in training dataset from January, 2009 to May, 2017.

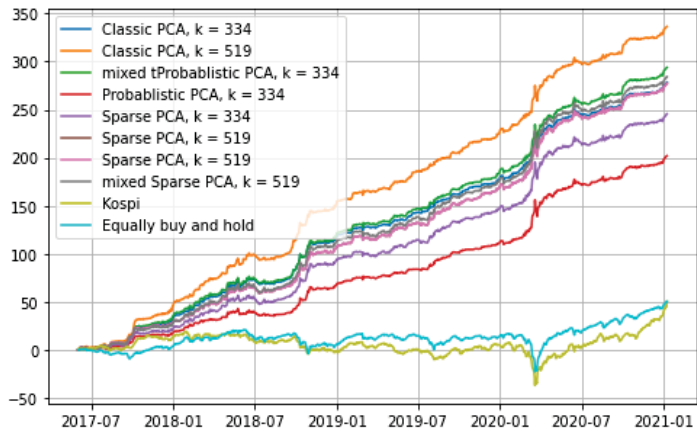


Figure 7.2: Profit and loss(%) of portfolios in Test dataset from June, 2017 to December, 2020.

Table 7.2: Summary of conditional optimized portfolio's performance in the training set.

Type	Classic PCA		Probabilistic PCA	Sparse PCA		mixed probabilistic PCA	mixed sparse PCA	KOSPI
	519	334		519	334			
Number of PCs	519	334	334	519	334	334	519	-
Sharpe ratio/(Yr)	<b>8.62</b>	8.04	6.61	7.03	6.61	7.84	7.03	0.67
Profit and loss(%/Yr)	<b>182.3</b>	148.3	132.2	121.5	109.8	153.0	121.5	11.1
Turn over(/Yr)	131.1	129.1	128.2	119.1	<b>116.6</b>	131.4	119.1	-

Table 7.3: Summary of conditional optimized portfolio's performance in the test set.

Type	Classic PCA		Probabilistic PCA	Sparse PCA		mixed probabilistic PCA	mixed sparse PCA	KOSPI
	519	334		519	334			
Number of PCs	519	334	334	519	334	334	519	-
Sharpe ratio/(Yr)	<b>4.90</b>	4.40	3.54	3.96	3.59	4.35	3.78	0.72
Profit and loss(%/Yr)	<b>89.3</b>	73.9	53.6	73.6	65.2	78.0	75.4	13.4
Turn over(/Yr)	115.8	113.5	111.5	98.9	<b>95.8</b>	113.4	96.3	-

## Chapter 8

### Concluding Remark

In this thesis, we proposed a model that generates an optimal portfolio using systematic factors based the autoregressive model, then applied it to the Korean equity market. The systematic factors are estimated from the sample covariance of returns. We concluded that the using the sample covariance yield better performance than using the factors extracted from the sample correlation. Sparse or probabilistic models were employed to estimate the proper systematic risk. With the proposed model, we derived conditional mean-variance optimized portfolio, which outperformed its origin signal  $\hat{R}$ .

One of the contributions of our model is that it provided some key ingredients to construct a portfolio with high performance, and can readily be implemented to construct more powerful portfolios. To illustrate, we can use various types of shrinkage estimators for estimating the covariance matrix of returns when the dimension of a given dataset is large, see Choi (2019). Also, factor analysis can be another enticing option to estimate systematic risk factors more effectively, see Bhattacharya (2011). In addition, time-series models can be applied upon estimating factor loadings, then these can be utilized to observe the change of sample variance in Section 2.3. Ultimately, we believe that this thesis will provide insights to future researchers and practitioners regarding modeling high-dimensional financial datasets, especially Korean stock market.

## Bibliography

- [1] W. F. Sharpe, "Capital Asset prices: A Theory of Market equilibrium under condition of risk," *Journal of Finance.*, vol. 19, no.3, pp. 425-442, Sep 1964.
- [2] J. L. Treynor, "Market Value, Time and Risk," Unpublished manuscript., "Rough Draft", pp. 95-209, Aug 1961.
- [3] J. Lintner, "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets," *The Review of Economics and Statistics.*, vol. 47, no.1, pp. 13-37, Feb 1965.
- [4] J. Mossin, "Equilibrium in a Capital Asset Market," *Econometrica.*, vol. 34, no.4, pp. 768-783, Oct 1966.
- [5] F. Fama and K. French, "Common risk factors in the returns on stocks and bonds," *Journal of Financial Economics.*, vol. 33, no.1, pp. 3-56, Feb 1993.
- [6] M. Poterba, "Mean reversion in stock prices Evidence and Implications," *Journal of Financial Economics.*, vol. 22, no.1, pp. 22-59, Aug 1987.
- [7] S. Mukherji, "Are stock returns still mean-reverting?," *Review of Financial Economics.*, vol. 20, no. 1, pp. 22-27, Jan 2011.
- [8] P. Whittle, "Hypothesis Testing in Time Series Analysis". Thesis, Uppsala University, Almqvist and Wiksell, Uppsala, 1951.

- [9] G. E. P Box and G. Jenkins, *Time series analysis: forecasting and control.*, Prentice Hall, New Jersey. 1994.
- [10] B. Mandellbrot, “The Variation of Certain Speculative Prices,” *The Journal of Business.*, vol. 36, no.4, pp. 394-419, Oct 1963.
- [11] F. Engle, “Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation,” *Econometrica.*, vol. 50, no. 4, pp. 987-1007, Jul 1982.
- [12] T. Bollerslev, “Generalized autoregressive conditional heteroskedasticity,” *Journal of Econometrics.*, vol. 31, no.3, pp. 307-327. 1986.
- [13] W. Li, S. Ling and M. McAleer, “Recent theoretical results for time series models with garch errors,” *Journal of Economic Surveys.*, Vol. 16, no. 3, pp. 245-259, Feb 2002.
- [14] H. Markowitz, “Portfolio Selection,” *The Journal of Finance.*, vol. 7, no. 1, pp. 77-91, Mar, 1952.
- [15] F. Black and R. Litterman, “Global Portfolio Optimization,” *Financial Analysts Journal.*, vol. 48, no. 5, pp. 28-43, Oct 1992.
- [16] Y. G. Choi, J. Lim and S. Choi, “High-dimensional Markowitz portfolio optimization problem: emperical comparison of covariance matrix estimators”, *Journal of statitital computation aand simulation.*, vol. 89, no. 7, pp. 1278-1300, Feb 2019.
- [17] M. Avellaneda and J. H. Lee, “Statistical arbitrage in the US equities market,” *Qunatitative Finance*, vol. 10, no. 7, pp. 761-782, Aug 2010.
- [18] F. Fama and K. French, “A five-factor asset pricing model,” *Journal of Financial Economics*, vol. 116, no. 1, pp. 1-22, April 2015.

- [19] J. Guigarro-Ordóñez, “High-dimensional Statistical Arbitrage with Factor Models and Stochastic Control,” *Applied Mathematical Finance.*, vol. 26, no.4, pp. 328-358, Dec 2019.
- [20] I. T. Jolliffe, “Principal Component Analysis and Factor Analysis”, Springer Series in Statistics. Springer, New York, 1986.
- [21] R. Tibshirani, “Regression Shrinkage and Selection via the Lasso,” *Journal of the Royal Statistical Society. Series B.*, vol. 58, no.1, pp.268-288, 1996.
- [22] H. Zou, T. Hastie and R. Tibshirani, “Sparse Principle Component Analysis,” *Journal of Computational and Graphical Statistics.*, vol 15, no.2, pp. 265-286, June 2006.
- [23] Z. Kakushadze, “Mean-Reversion and Optimization,” *Journal of Asset Management*, vol. 16, no. 1, pp. 14-40, 2015.
- [24] M. E. Tipping and C. M. Bishop, “Probabilistic Principal Component Analysis,” *Journal of the Royal Statistical Society Series B.*, vol. 61, no. 3, pp. 611-612, Jan 2002.
- [25] T. P. Minka, “Automatic choice of dimensionality for PCA,” MIT Media Laboratory, Massachusetts, US, Technical Report. No. 514. Dec. 2000.
- [26] A. Bhattacharya and D. B. Dunson, “Sparse Bayesian infinite factor models,” *Biometrika*, vol. 98, no.2, pp. 291-306, Jun 2011.
- [27] J. Sherman and W. J. Morrison, “Adjustment of an Inverse Matrix Corresponding to a Change in One Element of a Given Matrix,” *The Annals of Mathematical Statistics.*, vol. 21, no.1, pp. 124-127, Mar 1950.
- [28] M. A. Woodbury, “Inverting Modified Matrices,” Statistical Research Group, Princeton University, Princeton, Technical Report 42, 1950.

# 초 록

본 논문에서는 한국 주식시장의 모델 기반 통계적 차익거래전략을 연구하였다. 고차원 자료로부터 체계적 위험을 추정하기 위해 주성분 분석기법을 사용하였다. 금융자료의 주요 현상으로 관찰되는 평균 회귀와 분산 군집화와 같은 특성들이 추정된 체계적 위험 요소와 함께 평가되었다. 이런 평가로부터 개별위험기반에 확률 과정 모형을 적용한 기존 연구들과는 달리 체계적 위험 기반의 자기 회귀 모형이 한국 주식의 일간 수익률을 설명하기 위해 제안되었다. 제안된 모형을 기반으로 Markowitz의 평균 분산 최적화 포트폴리오를 조건부 기대 평균 분산 최적화 방식의 포트폴리오 전략으로 개선하였다. 분석에 사용된 데이터는 한국 주식시장의 2009년 1월부터 2020년 12월 자료이며, 테스트를 포함한 분석 기간 전체에서 제안된 조건부 최적화 포트폴리오의 성능이 기존의 KOSPI와 다른 단순 신호기반의 전략들의 성능을 크게 상회하는 결과를 보여주었다.

**주요어:** 통계적 차익거래기법, 자기회귀 모형, 주성분 분석, 희소 주성분 분석, 평균 분산 포트폴리오 최적화, 자산가격결정 모형, 한국 유가증권

**학번:** 2019-24588