공학박사 학위논문

New criterion on portfolio selection based on trimmed clusters of stock market

금융 시장 클러스터 분석에 기반한 포트폴리오 관리

2016 년 8 월

서울대학교 대학원

산업공학과

정 승 재

New criterion on portfolio selection based on trimmed clusters of stock market

지도 교수 장 우 진

이 논문을 공학박사 학위논문으로 제출함

2016 년 6 월

서울대학교 대학원

산업공학과

정 승 재

정승재의 공학박사 학위논문을 인준함

2016 년 6 월

위 원 장 박 용 태 (인)

부위원장 장 우 진 (인)

위 원 조 성 준 (인)

위 원 이 재 욱 (인)

위 원 이 덕 주 (인)

Abstract

Decades have passed since the financial market began to receive attention from academia. Financial Economics became a solid branch of Economics, and statistical tools and Econometrics were exhaustively employed to analyze the financial data from every possible directions. A framework proposed using mathematical models was a catalyst for expansion of the market. Analytical tools from other fields such as Signal Processing and Physics discovered phenomena ubiquitous in financial data and established the stylized facts. Though these studies did deepen the understanding of the financial market, they weren't sufficient to prevent financial crises. Institutional investors and policy makers helplessly watched the market tumbles and the academia was unable to provide a clear answer and often held responsible for crises. An apparent conclusion was that the financial market is far from being fully grasped, and further study is necessary. Recently, Network theory gained popularity as a tool to interpret the financial market structure. Analytical methods found in this field such as hierarchical tree and Minimum Spanning Tree were effective to visualize relative positions and interaction between

assets. A network analysis begins with a similarity/dissimilarity measure to represent the system of objects. Statistical correlation between assets were extensively studied and accepted as a good quantity to measure the similarity/dissimilarity of assets. Another approach to utilize a dissimilarity measure is data mining. Data mining methods are particularly useful to process large amount of data in an exploratory research. Given that the financial market is yet to be explained such approach might provide an insight which was overlooked before. Therefore, clustering analysis, one of the most well-known data mining methods, was applied to the financial market.

In this study, correlation coefficients between stocks were measured and transformed using a distance function. A well-established distance function preserves the topology of the original correlation matrix. It is a good metric to see the stocks as they were. Clustering analysis was then performed on dissimilarity matrices, which are correlation matrices transformed by a distance function. Clustering analysis is designed to put similar objects together in a cluster. Though not in a quantitative way as the clustering analysis, the investors and market participants already have a framework to group similar firms together. The categorization of firms using industrial sector such as the one given by MSCI's Global Industry Classification Standard is accepted as the standard approach to group firms

together. Investors would compare the firms in the same sector and add the most promising ones to their stock portfolio.

One of the objectives of this study is to test whether the quantitative methods agree with the traditional classification by sectors. Firms were grouped by their correlations in stock returns and the members of the cluster were individually identified by their industrial sector. When a small data set of largest 30 firms by market capitalization in Korea were used to create a dendrogram, firms in the same sector were often found next to each other on the tree which suggests they are close to each other. There were few exceptions and the overall structure of trees varies for different correlation coefficients but a large part of the data would agree with the classification by the sector. However, when a clustering analysis was performed on a larger data set of 200 firms in Korea, most clusters were made of firms from different sectors and clusters rarely had more than 75% covered by a single sector which implies even within a sector, there is no clear dominant pattern which the members of the sector follow.

A portfolio of stocks were constructed based on the clustering analysis. A hypothesis was that if the clustering analysis was able to capture the market structure, the portfolio created based on this information should outperform benchmarks such as the market index. The largest 200 firms by market capitalization in Korea were used to for the analysis, and portfolios of 10, 20, and

30 stocks were built and their performance was recorded. Stocks were chosen randomly from each clusters and the average performances of 1000 such portfolios were compared to the benchmarks. Since the stocks were chosen randomly, another benchmark, a portfolio of stocks randomly chosen from the entire data set was created. The purpose of the random portfolio is to determine whether there is a statistical difference in choosing stocks from portfolio or choosing in a completely random fashion. All clustering portfolios were able to outperform the market index but many failed to beat the random portfolio in terms of return-to-risk ratio. One of the possible explanation was found that the clustering analysis was able to identify a group of underperforming stocks and by choosing equal number of stocks from each clusters, the clustering portfolio had a relatively larger number of underperforming stocks.

For an investor, the purpose of creating a stock portfolio is not to analyze the market structure but to buy diverse stocks with varying risk profiles thereby generating positive excess returns with an acceptable level of downside risk. Therefore, a trading simulation was performed to see if the clustering portfolios can be used to serve this purpose. Correlations of stocks were estimated using the historical data before a portfolio was launched, and then the portfolios were constructed in the same manner as the previous section. The portfolios were launched after the period of correlation estimation, so no information

regarding the period of investment was incorporated in the portfolios. Although the clustering portfolios did outperform the market index, none of them were able to beat the random portfolio. A marked underperformance of clustering portfolios was detected for most of the portfolios. Detailed analyses of each portfolios and their clusters revealed that there were clusters of underperforming stocks and the portfolios had a disproportionately large number of underperforming stocks in their portfolio. By trimming down the underperforming clusters and thereby removing them from the portfolio construction step, the clustering portfolios were able to beat the random portfolio. The framework was formalized and using the US market, an extended portfolio management over 20 years were simulated. Clustering portfolios were constructed in 1990 and were managed until the end of 2015. Three rebalancing periods of 3 months, 6 months and 12 months were chosen and the assets were reallocating every rebalancing period to study the effect of rebalancing frequency. Three correlation estimation periods of 1 year, 3 years and 5 years were chosen and correlation coefficients were estimated over a given period to study how changing correlation estimation period would affect the performance of portfolio. Many clustering portfolios were unable to outperform the market index, and it was found that neither rebalancing period nor correlation estimation period had a linear relationship with the performance of portfolios. The cluster

trimming process was formalized with rules and when a cluster with the most firms with net earnings loss over a long correlation estimation period of 3 years or 5 years was removed, the average return and return-to-risk ratio was improved. The result makes sense because firms with persistent earnings loss are likely to be struggling and adding them to portfolio is likely to be detrimental for portfolio's performance. Another rule found was that when clusters with more firms with net earnings loss than firms with net earnings gain were removed, the performance of portfolios was improved significantly. The two rules were applied simultaneously and all clusters which satisfied the conditions were removed. The portfolios created without those clusters were able to outperform other clustering portfolios and the benchmark index.

The purpose of this research was to analyze the market structure using clustering analysis based on correlation coefficients and propose a framework to create a stock portfolio. It was found that the classification by sector is insufficient to create a diversified portfolio. A framework to construct a portfolio based on clustering analysis was proposed and the trimming process to remove clusters of inferior stocks was introduced.

Keywords : Correlation analysis, Clustering analysis, Portfolio management, Trading simulation
Student ID : 2012-21072

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background and Motivation

The financial market drew an enormous amount of attention from academia. The market represents a complex system and it is a good source of both time series and cross-sectional data; As a result, researchers from different fields tried out methods from their own field to the market. For example, economists used Econometric methods such as regression to analyze financial assets and successfully decomposed asset returns into several factors (Carhart, 1997; Fama & French, 1993, 2015; Fama & Macbeth, 1973; Lintner, 1965a, 1965b; Merton, 1973a; Roll, 1977; Sharpe, 1964). Mathematical models were employed to price financial instruments with uncertainty (Black, 1976; Black & Litterman, 1992; Black & Scholes, 1973; Merton, 1973b). Financial data such as stock price are recorded as a function of time and naturally form time series, so it shouldn't be surprising

that time series analyses are popular when studying financial data. It began with a simple autoregressive moving average (ARMA) model, and models which compensate for ARMA's weaknesses were developed, generalized and expanded (Bollerslev, 1986; Engle, 1982; Engle & Ng, 1993; Glosten *et al.*, 1993; D. B. Nelson, 1991). Statistical analysis of financial time series made many important discoveries such as long memory property of financial time series, volatility clustering, regime switching behavior and other evidences of non-normality (Cont, 2001; Ding *et al.*, 1993; Durland & McCurdy, 1994; Hamilton, 1989; Maheu & McCurdy, 2000; Mandelbrot, 1997). These discoveries are now known as the stylized facts of financial time series. Methods developed in Physics to study phenomena such as Multifractality and entropy were employed to study financial time series (Di Matteo, 2007; Jiang & Zhou, 2008; Jizba *et al.*, 2012; Kantelhardt *et al.*, 2002; Maasoumi & Racine, 2002; Mandelbrot, 2013; Marschinski & Kantz, 2002; Zhou, 2009; Zunino *et al.*, 2009).

Recently, network and graph theory became another popular tool to examine the structure of financial market. It began with a study of hierarchical structure of the market, where the author found that using a correlation matrix of stocks, an economic taxonomy of stocks was discovered (Mantegna, 1999). A similar idea was investigated using a high-frequency dataset of stocks (Bonanno *et al.*, 2001). Filtered networks of stocks were

estimated over different time horizons and concluded that the market was progressively structured at different time horizons (Tumminello *et al.*, 2007). The minimum spanning tree and planar maximally filtered graph were constructed with a moving window and their dynamics over time were studied (Di Matteo *et al.*, 2009). A partial correlation network was created using stocks listed in New York Stock Exchange and it was found that the financial sector plays the most influential role in the market (Kenett *et al.*, 2010). A filtered network was built and a stock portfolio consists of peripheral stocks based on the network was able to outperform the entire set of stock or a stock portfolio made of high centrality in terms of return-to-risk ratio (Pozzi *et al.*, 2013). A minimum spanning tree is studied over time and during a financial crisis, the tree shrank strongly but the optimal Markowitz portfolio was lying on the outskirts of the tree for all times (Onnela *et al.*, 2002). A network of firms was compared against a random graph to examine non-random features of the network (Onnela *et al.*, 2004). An empirically obtained minimum spanning tree of stocks was matched against a simulated market and features that cannot be reproduced by a random market model were observed (Bonanno *et al.*, 2003). A Granger-causality network was constructed using firms in the financial sector, and the authors found that sub-sectors of the financial sector became highly interrelated and using this information they

were able to quantify the financial crisis periods (Billio *et al.*, 2012).

However, even after vast amount of research, the market remained elusive and recent events such as 2007–2008 financial crisis and 2011 European sovereign crisis proved that it is yet to be fully understood and different approach is required to shed a light.

With the rise of Big Data, the data mining methods found many applications in engineering, business and finance. Unlike statistical methods, data mining usually does not require conditions such as a random variable being normally distributed. However, it does require a researcher to determine input variables such as number of cluster. The unsupervised methods such as clustering are data-driven in nature so it is ideal for studying a dataset where a model-based approach or approach based on conventional wisdom had a challenge. Given the sheer amount of study already performed on stocks and the financial market, one would expect that the market could be explained but almost every decade, an event or two occurs with seemingly no sign of warning and ask for another set of explanation. One could deduce that the traditional methods may have overlooked certain properties and it is critical to reveal them to better understand the market. An approach which does not take any assumption or domain knowledge and starts from the scratch may prove useful in this situation.

Another aspect to consider when it comes to Finance is that an idea or a model developed in academia may or may not be applicable to the real world. If a model requires too many assumptions which deviate from the real world, it is likely that practitioners would reject the model. This has been the case for many models developed before. Though some of them were able to provide at least a framework to approach a problem, many studies were irrelevant to practitioners. Since the market crash and financial turbulence are not a theory but events observed in the real world, a financial study with a real world application in mind is needed.

1.2 Research Objectives

The primary objective of this research is to analyze the stock market and study its structure. Though similar studies were performed before, they were insufficient to fully explain the market, and market participants could rarely make use of them. In this study, a different approach is taken and the previous results were revisited. The study was focused on empirical analysis using stock price data, and making assumptions was avoided as much as possible. Another important goal of this research is to develop an algorithm or framework which can be directly applied to the real world. The specific goals of this research are as follows.

The first goal is to study the firms in the market using dissimilarity measures. Stock returns were computed from the stock price data and correlation between all pairs of stocks were estimated. The estimated coefficients were analyzed and through a filtering procedure, a noise component was removed from the correlation matrix. A distance function was adopted and dissimilarity matrices were built based on the filtered correlation matrix. The dissimilarity matrices were then used to build hierarchical trees to visually study the proximity between firms. The purpose of this step is to investigate whether the different

dissimilarity metrics were unanimous in explaining the market structure.

The second step is the clustering analysis of the dissimilarity matrix. Clustering analysis is an unsupervised, data-driven method which does not require any prior information, and the result is purely based on the input data. The main question was whether the data-driven structure agrees with the conventional wisdom. A hypothesis was that since the firms in the same sector are exposed to the same environment and risk, they are likely to be clustered together. For example, oil companies are inevitably linked to the price of crude oil and overall demand of the market, so their stock movements must be correlated to those conditions, which lead to a high-correlation between the oil companies.

When the market structure is revealed, a portfolio can be built. The idea of portfolio selection is that by choosing the firms that are least related to each other, one can maximize the expected return for a given level of risk. A strength of clustering analysis is that one can naturally pick stocks with smaller correlation by choosing stocks in different clusters. Also, when clustering analysis was performed based on stock returns, then the firms with poor performance could be clustered together or form a small cluster of outliers which can be removed from portfolio selection. Therefore, if the market structure revealed was true, then the portfolio built based on the information will outperform a benchmark such as the market index. This step was to confirm

whether dissimilarity matrices and clustering analysis were able to capture the market structure; hence, the analysis was done in an *ex-post* manner.

The next step is a trading simulation. Using only the past information available at the time, a portfolio was constructed and passively managed for a year. The performance of the portfolio which is measured by annual return-to-risk ratio was compared to the benchmarks. If the clustering analysis was able to separate the firms in terms of their fundamental differences, then the differences would remain for some time. A portfolio of fundamentally different firms should have a superior return-to-risk profile. The main objective of the research is to study if it is possible to achieve excess return using only the analysis of historical data.

As the final step, portfolio management over a long period of time was performed. The previous step provided a detailed view of portfolio management over a year. In this section, based on the findings from previous section, strategies for portfolio management were developed and a portfolio was constructed and managed for more than 20 years with periodic rebalancing of portfolios. The purpose of this study is to provide an empirically tested framework to build a stock portfolio.

1.3 Organization of the Research

The dissertation is organized in five chapters.

Chapter 1 introduces the background, motivation and research objectives of the research.

Chapter 2 reviews the literature related to the methods used in this study, mainly correlation analysis, clustering analysis, and portfolio selection.

Chapter 3 examines the stock returns using correlation analysis. A filtering procedure based on Random Matrix Theory was performed and filtered correlation matrices were used to build dissimilarity matrices. Then, hierarchical trees were constructed based on the dissimilarity matrices.

Chapter 4 investigates the market structure. Clustering analysis was performed on the dissimilarity matrix to group similar stocks together. A portfolio analysis was followed and trading simulation was implemented to provide a framework for practical purpose.

Chapter 5 concludes the research with summary, implications, and future work.

Chapter 2

Literature Review

2.1 Financial Time Series and Correlation

A vast amount of financial data is available in time series such as daily closing price of stocks or commodities. It is obvious that tools for time series analysis can be used for studying financial time series as well and many important discoveries were made this way. For example, the distribution of financial return series have heavier tail than the Gaussian distribution (Carr *et al.*, 2002; Cont, 2001; Eraker *et al.*, 2003; Gabaix *et al.*, 2003; Gopikrishnan *et al.*, 1999; Longin & Solnik, 1995; Mantegna & Stanley, 1995; D. B. Nelson, 1991; Plerou, Gopikrishnan, Nunes Amaral, *et al.*, 1999; Skjeltorp, 2000; Stanley *et al.*, 2000). Figure 1 shows the histogram of daily log returns of Korea Composite Stock Price Index (KOSPI) from 2003 to 2015 with a Gaussian curve fit to the histogram. A log return $r_{i,t}$ is computed by

$$r_{i,t} = \log(S_t) - \log(S_{t-\Delta t}) \qquad (2-1)$$

where S_t is the daily closing price of KOSPI at time t, and Δt is the time interval which is one day throughout this research. It is visually clear that the histogram has heavier tail than the Gaussian distribution, which is supported by its sample kurtosis value of 9.6834.

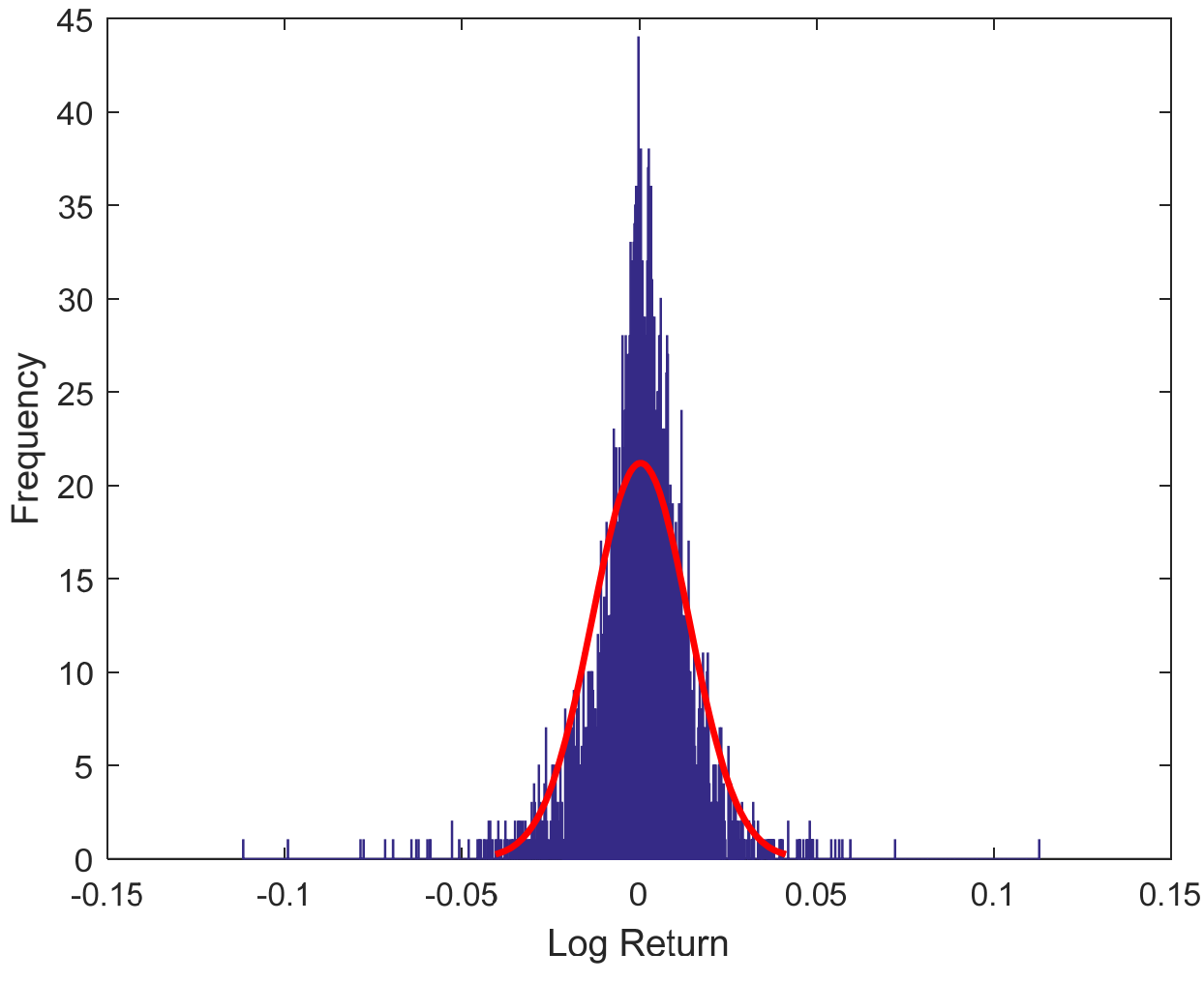

Figure 1: Histogram of KOSPI returns

The autocorrelation of financial time series was extensively studied and it was found that though the log return series had no autocorrelation, the absolute value of returns did have autocorrelation and multiple attempts were made to model this property (Baillie, 1996; Barkoulas *et al.*, 2000; Bollerslev & Mikkelsen, 1996; Cheung & Lai, 1995; Ding *et al.*, 1993; Granger

& Hyung, 2004; Henry, 2002; Lobato & Savin, 1998). The autocorrelation functions of log returns and the absolute value of log returns are plotted in Figure 2. The autocorrelation is computed by

$$\rho_{\Delta t} = \frac{\langle (r_i - \langle r_i \rangle)(r_{i-\Delta t} - \langle r_{i-\Delta t} \rangle) \rangle}{\sqrt{\sigma_i^2 \cdot \sigma_{i-\Delta t}^2}} \qquad (2-2)$$

where $\langle r_i \rangle$ represents the mean value of random variable r_i over a given period, and σ_i denotes the standard deviation of r_i. Notice that the absolute value of returns has persistent autocorrelation over long lags.

Another stylized facts of finance is that a large fluctuation of price tends to be followed by large changes, a well-known phenomenon called volatility clustering (Bentes *et al.*, 2008; Campbell *et al.*, 1998; Jacobsen & Dannenburg, 2003; Tseng & Li, 2011). Figure 3 plots the KOSPI return, and it is clear that some parts of return series have a greater variance for a short period of time then other parts. All these stylized facts support the idea that an asset returns do not follow the Gaussian distribution.

When two financial time series are considered together, their correlation is an important subject for examination. It was found that the correlation between assets or an asset and the market is

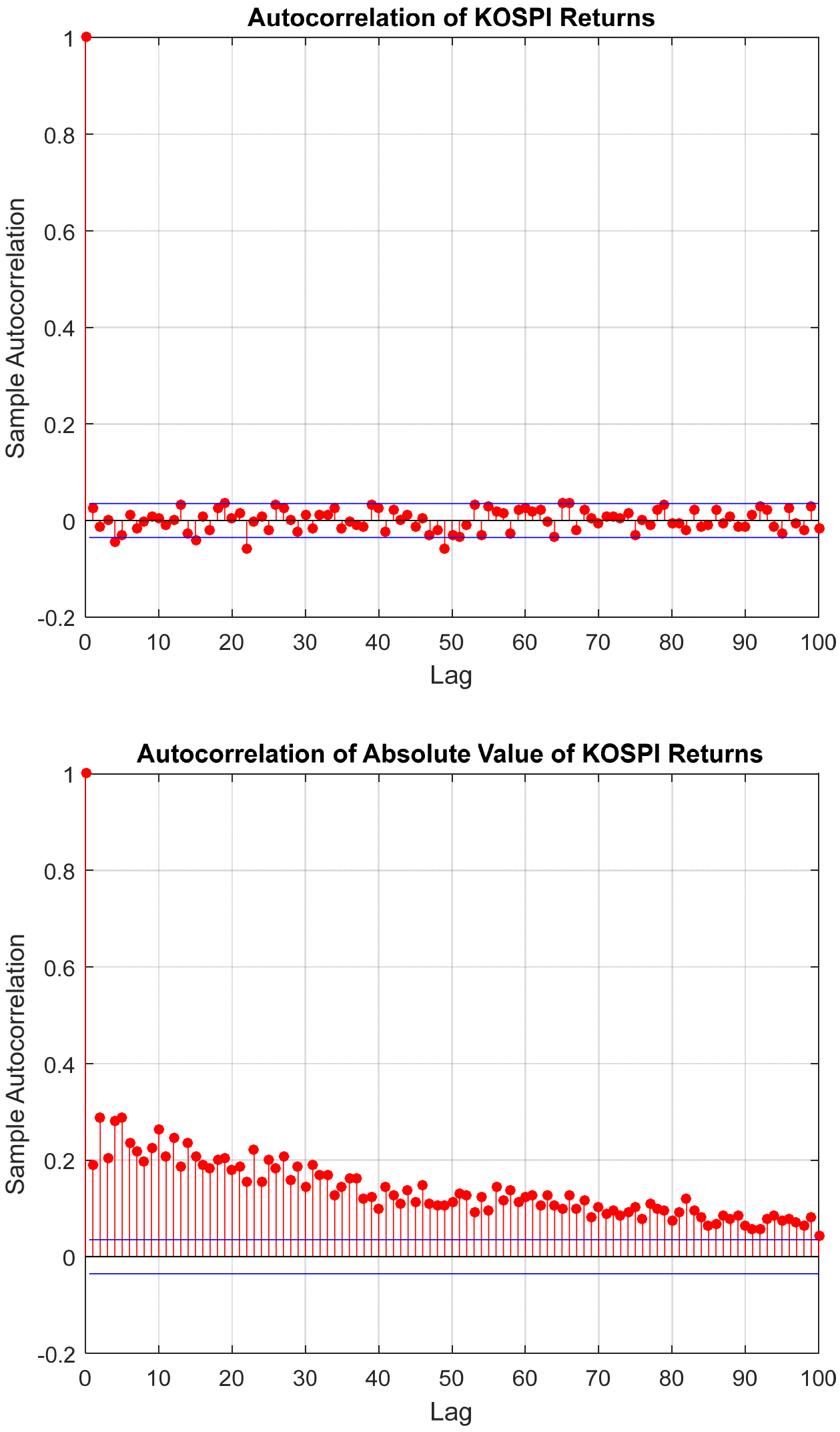

Figure 2: Sample autocorrelations of KOSPI returns

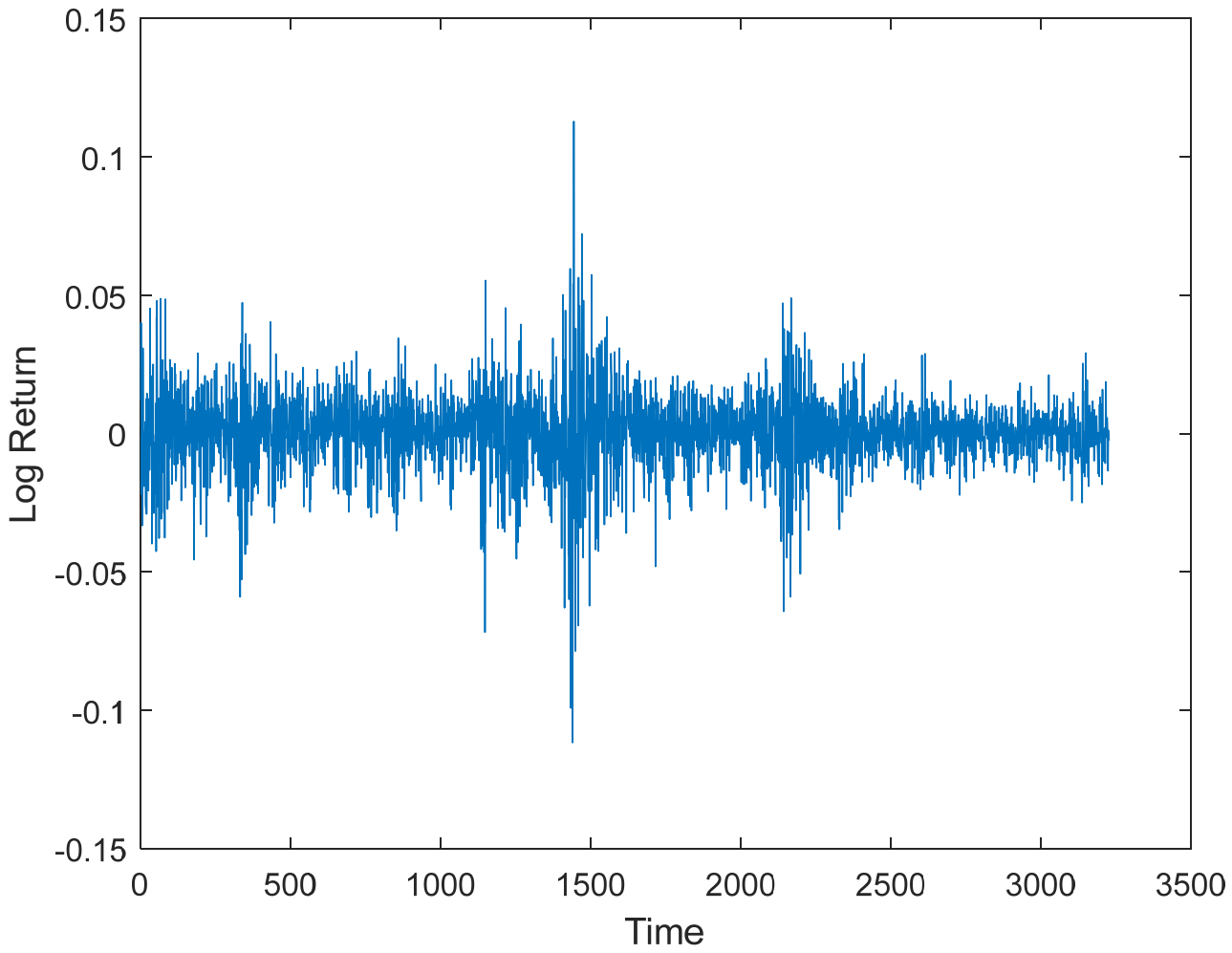

Figure 3: Return series of KOSPI from 2003 to 2015

time-varying and becomes larger during a financial crisis (Ang & Chen, 2002; Caporale *et al.*, 2005; Longin & Solnik, 1995; Tseng & Li, 2011). Similar phenomena were observed that the stock markets around the world have higher correlations when the markets are volatile (Arestis *et al.*, 2005; Bertero & Mayer, 1990; Chiang *et al.*, 2007; Eun & Shim, 1989; Ramchand & Susmel, 1998). Figure 4 shows the correlation coefficients between KOSPI returns and Samsung Electronic returns estimated using moving window with window size of 252 days from 2003 to 2015. During this period, the correlation coefficients varied from 0.8892 to −0.1660.

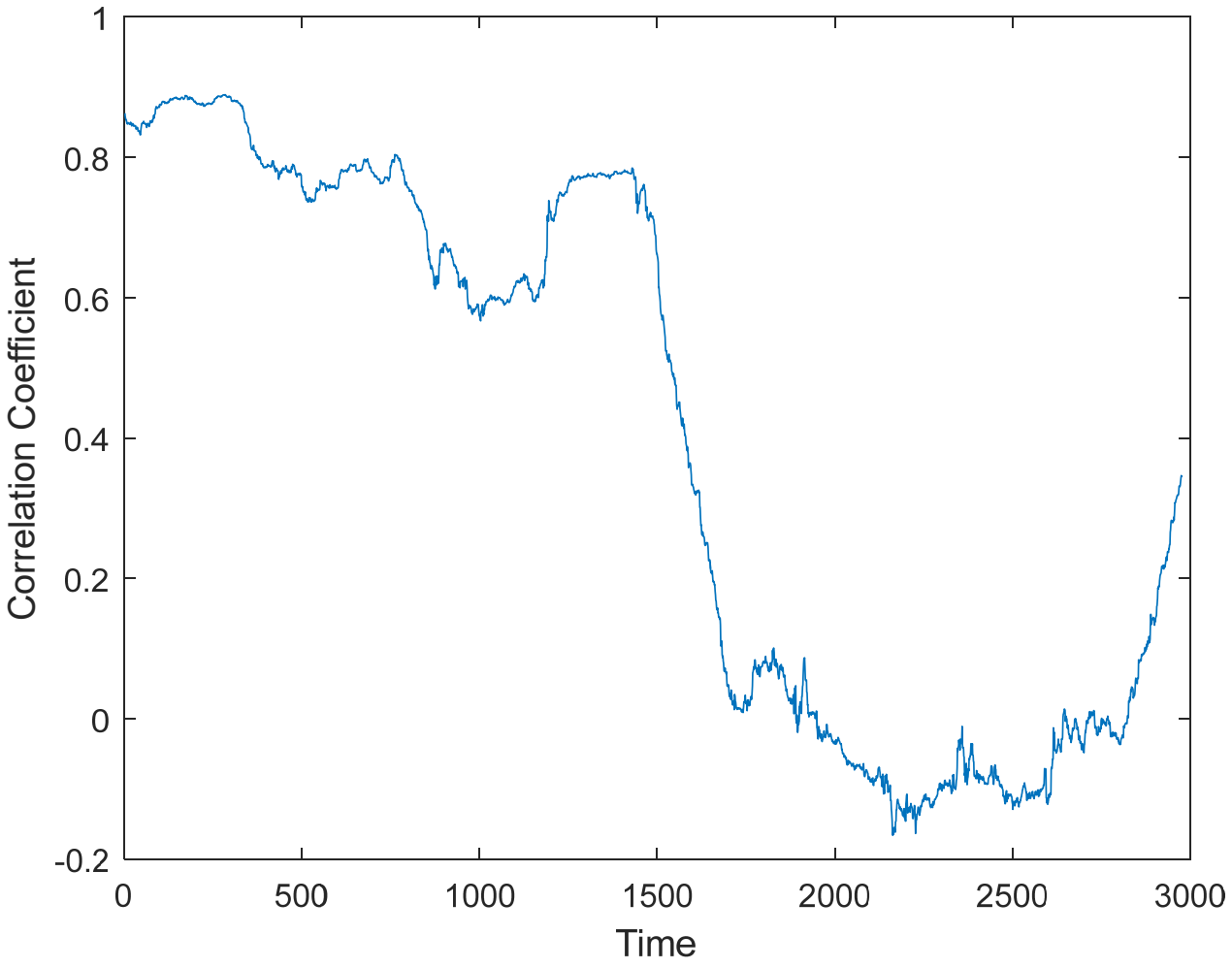

Figure 4: Time-varying correlation between KOSPI returns and Samsung Electronics' returns

When a large number of stocks are considered together, it is difficult to visualize the correlation between those stocks because the associated correlation matrix naturally becomes an N-dimensional space where N is the number of stocks which could easily be well over a hundred. Therefore, a different approach is often used to study this kind of dataset. Network Theory is a popular choice to visually study the multiple assets together. The minimum spanning tree and its extensions are capable of reducing the dimension a correlation matrix into two and visually plot the matrix of stocks (Coelho *et al.*, 2007; Micciche *et al.*, 2003; Onnela, Chakraborti, Kaski, & Kertesz, 2003). The network of correlation matrix was used to investigate

other aspects of Financial Economics such as systemic risk and hierarchical structure of the market. (Billio *et al.*, 2012; Bonanno *et al.*, 2004; Tumminello *et al.*, 2010) Figure 5 plots the minimum spanning tree of 200 largest firms in KOSPI during the year 2015. Though it is difficult to read, the plot provides a coarse map of how firms are connected to each other.

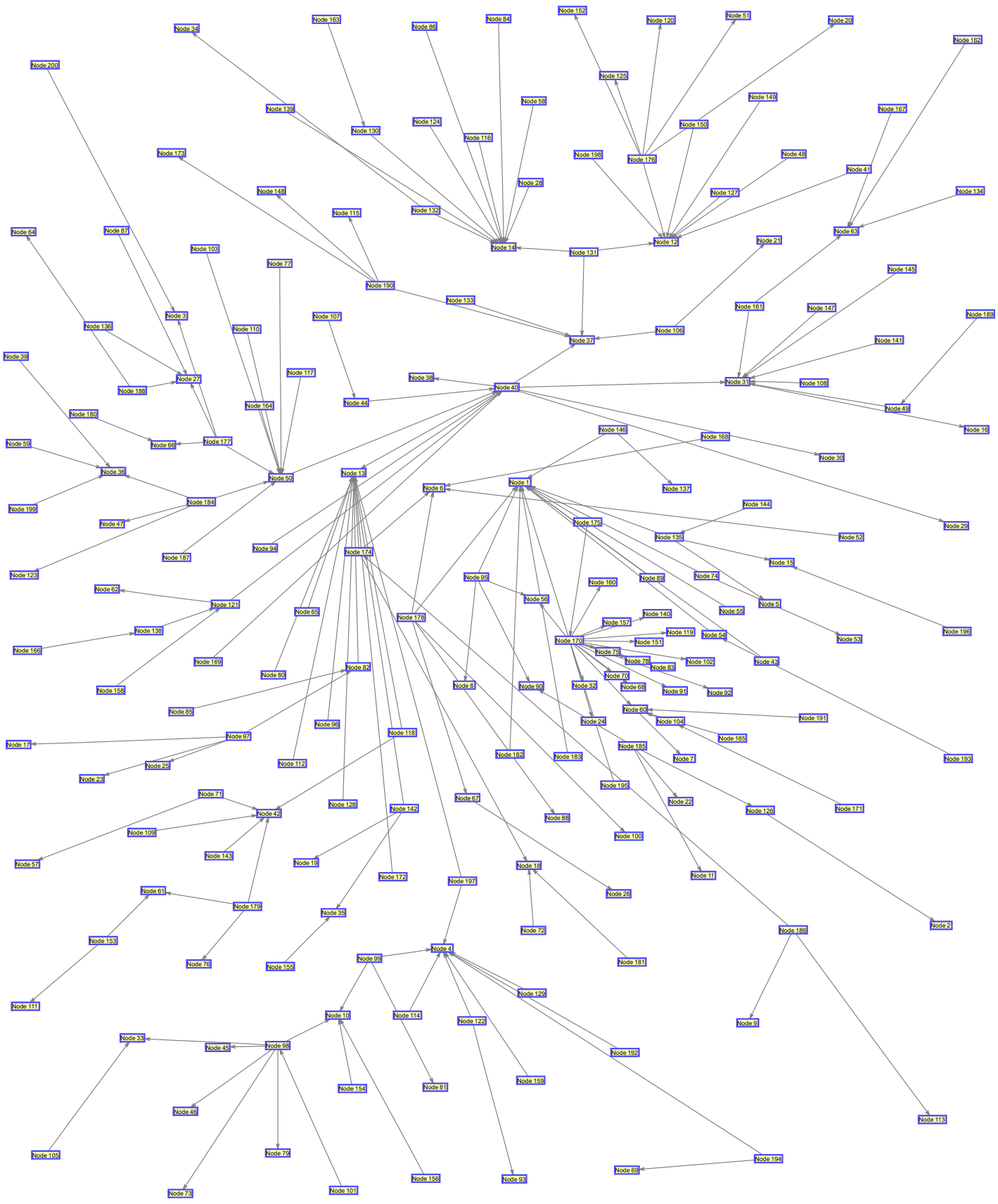

Figure 5: Minimum spanning tree of largest 200 firms in KOSPI

2.2 Clustering Financial Time Series

Clustering is a process of grouping a large number of objects, often infeasible to be done manually by human experts, into a number of groups or clusters. The objects assigned to the same cluster have high similarity, and they are dissimilar to objects in other clusters. Clustering analysis has many application areas such as biology, business, web search and so on (Han *et al.*, 2011).

Clustering is known as an unsupervised method because the label for clusters are not predetermined. The researcher is assuming that no prior or domain knowledge is available and only choose input variables and a distance measure. It was proved that Clustering analysis was useful to study a high-dimensional correlation matrix found in the financial market. It was applied to different data sets with different goals in mind and successfully delivered some results. A fixed income market, more specifically the US Treasury bonds, was studied and it was found that the correlations between fixed income securities were highly dependent on the assets' maturities (Bernaschi *et al.*, 2002). A mutual fund database was analyzed and the clustering analysis was used to classify the mutual fund's style and concluded that the method was able to group the similar funds together in accordance with the existing institutional classification. A dissimilarity measure based on tail dependence was studied for

clustering and this approach was useful to construct a portfolio which is more stable during a financial crisis (De Luca & Zuccolotto, 2011, 2014). A clustering analysis based on variance ratio test statistics was performed and when used for international stock market, it was able to distinguish different markets by their size and the level of development (Bastos & Caiado, 2014). Another study was focused on reliability of the portfolio in terms of predicted risk and realized risk and showed that the use of clustering methods can improve the reliability (Tola *et al.*, 2008).

A few options are available for which method of clustering to choose. K-means is one of the most popular clustering methods. As it is obvious from its name, K-means algorithm partitions a data set into k clusters. The objects within a cluster is close to the mean, or the centroid of a cluster; hence, it is known as a centroid-based partitioning algorithm.

In this study, however, a hierarchical clustering method was chosen. Agglomerative hierarchical clustering method was implemented to cluster the financial data. The hierarchical clustering method uses a bottom-up strategy where it begins with objects to cluster and merge two objects with closest proximity to be the first cluster. The process is iterative, and objects and clusters are merged together and become larger in size but fewer in number and eventually, all objects are clustered into a single cluster which is equivalent to the entire data set.

The researcher is free to choose a cutoff level to pick a desired number of clusters or a level of proximity.

The hierarchical method is chosen because it begins with building a dendrogram which is extensively studied in applications of correlation matrix using Network Theory (Mantegna, 1999; Onnela *et al.*, 2002; Onnela *et al.*, 2004; Tumminello *et al.*, 2010). These studies proposed that there is a hierarchical structure in the market. Figure 6 shows the dendrogram of the 30 largest firms in KOSPI. Firms can be separated into a different number of groups by drawing a horizontal line on a dendrogram which is equivalent to the cutoff level and putting firms below each 'branch' into the same cluster.

There are a number of ways to determine distance between clusters. In this research, the average distance $\mathbf{dist_{avg}}$ is used to compute the distance between clusters which is given by

$$\mathbf{dist_{avg}}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{o_i \in C_i, o_j \in C_j} |o_i - o_j| \qquad (2-3)$$

where C_i refers to cluster i and o_i denotes to the objects in the cluster i. n_i is the number of objects in cluster C_i, and $|o_i - o_j|$ is the Euclidean distance between o_i and o_j. Other distance measures include minimum distance, which measures the distance between two objects in each clusters with minimum

distance, and maximum distance, the distance between two objects which are the farthest away from each other.

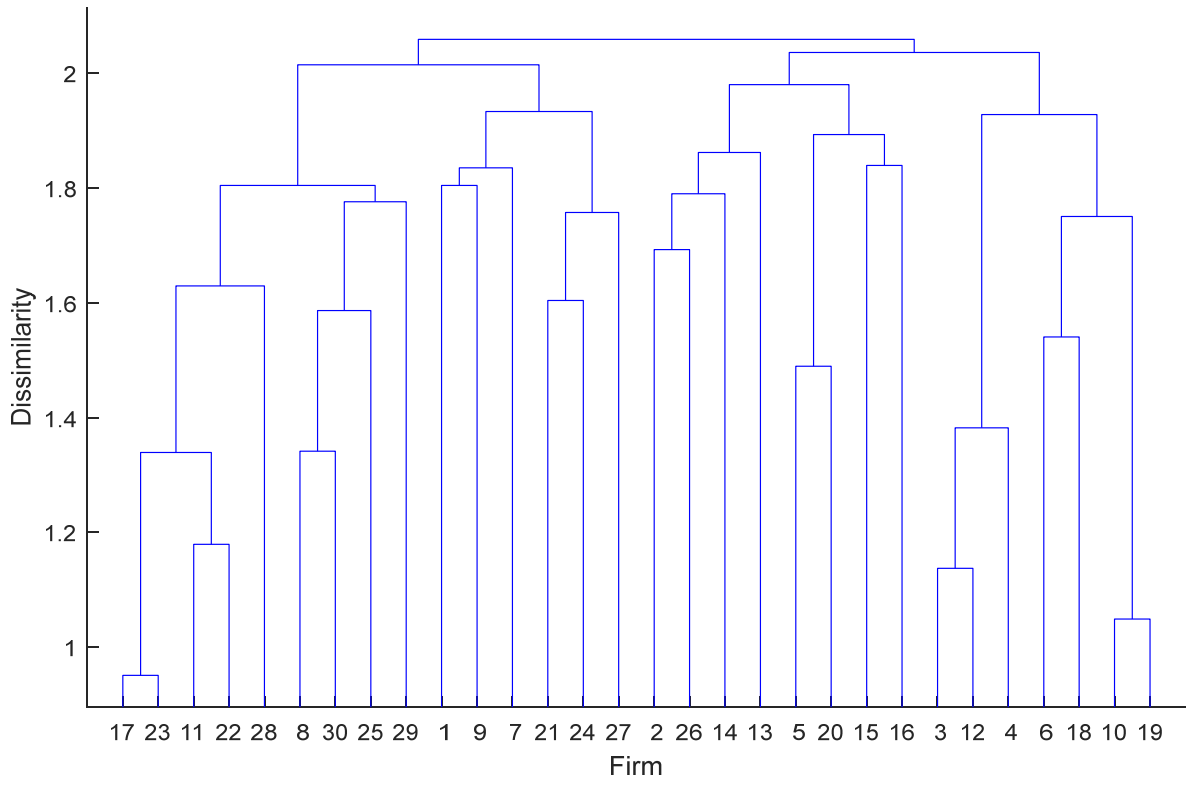

Figure 6: Dendrogram of largest 30 KOSPI firms in 2015

The next step is to determine a cut off level. Choosing a cut off level for a hierarchical clustering is equivalent to choosing a number of clusters. It is a critical step because it can dramatically change the quality of clusters. When the ground truth or domain knowledge is available, one can use this information to determine a number of cluster to use and use the knowledge to evaluate the clusters. However, much of the ground truth in the financial market is questionable at best. Therefore, an intrinsic method, which only uses the data set available to evaluate how well the clusters are separated, is more appropriate for the financial data. The cutoff level is determined by the Silhouette Coefficient (Han *et al.*, 2011). Silhouette coefficient measures how similar an object is to the other objects in the same cluster compared to the

objects in the neighboring cluster. The coefficient $s(i)$ is given by

$$s(i) = \frac{b(o_i) - a(o_i)}{\max[a(o_i), b(o_i)]} \qquad (2-4)$$

where $b(o_i)$ is the lowest average dissimilarity of object o_i to the closest cluster aside from its own cluster, while $a(o_i)$ is the average dissimilarity of o_i with the other members of the cluster it belongs to. A dissimilarity between objects was computed using correlation coefficients. From the definition it is clear that $s(i)$ can have any value between -1 and 1, and the higher value of $s(i)$ means that the clusters are compact.

2.3 Modern Portfolio Theory and Portfolio Analysis

Harry Markowitz developed a theory of portfolio selection to construct an optimal portfolio in terms of mean-variance optimization (Markowitz, 1952). The model seeks to minimize a portfolio's variance for a given value of expected return. In terms of optimization problem, it can be formulated by

$$\min \frac{1}{2} w^{'} \sum w$$
$$\text{s.t.} E(R_p) = \mu \ and \sum_i w_i = 1 \qquad (2-5)$$

where w is the weight vector of assets, $\sum$ is the variance-covariance matrix of the assets, $E(R_p)$ is the expected return of the portfolio.

Since its introduction, the Markowitz model drew attention from academia and several attempts were made to expand the theory (Brennan *et al.*, 1997; Fernholz & Shay, 1982). An empirical study of Markowitz's model was performed in an *ex-post* manner and showed that for two different portfolio of assets, an optimal portfolio outperformed a benchmark once but didn't show any statistically significant improvement for the other one (Jorion, 1992). Another study suggests that when skewness is

incorporated, the construction of optimal portfolio changes dramatically (Chunhachinda *et al.*, 1997). The mean-absolute deviation portfolio optimization model solved the problem of quadratic function and generated a portfolio similar to that of the Markowitz's model (Konno *et al.*, 1993; Konno & Yamazaki, 1991). A portfolio based on GARCH model which accounts for heteroscedasticity in variance and covariance performed better than a benchmark (Pojarliev & Polasek, 2003). The number of stocks required to construct a diversified portfolio is also studied under the framework of Markowitz's model and Capital Asset Pricing Model (Statman, 1987).

Though the model is theoretically useful, it is challenging to implement it due to several difficulties. For example, the variance-covariance matrix has to be estimated and it usually suffers from the curse of dimensionality where the estimated matrix is no different from a random matrix. The problem can be mitigated using filtering methods but it still requires a large amount of observations which are applicable only for the firms with long history. Also, as the number of asset grows, the quadratic function becomes computationally difficult to solve. Even after numerous improvements were made and established as the Modern Portfolio Theory, the model is still not widely accepted by practitioners as it still suffers from drawbacks such as sensitivity to the input parameters (Best & Grauer, 1991). Most importantly, the future return of an asset is unknown and

so is the portfolio's future return; one may be able to determine its expected value, but it is very unlikely the realized value would meet the predicted value.

Other attempts were made to build a portfolio of stocks. As it was described in the previous section, the network approach proved to be useful to study the correlation matrix of stocks and it was also applied to construct a portfolio. For example, dynamic asset trees were constructed for portfolio analysis and it was found that the assets for optimal portfolio were located at the outskirts of the tree for all times (Onnela *et al.*, 2002; Onnela, Chakraborti, Kaski, Kertesz*, et al.*, 2003). A planar maximally filtered graph was constructed using a correlation matrix and the centrality measures and peripherality of the firms were computed. It turned out a portfolio made of firms with high peripherality, or the ones located at peripheries of a network, performed better than a portfolio made of firms with high centrality (Pozzi *et al.*, 2013). Clustering analysis was also used to build optimal portfolios. K means, Fuzzy C-means, and self-organizing map were used to cluster stocks in Indian market, and a Markowitz portfolio was built (Nanda *et al.*, 2010). A new similarity measure was developed to study correlation in stock and used to cluster stocks and create an optimal portfolio (He-Shan & Qing-Shan, 2007). Most of these studies were conducted using historical data and prediction or out-of-sample testing was conducted only by a few (Pozzi *et al.*, 2013). Also, many

used a data set of firms that were continuously traded over the period of study, which already prove that the firms didn't bankrupt or delisted in a given period, which means the analyses suffered from the survivorship bias.

Chapter 3

Dissimilarity Metrics

3.1 Correlation Analysis of Financial Time Series

Correlation analysis is a study of statistical relationship between two random variables. As it was discussed in Chapter 2, correlation analysis is widely used in Quantitative Finance to study phenomena such as long-range dependence of asset returns and contagion of international market. Correlation analysis is especially important in the stock market analysis because most stocks are somehow connected to each other and shows varying degrees of correlation.

Most of the correlation coefficients used in previous literature were the Pearson correlation coefficient. It is one of the simplest form of correlation coefficients which measures linear correlation between two random variable and is given by

$$\rho\left(\mathrm{r}_i, \mathrm{r}_j\right) = \frac{\langle (r_i - \langle r_i \rangle)(r_j - \langle r_j \rangle) \rangle}{\sigma_i \cdot \sigma_j} \qquad (3-1)$$

where $\langle r_i \rangle$ represents the mean value of random variable $\mathrm{r_i}$ over a given period, and σ_i denotes the standard deviation of it. The random variable r_i is the log return series throughout this paper. In this chapter of the study, the 30 largest firm by market capitalization in KOSPI were used for analysis. The daily returns during the year 2015 are used to compute means, standard deviations and correlation coefficients. Figure 7 shows the Pearson correlation matrix of the firms. The correlation values are color coded for visualization in which red represents a positive correlation coefficient, while blue represents a negative correlation coefficient. The white block is for the correlation coefficient value of zero. It is clear that all firms have correlation with at least one other firm.

Though the Pearson correlation analysis is useful for simple analysis, it has a few weaknesses. One of them is the fact that it cannot deal with common factors affecting the two random variables of interest. For the financial market, the empirical data collected cannot have controlled environment; therefore, the data can be acquired from different time period and different geological location, but it is impossible to control every external factors, some of which are random variable themselves and have influence on the two random variables of interest. This could prove to be a problem especially in an empirical study of Finance.

For example, Figure 7 suggests that many firms in Korea have positive correlations with each other, which means they grow together or shrink together. However, it is impossible to tell if it is because of their individual performance or the entire market's growing and shrinking.

When it comes to analysis stock market, it is more important to better understand the individual performance of each stock rather than the performance stemmed from growing or shrinking market. Hence, in this paper an alternative coefficient was employed. Partial correlation analysis is designed to overcome this weakness of common factor. A partial correlation coefficient is a correlation between two random variable 'I' and 'j' when a common factor 'm' is removed from both of them. It can be seen as a correlation between residuals of 'i' and 'j'. In terms of equation, it is given by

$$\rho\left(\mathrm{r}_i, \mathrm{r}_j : r_m\right) = \frac{\rho\left(\mathrm{r}_i, \mathrm{r}_j\right) - \rho\left(\mathrm{r}_i, \mathrm{r}_m\right) \cdot \rho\left(\mathrm{r}_j, \mathrm{r}_m\right)}{\sqrt{\left(1 - \rho^{2}(\mathrm{r}_i, \mathrm{r}_m)\right) \cdot \left(1 - \rho^{2}\left(\mathrm{r}_j, \mathrm{r}_m\right)\right)}} \qquad (3-2)$$

Figure 8 shows the partial correlation matrix of the 30 largest firms in Korea, constructed in a similar manner to that of Figure 7 but using partial correlation coefficients when the common factor of the KOSPI index which serves as a proxy for the entire market was removed and it has notably more blue blocks than the one made of the Pearson correlation coefficients.

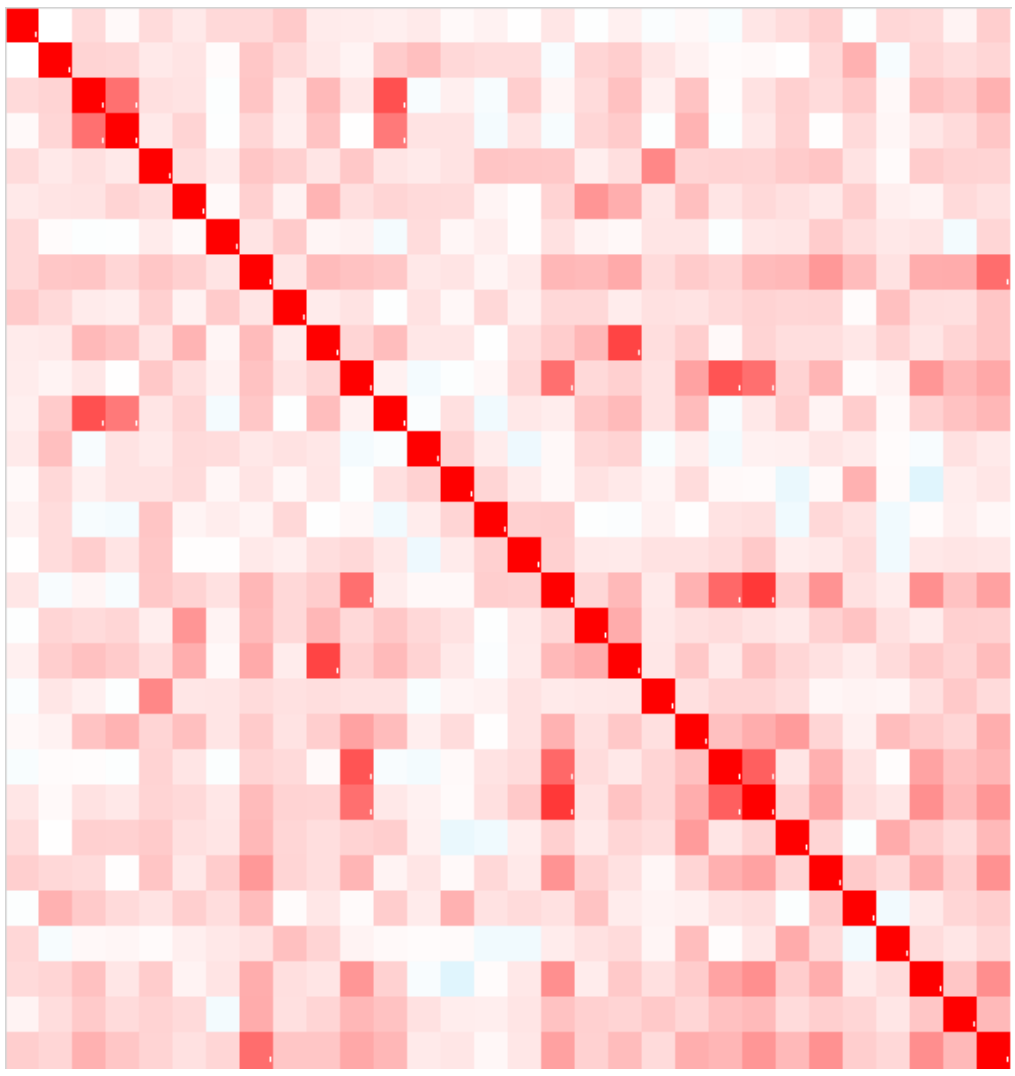

Figure 7: Pearson correlation matrix of largest 30 firms in Korea

The difference between figure 7 and figure 8 is prominent, and one could wonder where it is coming from. The histogram of correlation between the KOSPI index and its constituents were plotted in Figure 9. As one can see, most of the correlation coefficients between the index and firms were positive. Therefore, one can conjecture that even if two stocks had no real correlation between them, their respective positive correlation with the market could give them an indirect correlation.

Though not as much as the Pearson correlation analysis, the partial correlation analysis began to play its role and successfully demonstrated its use in financial studies (Kenett *et al.*, 2015).

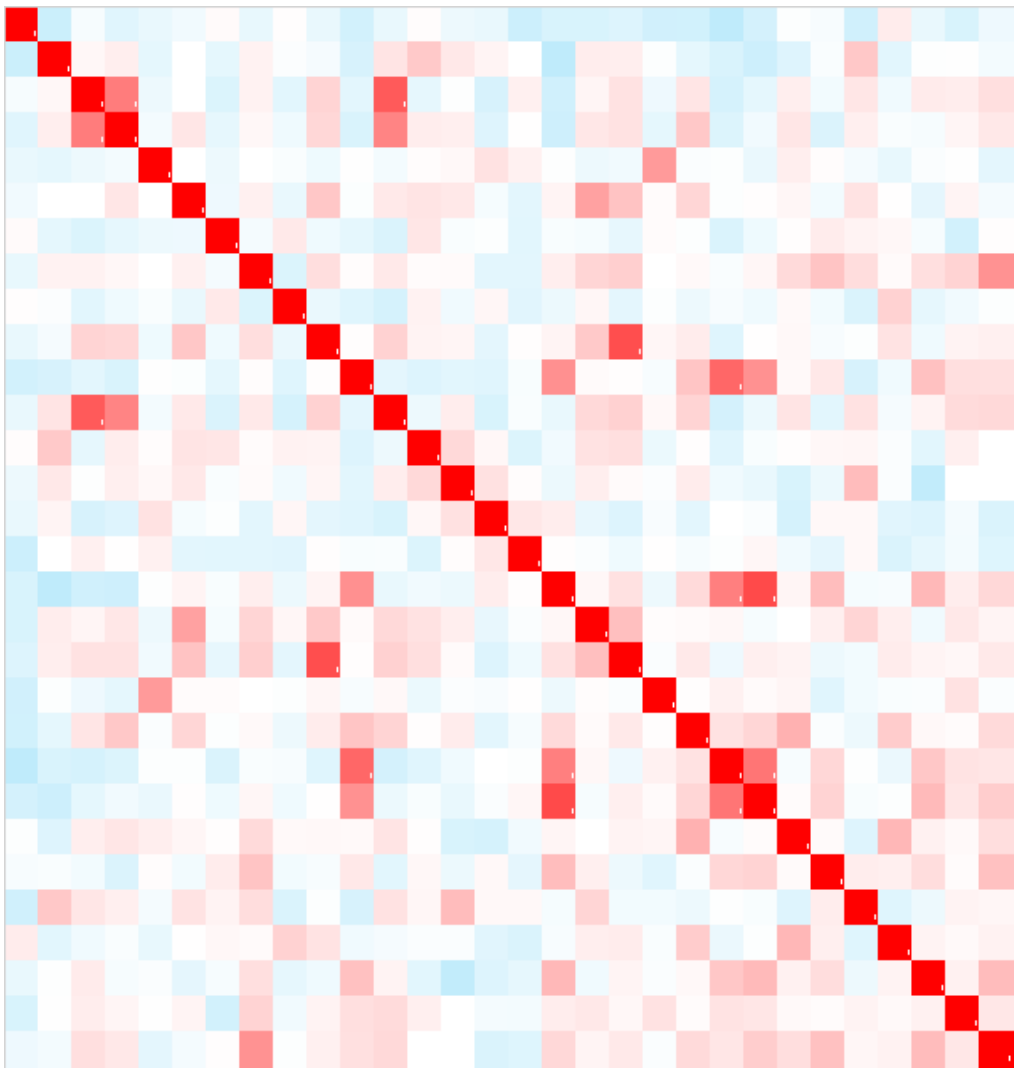

Figure 8: Partial correlation matrix of largest 30 firms in Korea

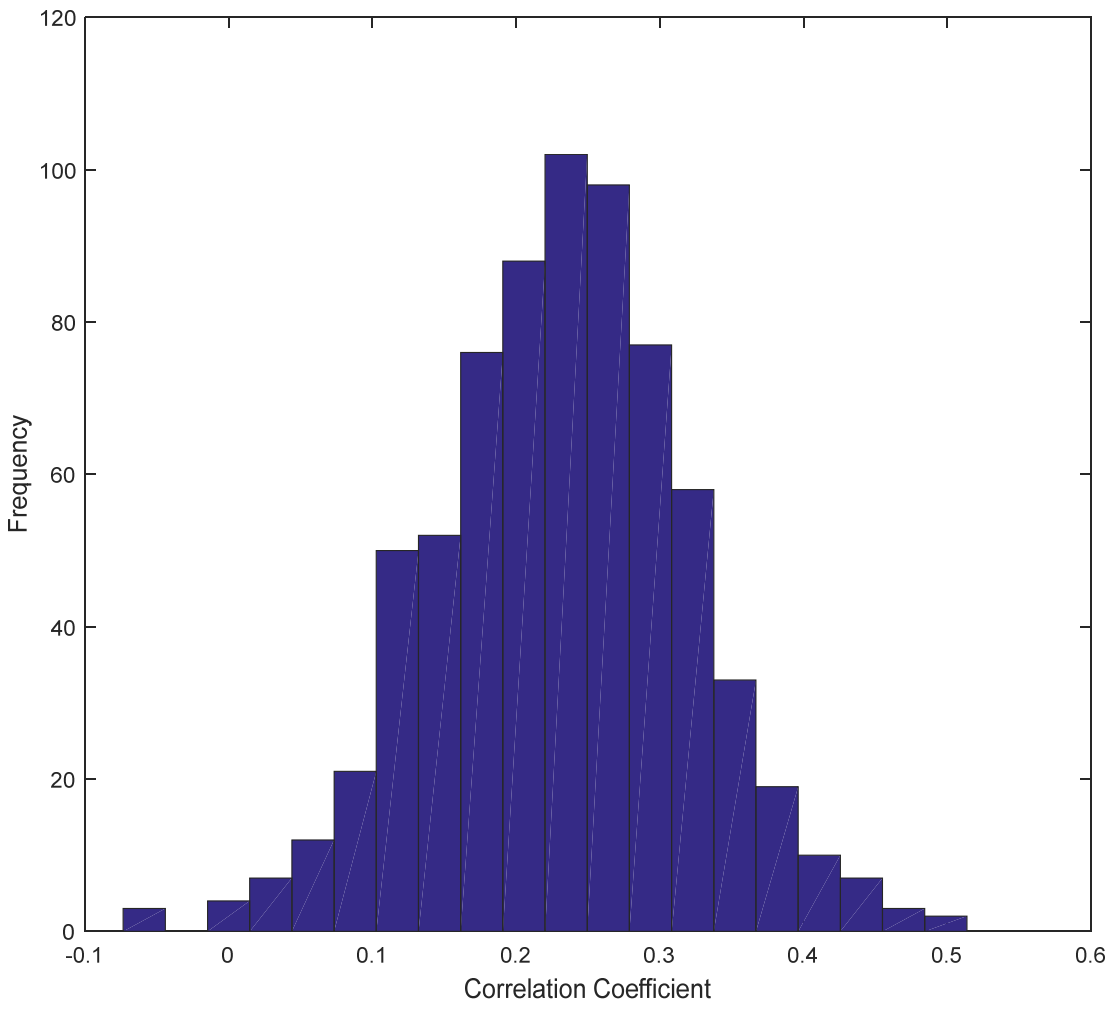

Figure 9: Histogram of correlation coefficients between KOSPI and all 1089 listed Firm in 2015

An interesting research was performed to study the effect of index using the Pearson correlation and the partial correlation (Shapira *et al.*, 2009). The authors verified that the index functioned as a cohesive force between stocks and correlations between stocks were largely because of the correlation between each stock and the index. The index cohesive effect was applied to identify a dominating sector within a market (Kenett *et al.*, 2011; Kenett *et al.*, 2010). The network approach also adopted the partial correlation analysis to construct node-node correlation matrices and showed that an insignificant node can be removed without disrupting the network (Kenett *et al.*, 2012). Another study proposed Sector Dominance Ratio using Pearson and partial correlations and empirically showed that the financial sector exhibit strong dominance for US and UK markets (Uechi *et al.*, 2015).

Note that when the market index is chosen to be the common factor, the returns used to compute the partial correlation is the residuals from the Capital Asset Pricing Model (Lintner, 1965a, 1965b; Sharpe, 1964). If the CAPM did hold, and the firm's returns can be explained by the market factor and its coefficient, Beta, then the residuals should be a series of white noise with zero mean, along with the correlation between them. However, as it is visually clear from figure 8, the correlation coefficients were mostly non-zero which shouldn't be surprising as the CAPM doesn't perform well with empirical data.

One of the well-known stylized facts of the financial time series is that it has heavier tail than Gaussian distribution which implies an extreme event is more likely to occur (Cont, 2001). The Pearson correlation measures linear combination and it is susceptible to outlier as it was shown by the famous Anscombe's Quartet (Anscombe, 1973). Hence, some of the previous research used a more robust correlation measure such as Kendall's rank correlation coefficient or Spearman's rank correlation coefficient (Bonanno *et al.*, 2004; Kalay, 1982; Micciche *et al.*, 2003; Onnela, Chakraborti, Kaski, Kertesz*, et al.*, 2003). In this study, the Kendall's rank correlation coefficient were used as an alternative (Kendall, 1938). Also known as Kendall's tau coefficient, it is a measure of rank correlation which is computed by

$$\rho_{\tau}(\mathrm{r}_i, \mathrm{r}_j) = \frac{n_c - n_d}{n(n-1)/2} \qquad (3-3)$$

where n_c refers to the number of concordant pairs, n_d is the number of discordant pairs and n is the number of observations. A concordant pair refers to a pair of observations $(\mathrm{r}_{i,t-\Delta t}, \mathrm{r}_{j,t-\Delta t})$ and $(\mathrm{r}_{i,t}, \mathrm{r}_{j,t})$ where the ranks for both elements are the same. For example if $\mathrm{r}_{i,t-\Delta t} > \mathrm{r}_{i,t}$ and $\mathrm{r}_{j,t-\Delta t} > \mathrm{r}_{j,t}$ or $\mathrm{r}_{i,t-\Delta t} < \mathrm{r}_{i,t}$ and $\mathrm{r}_{j,t-\Delta t} < \mathrm{r}_{j,t}$, then the pair are said to be concordant. Likewise, if $\mathrm{r}_{i,t-\Delta t} > \mathrm{r}_{i,t}$ and $\mathrm{r}_{j,t-\Delta t} < \mathrm{r}_{j,t}$ or $\mathrm{r}_{i,t-\Delta t} < \mathrm{r}_{i,t}$ and $\mathrm{r}_{j,t-\Delta t} > \mathrm{r}_{j,t}$, then the pair are said to be discordant.

Though Kendall correlation may be more robust against outliers, it still suffers from the same weakness of common factor indirectly affecting random variables of interest. Therefore partial Kendall correlation which can be computed in the same manner with its Pearson counterpart is employed. The equation for partial Kendall correlation is given by (P. I. Nelson & Yang, 1988):

$$\rho_{\tau}(r_i, r_j : r_m) = \frac{\rho_{\tau}(r_i, r_j) - \rho_{\tau}(r_i, r_m) \cdot \rho_{\tau}(r_j, r_m)}{\sqrt{\left(1 - \rho_{\tau}^{2}(r_i, r_m)\right) \cdot \left(1 - \rho_{\tau}^{2}(r_j, r_m)\right)}} \qquad (3-4)$$

Figure 10 and figure 11 are prepared in the same manner prepared for figure 7 and figure 8, respectively. Note that the overall color scheme of the Pearson correlation matrices is thicker. The Pearson correlation matrix has more red than the Kendall correlation matrix and the partial correlation matrix has more blue than the partial Kendall correlation matrix. From these figures, one can deduce that any analyses performed on these matrices may not agree with each other though they were created using the exact same data set for the same purpose to study the market structure. Hence, subsequent analysis were performed using all four correlation coefficients.

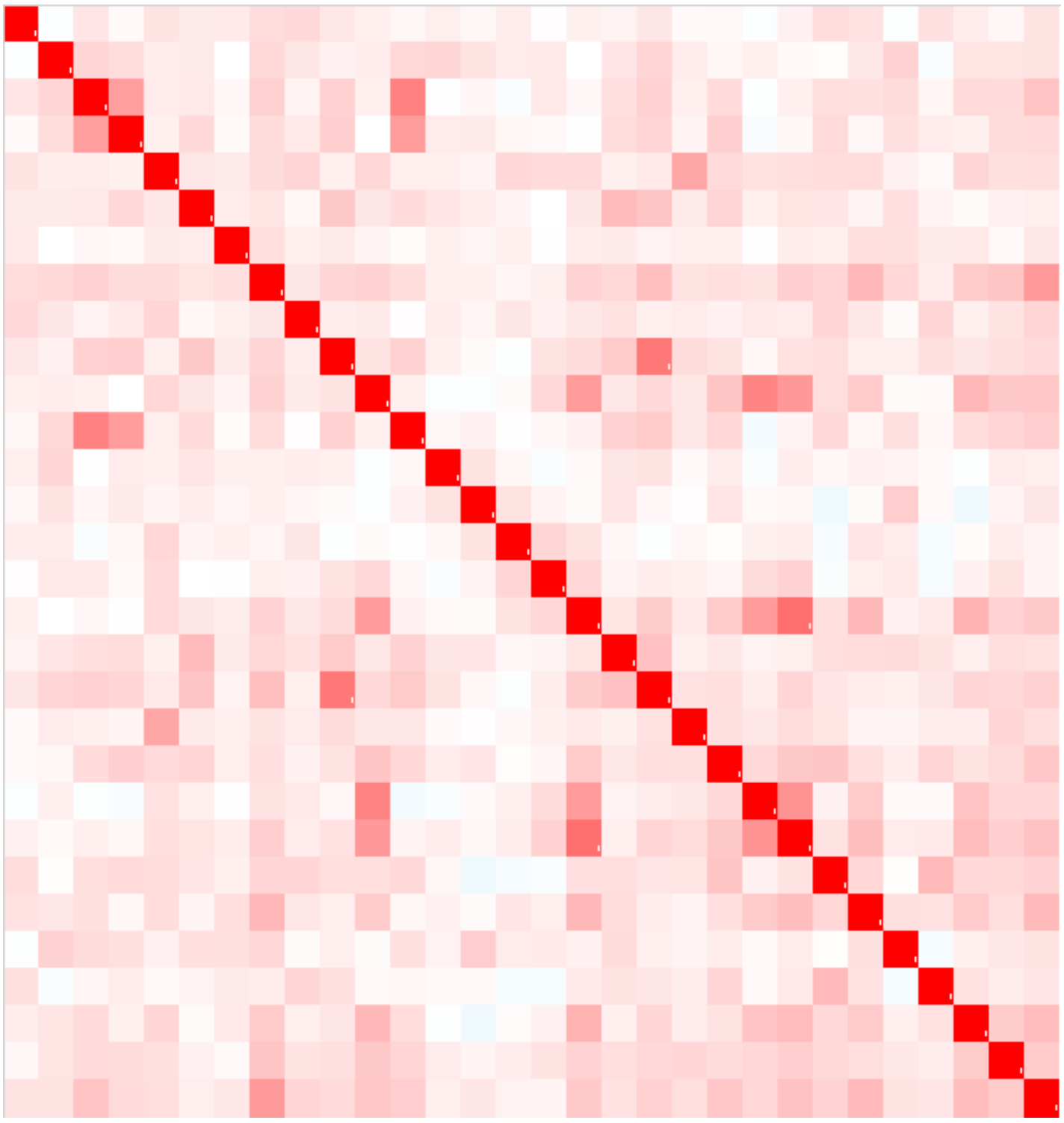

Figure 10: Kendall correlation matrix of largest 30 firms in Korea

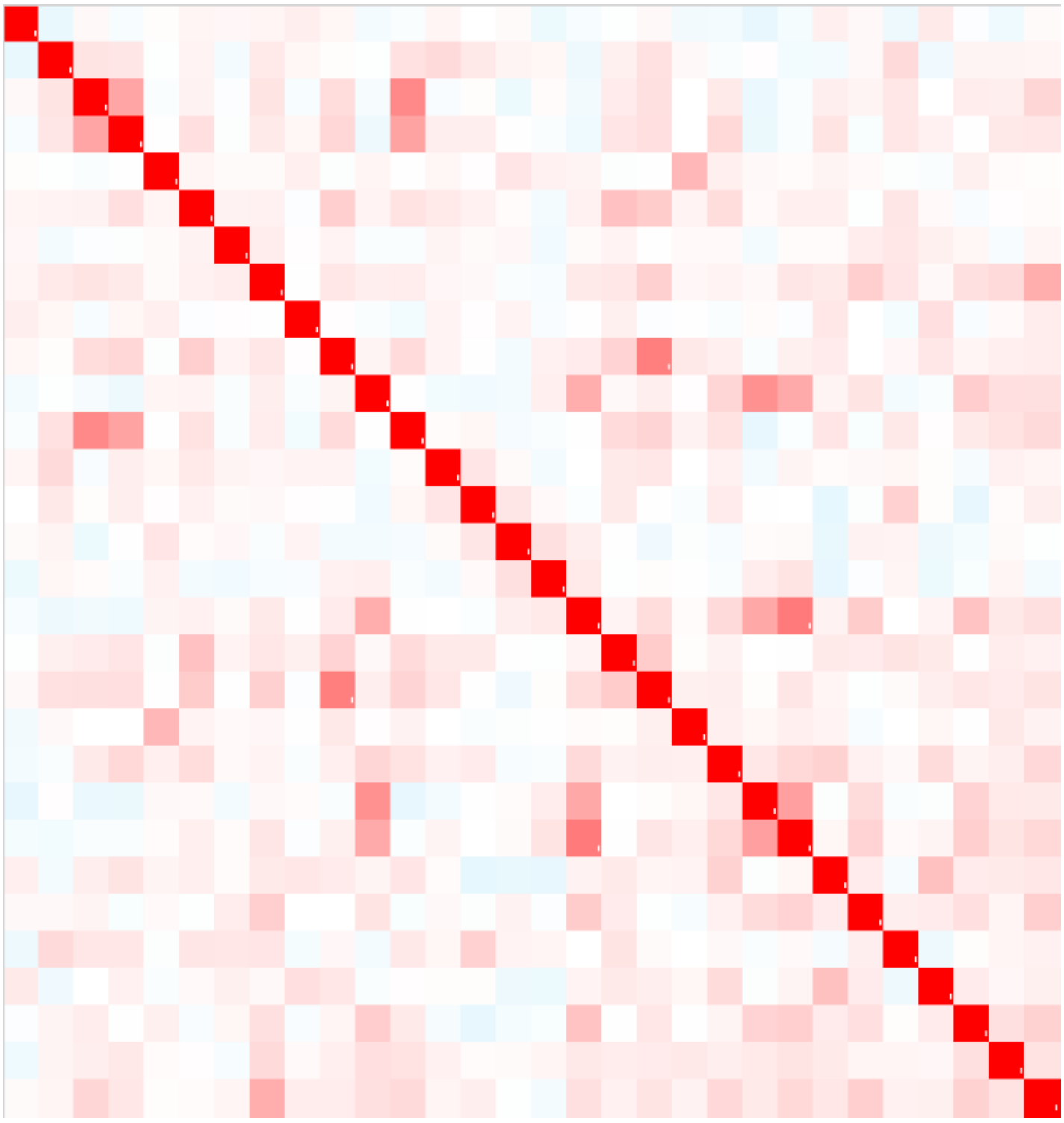

Figure 11: Partial Kendall correlation matrix of largest 30 firms in Korea

3.2 Random Matrix Theory and Filtered Correlation Matrix

In this study, four different type of correlation matrices, Pearson correlation, partial correlation, Kendall correlation, and partial Kendall correlation were estimated for each study. When estimating the coefficients, one should be cautious as it is likely to fall into the curse of dimensionality. The number of coefficients to be estimated for each matrix is $N(N-1)/2$ where N is the number of firms. However, one can only expect to have $N \times T$ data to estimate the variables where T is the length data or number of observations; therefore, if the T is not large enough, it may result in a spurious correlation. Many approaches were developed to mitigate this problem and one of the most promising methods was the application of random matrix theory (Laloux *et al.*, 1999). An extensive research was performed on different filtering methods of the correlation matrix and the random matrix theory was one of the top performers (Pafka & Kondor, 2004).

Random matrix theory compares the eigenvalues of the correlation matrix created using purely random variables to the eigenvalues of the correlation matrix from data. For N by T matrix of random variables, in the limit of $T \to \infty$ and $N \to \infty$

with $Q = \frac{T}{N}$ remains finite, the eigenvalues λ have probability distribution given by

$$\rho(\lambda) = \frac{Q}{2\pi\sigma^2}\frac{\sqrt{((\lambda_{max} - \lambda)(\lambda - \lambda_{min}))}}{\lambda} \quad (3-5)$$

$$\lambda_{max} = \sigma^2\left(1 + \frac{1}{Q} + 2\sqrt{\frac{1}{Q}}\right) \quad (3-6)$$

$$\lambda_{min} = \sigma^2\left(1 + \frac{1}{Q} - 2\sqrt{\frac{1}{Q}}\right) \quad (3-7)$$

where σ^2 is the variance of the $N \times T$ dataset. The eigenvalues larger than λ_{max} are considered to have non-random information. Figures from 12 to 15 shows the eigenvalues of the four correlation matrices. It is clear that for all four cases, there were eigenvalues which deviate from the random matrix which suggest they contain some non-random information. Many studies of financial correlation matrix were performed using random matrix theory (Gopikrishnan *et al.*, 2001; Kenett *et al.*, 2009; Pafka & Kondor, 2004; Plerou *et al.*, 2002; Plerou, Gopikrishnan, Rosenow*, et al.*, 1999; Sandoval & Franca, 2012; Uechi *et al.*, 2015; Utsugi *et al.*, 2004). An interesting method

was developed to filter out the noise part of a correlation matrix using random matrix theory (Kim & Jeong, 2005; Mel, 2015). The authors developed a method to filter out the market-wide effect and the random noise, and construct a correlation matrix of stock groups with non-trivial correlations by removing the largest eigenvalue which represents the market-wide effect and replacing the eigenvalues inside the min-max range with their mean. This approach was proved to be useful in an empirical study, so the filtering procedure was employed for this study. The correlation matrices were decomposed into their eigenvector and eigenvalues. The theoretical values of eigenvalues were computed and compared to the empirically obtained ones. All eigenvalues smaller than λ_{max} defined by equation 3-6 were replaced to the mean value of eigenvalues smaller than λ_{max} to remove the noise component. The largest eigenvalue was removed to eliminate the market-wide effect. The modified eigenvalues and eigenvector were used to reconstruct a correlation matrix. The correlation matrix is removed of noise components and the market-wide effect; therefore, it is called a filtered correlation matrix. All four correlation matrices were filtered through this procedure to create four filtered correlation matrices.

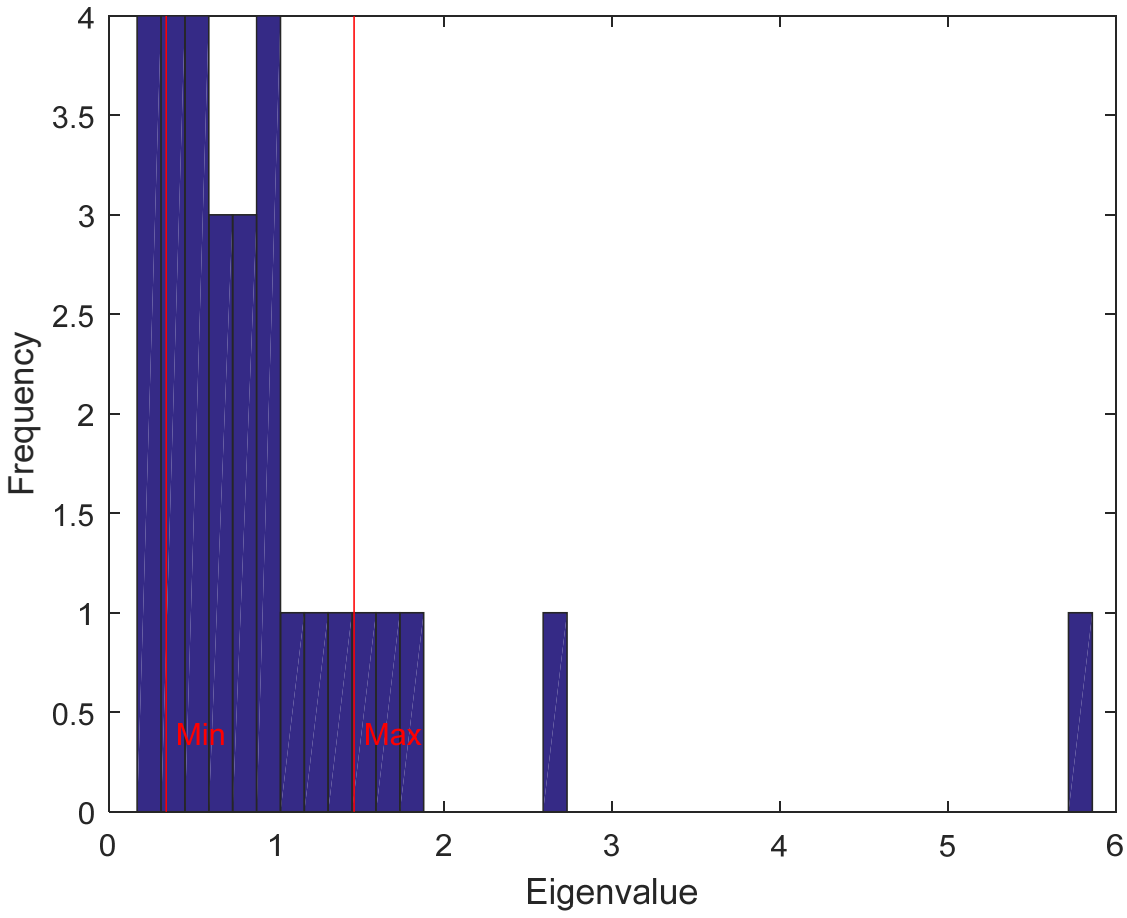

Figure 12: Histogram of eigenvalue distribution for Pearson correlation matrix in 2015

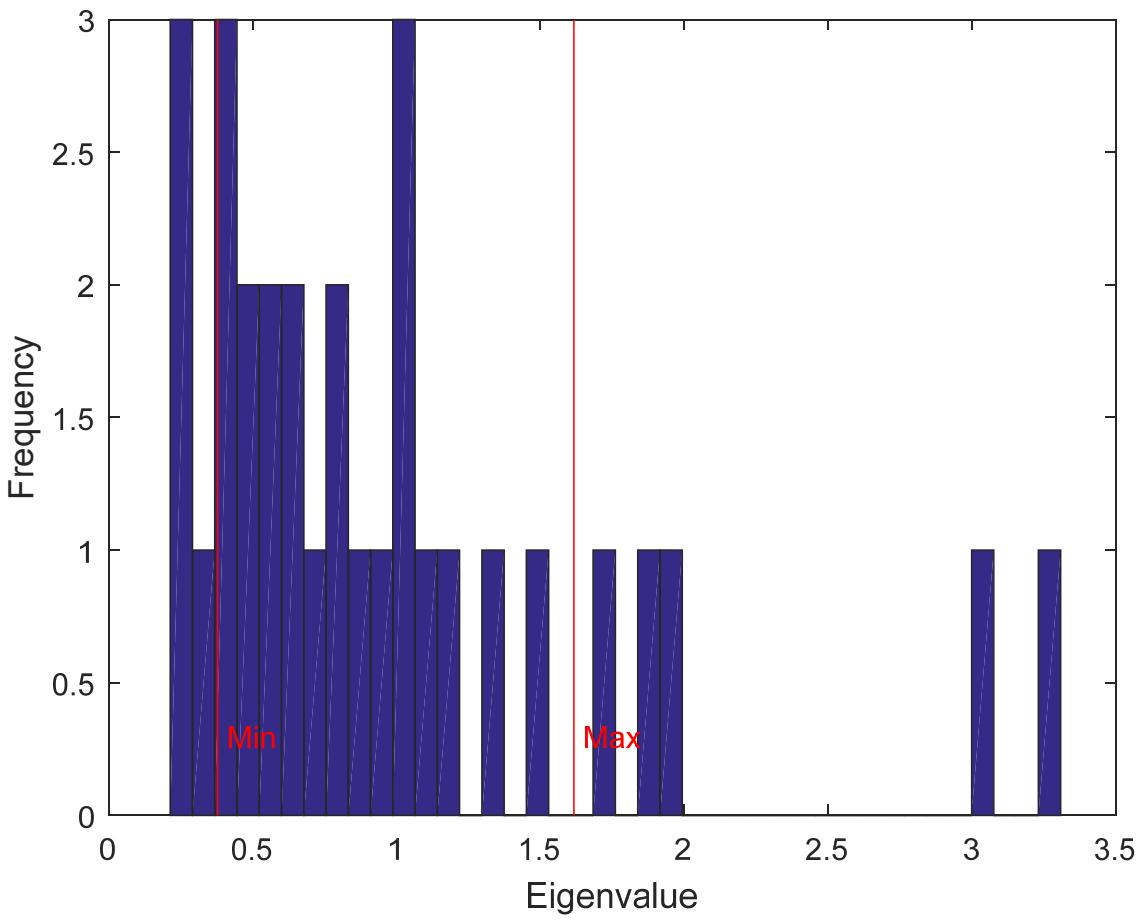

Figure 13: Histogram of eigenvalue distribution for Partial correlation matrix in 2015

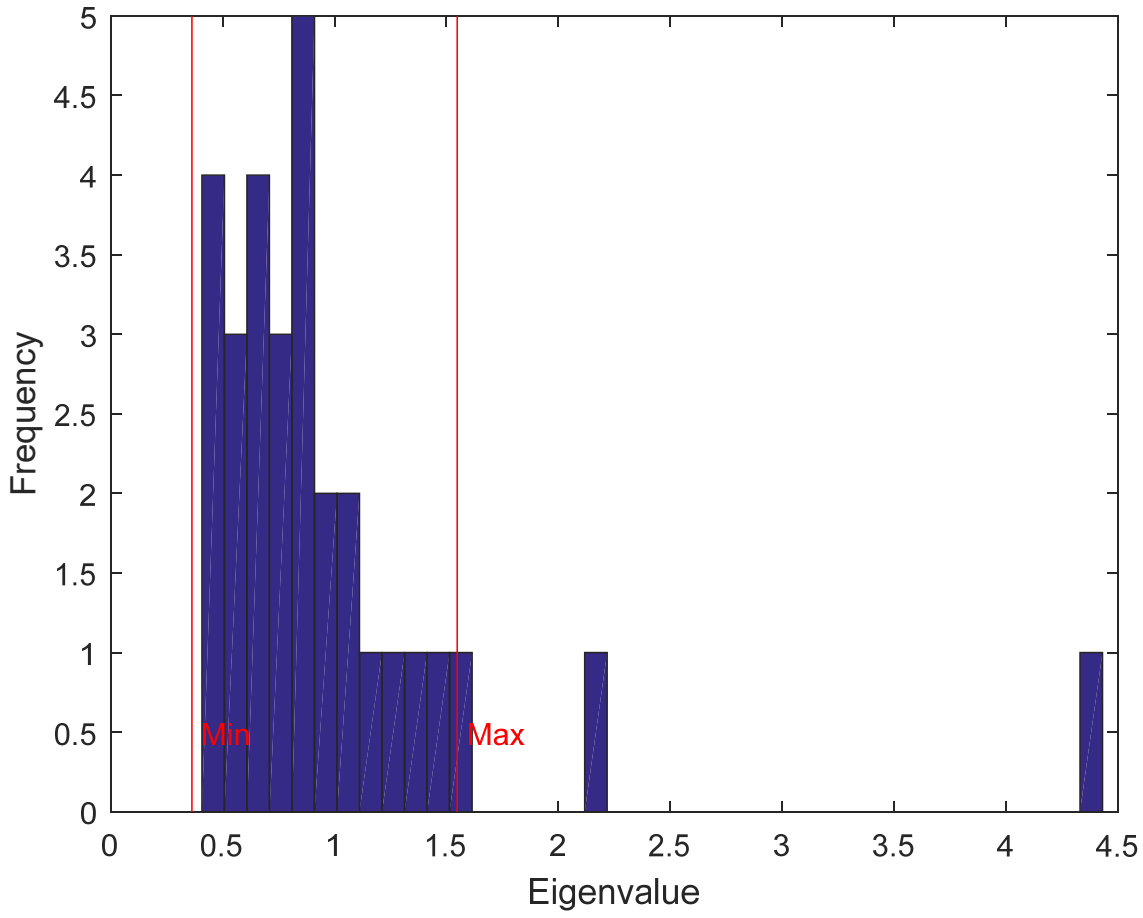

Figure 14: Histogram of eigenvalue distribution for Kendall correlation matrix in 2015

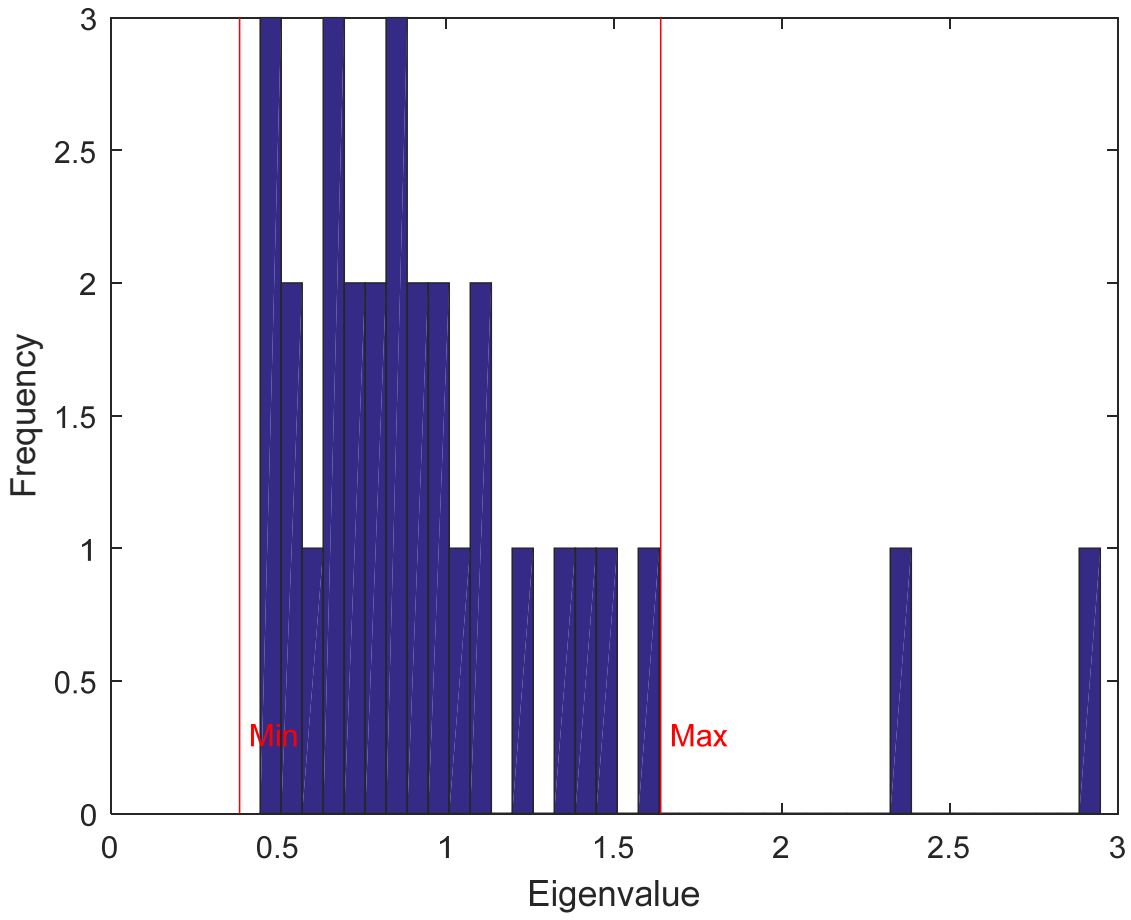

Figure 15: Histogram of eigenvalue distribution for Partial Kendall correlation matrix in 2015

3.3 Dissimilarity Metrics and Dendrogram

A Euclidean distance between two vector r_i and r_j is computed by

$$d(r_i, r_j) = \sqrt{(r_{i1} - r_{j1})^2 + (r_{i2} - r_{j2})^2 + \cdots + (r_{iT} - r_{jT})^2} \qquad (3-8)$$

where T is the number of observations. In this study, dissimilarity between all pairs of firms were computed using the Euclidean distance. The dissimilarity is based on correlation coefficients. However, a correlation coefficient itself is not a suitable distance metric as it fails to satisfy the three axioms of a metric (Mantegna, 1999). The three axioms are

$$d(r_i, r_j) = 0 \; if \; and \; only \; if \; i = j \qquad (3-9)$$

$$d(r_i, r_j) = d(r_j, r_i) \qquad (3-10)$$

$$d(r_i, r_j) \leq d(r_i, r_k) + d(r_k, r_j) \qquad (3-11)$$

The previous study by Mantegna (1999) employed a distance function of the correlation coefficient which is given by

$$d(r_i, r_j) = \sqrt{2(1 - \rho_{i,j})} \qquad (3-12)$$

The four filtered correlation matrices were transformed into the dissimilarity matrices using the equation (3-11), and used to build dendrograms. Hierarchical tree and similar methods are widely adopted in Finance to visually study the structure of the market (Bonanno *et al.*, 2001; Mantegna, 1999; Onnela *et al.*, 2002; Onnela *et al.*, 2004). One can visually confirm which firms were placed next to each other and the hierarchical structure of the market. Figures from 16 to 19 are the dendrograms of the largest 30 KOSPI firms using the dissimilarity matrices. The figures gave many interesting interpretation. First of all, the correlation matrices do agree with some of the conventional wisdom of classification by industrial sector. For example, Hyundai Motor (현대차), Kia Motor (기아차), and Hyundai Mobis (현대모비스), all of which are belong to the same automaker industry and practically one company since Hyundai Mobis and Kia Motor are subsidiaries of Hyundai Motor, formed the first cluster. Also, for all four correlations, the four firms in the banking industry (Shinhan Financial Group (신한지주), KB Financial Group (KB 금융), Samsung Life Insurance (삼성생명), Samsung Fire & Marine Insurance (삼성화재)) were always found next to each other. Three defensive stocks (Korea Electric Power Corporation (한국전력), KT&G, SK Telecom (SK 텔레콤)) and four oil-related stocks (LG Chem (LG 화학), SK Innovation (SK 이노베이션), Lotte Chemical Corp (롯데케미칼), S-Oil) were found in the close proximity. The finding was consistent with

what the previous studies suggest where the firms in the same sector are found in a close proximity and it means that the correlation matrices do have viable information.

Assuming that the correlation matrices do contain economic information, there was another finding worth pointing out which all four correlation matrices agree. Lotte Shopping (롯데쇼핑) is a company mainly engaged in department stores and other stores which makes it a consumer discretionary. Samsung C&T (삼성물산) is primarily engaged in building construction and it falls into the category of industrials. Samsung SDS (삼성에스디에스) provides IT consulting and business solutions and it belongs to IT sector. Three firms, seemingly with no connection, were located close to each other for all four correlations which suggests their returns are correlated even though their business may not be.

Another disagreement between correlation matrices and the conventional wisdom of classification by business operation was found in the industrials sector. POSCO and Hyundai Steel (현대제철) are both manufacturer of steel products and Hyundai Heavy Industries (현대중공업) operates in ship building and platform building. These firms are conventional studied together because the nature of their businesses are closely related to each other. However, four correlation matrices gave different answers for them. The Pearson correlation matrix and the partial correlation matrix agrees and the three firms were placed next

to each other in the dendrograms. The Kendall correlation matrix, on the other hand, separated all three of them far from each other and the partial Kendall correlation matrix put Hyundai Heavy Industries away from the other two.

So far, the analysis was based on the visual observation of dendrograms. When a greater number of stocks is considered, such method becomes unrealistic as there are well over hundreds of firm and the dendrogram practically becomes illegible. A different approach is required to obtain results in a presentable manner.

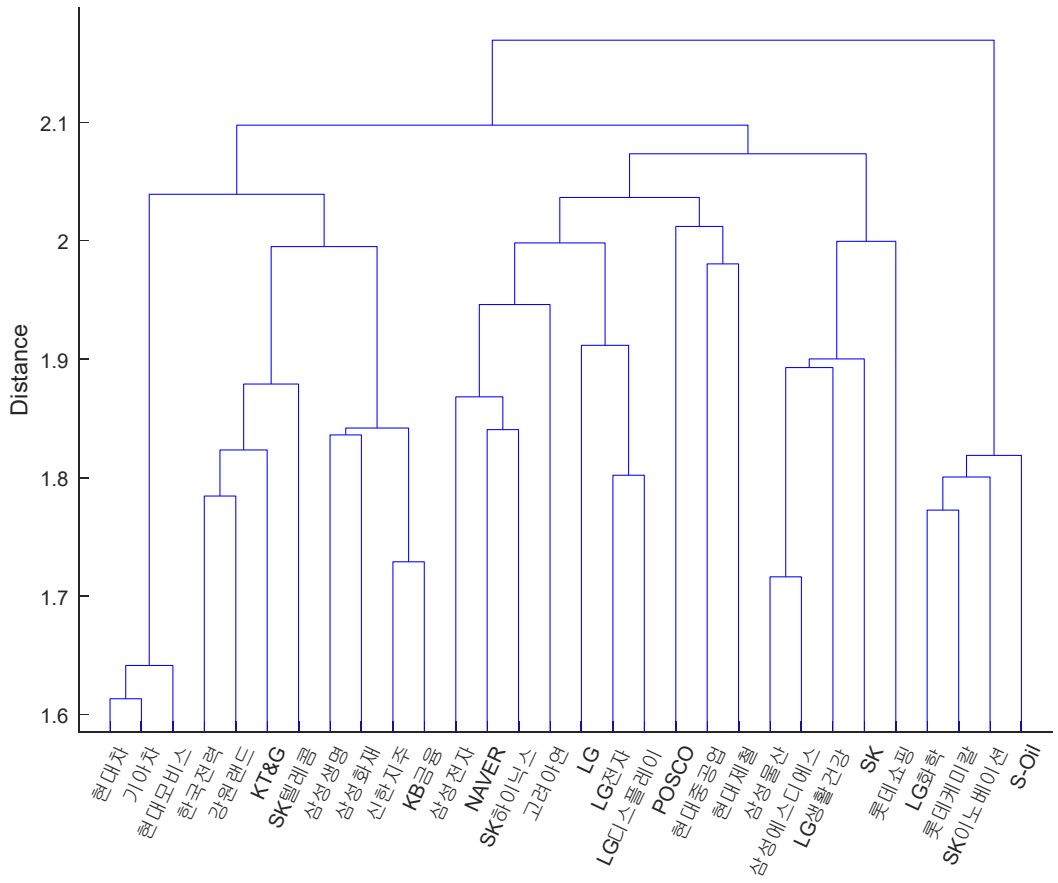

Figure 16: Dendrogram of the largest 30 KOSPI firms using Pearson dissimilarity matrix

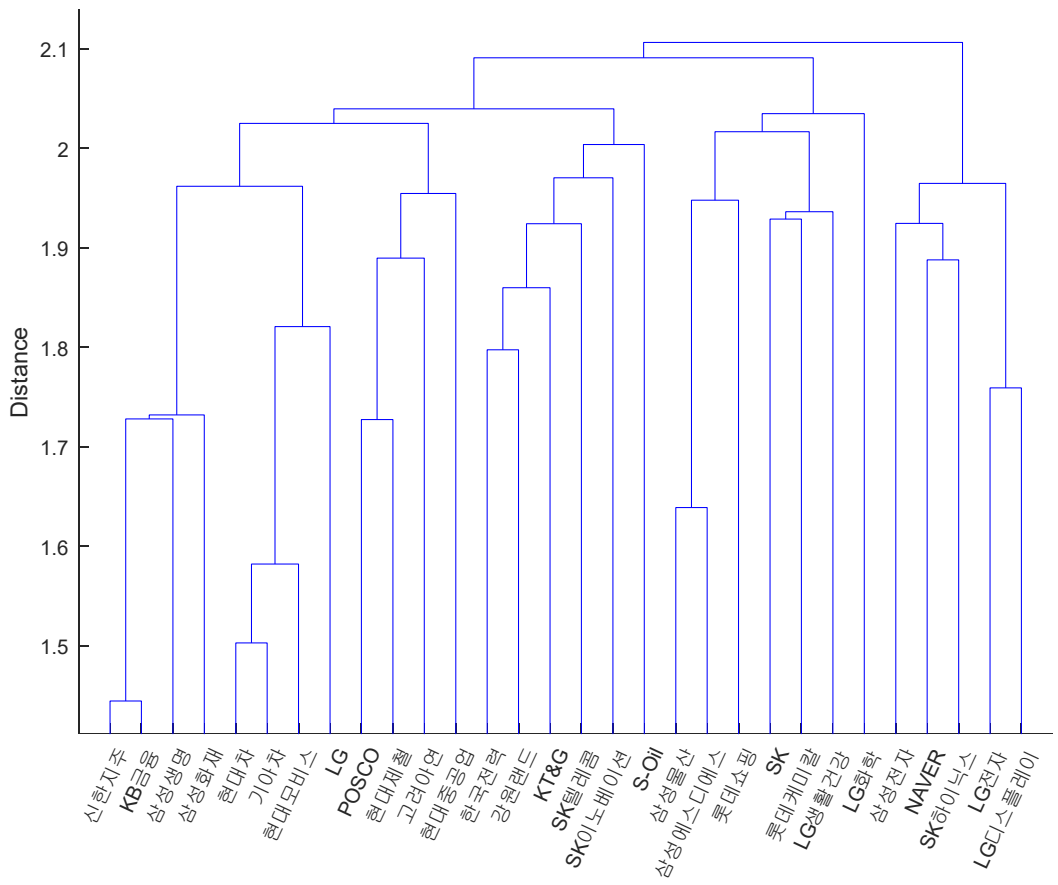

Figure 17: Dendrogram of the largest 30 KOSPI firms using Partial dissimilarity matrix

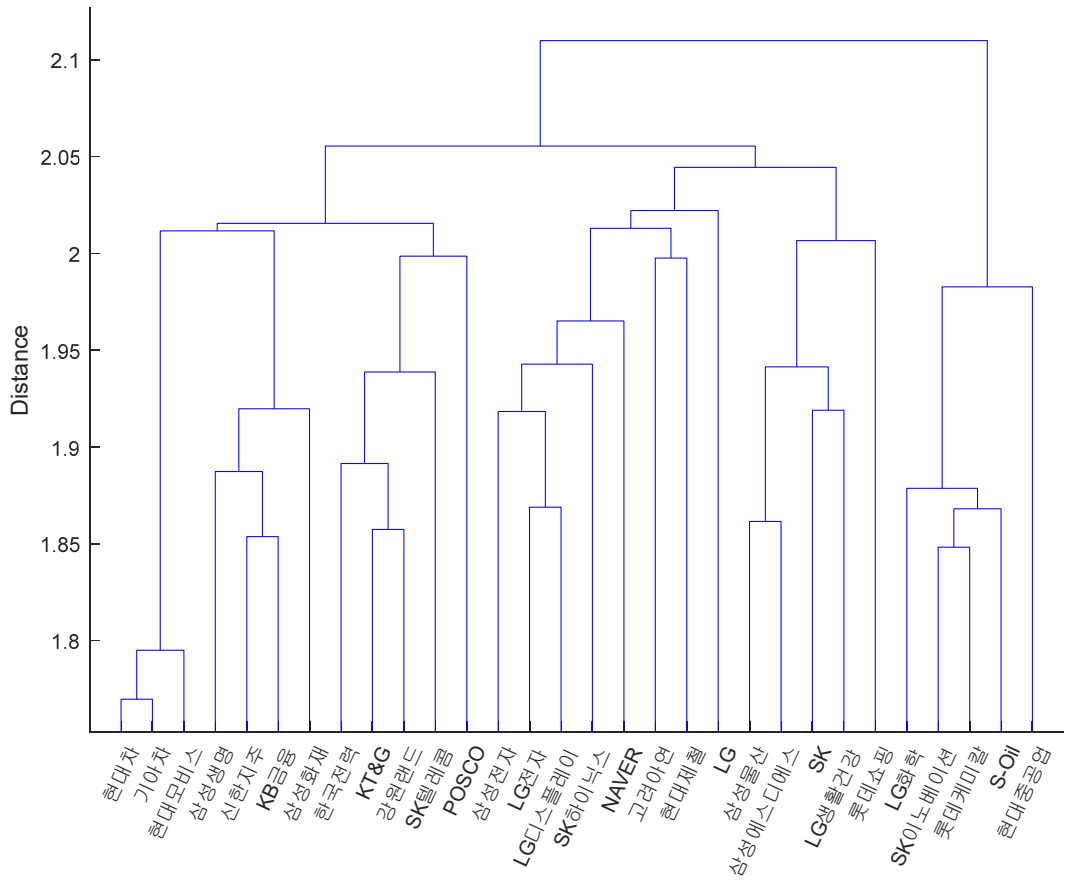

Figure 18: Dendrogram of the largest 30 KOSPI firms using Kendall dissimilarity matrix

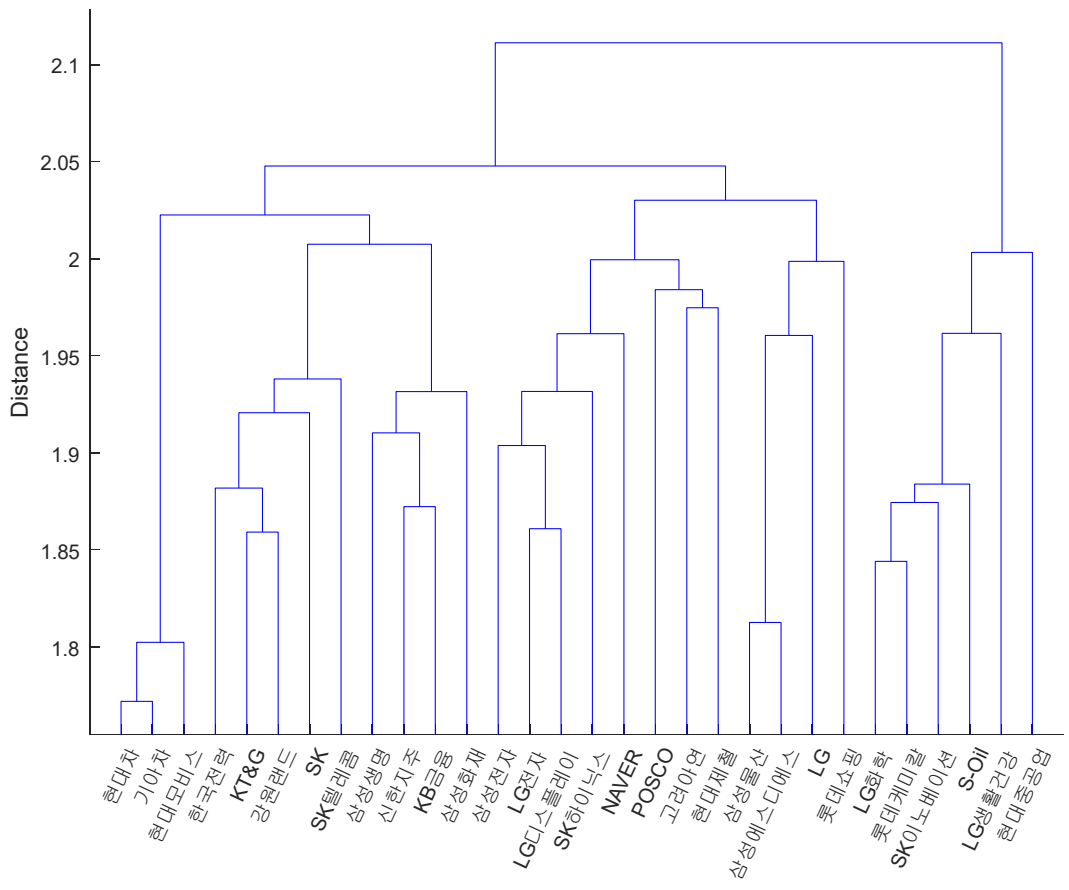

Figure 19: Dendrogram of the largest 30 KOSPI firms using Partial Kendall dissimilarity matrix

Chapter 4

Clustering and Portfolio Analysis

4.1 Return clusters and Industrial Composition

During the past decades, statistical physics became a popular tool for studying the financial market. One of the most prominent results was the use of minimum spanning tree and hierarchical tree to identify the hierarchical structure of financial markets (Mantegna, 1999). Using the Pearson correlation coefficients between stocks from Dow Jones Industrial Average and the Standard and Poor's 500, the author was able to identify a hierarchical structure and groups of stock which were mostly homogeneous in terms of industry the stocks belong to, and provided a meaningful economic interpretation. This was a pioneering application of Network Theory to study financial market. Since then many similar approaches were taken to study the other aspects of the financial market. The correlation of returns from high-frequency data which records transactions in

a period shorter than a day revealed that using different time horizon gave varying hierarchical structure (Bonanno *et al.*, 2001). Also, a network of partial correlation was studied and found that stocks belonging to the financial sector had the largest influence on the entire market (Kenett *et al.*, 2010).

Clustering analysis takes similar approach to study a large set of data. Clustering analysis of stock market also begins with estimating the correlation matrix of assets, and observe which assets are clustered together. A previous study using the Pearson correlation analysis successfully demonstrated that the firms in the Dow Jones Industrial Average are clustered in accordance with the industrial sector (Basalto *et al.*, 2005).

Much of the literature reviewed here generally agreed that the hierarchical structure or the clustering of stocks is consistent with the division by industrial sector, but the results obtained were usually done using a small set of data as it was the case for Chapter 3 of this study. When a larger set of data is considered, not every firm in the data would be financially stable and some of them may not have a clear business model. Therefore, clustering analysis of dissimilarity matrices was performed. The purpose of this step is to study whether the firms in the same sector are generally close to each other in terms of correlation of their returns. The largest 200 firms in KOSPI are chosen for this study. The correlation matrices were computed using daily returns during the year 2015, and the dissimilarity matrices are

prepared in the same manner as the previous section using random matrix theory to filter the correlation matrices and the filtered matrices were transformed using the distance function. Figures from 20 to 23 plot the resulting clusters, one for each dissimilarity matrices. The dimension of clusters were reduced to 2 using Principal Component Analysis for visualization purpose. A detailed list of constituents for each clusters were reported in the Appendix A. An obvious observation was that the plots do not agree with each other. Each clusters put a firm in a cluster and group them with different firms. Given that the matrices are prepared in an *ex−post* manner using historical data, one would expect such deterministic nature in data should result in an agreement. However, the result of analyses disagrees and one may hypothesize that a dissimilarity metric was better than the others to explain the market structure.

Second analysis was performed to study if the clustering results comply with the traditional idea of firms in the same sector are related to each other. Each firms has an industrial sector assigned to it. For example, Hyundai Motor, and Kia Motor, both of which are automobile manufacturers, belong to the same sector of consumer discretionary while Samsung Electronic is assigned to IT sector. Each firms is categorized in accordance with the Global Industry Classification Standard developed by MSCI and Standard and Poor.

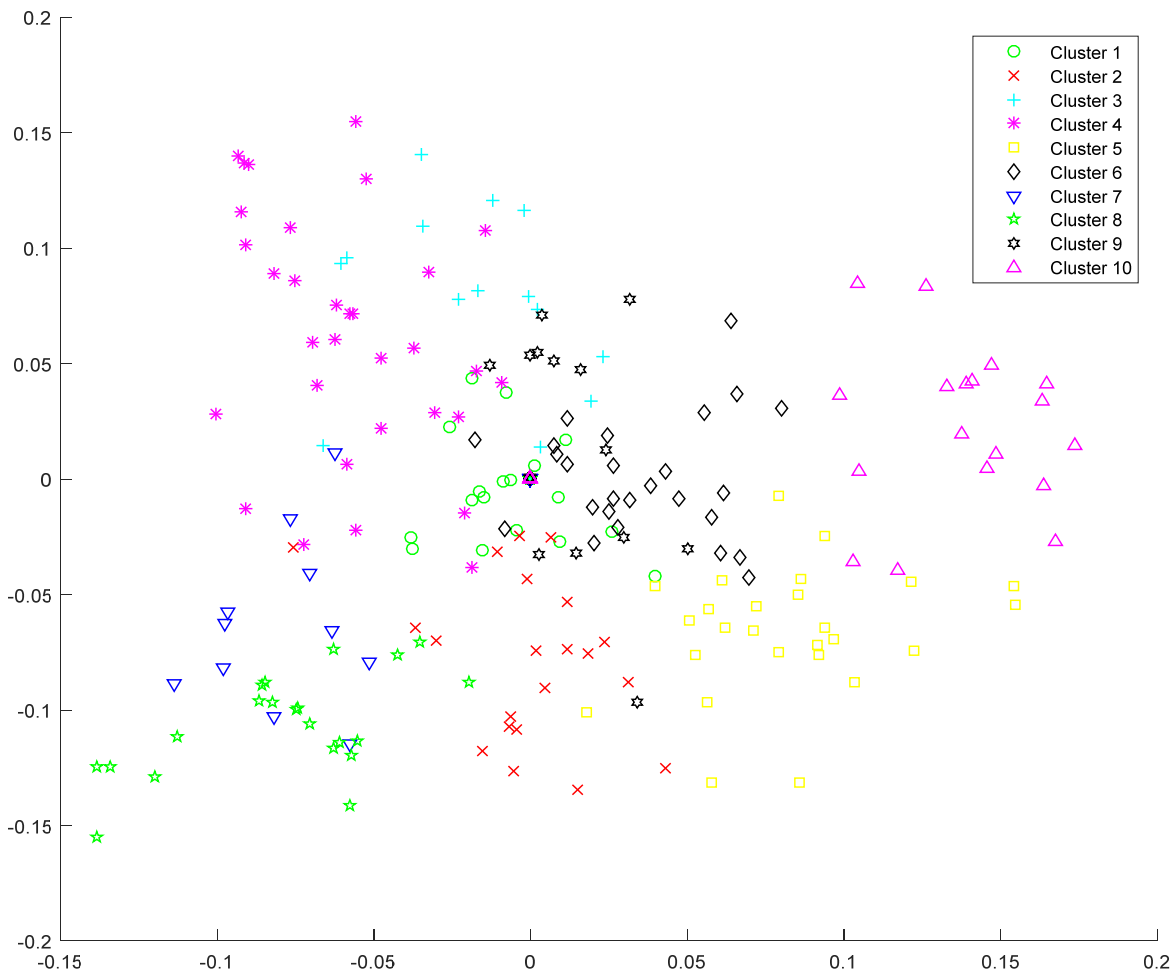

Figure 20: Clustering Pearson dissimilarity matrix

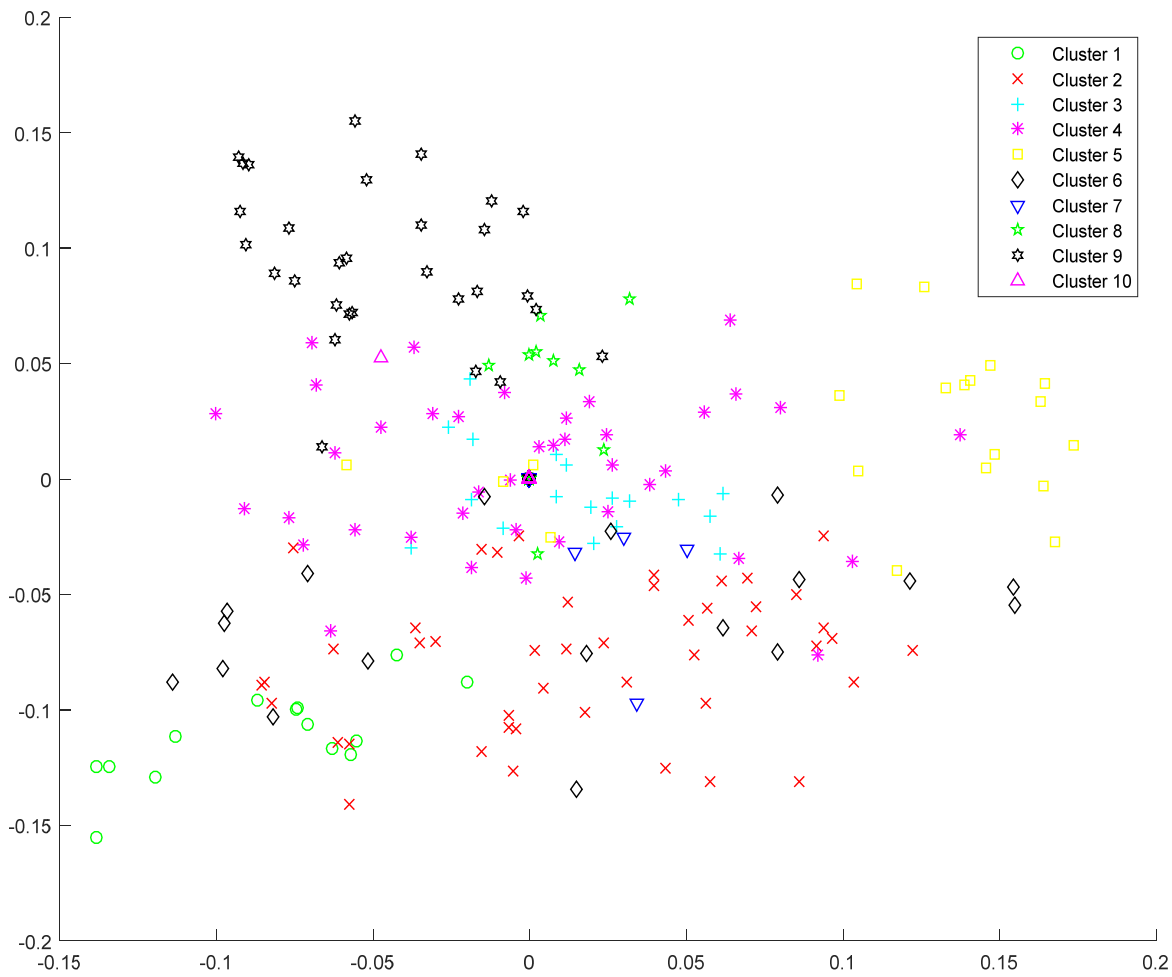

Figure 21: Clustering Partial dissimilarity matrix

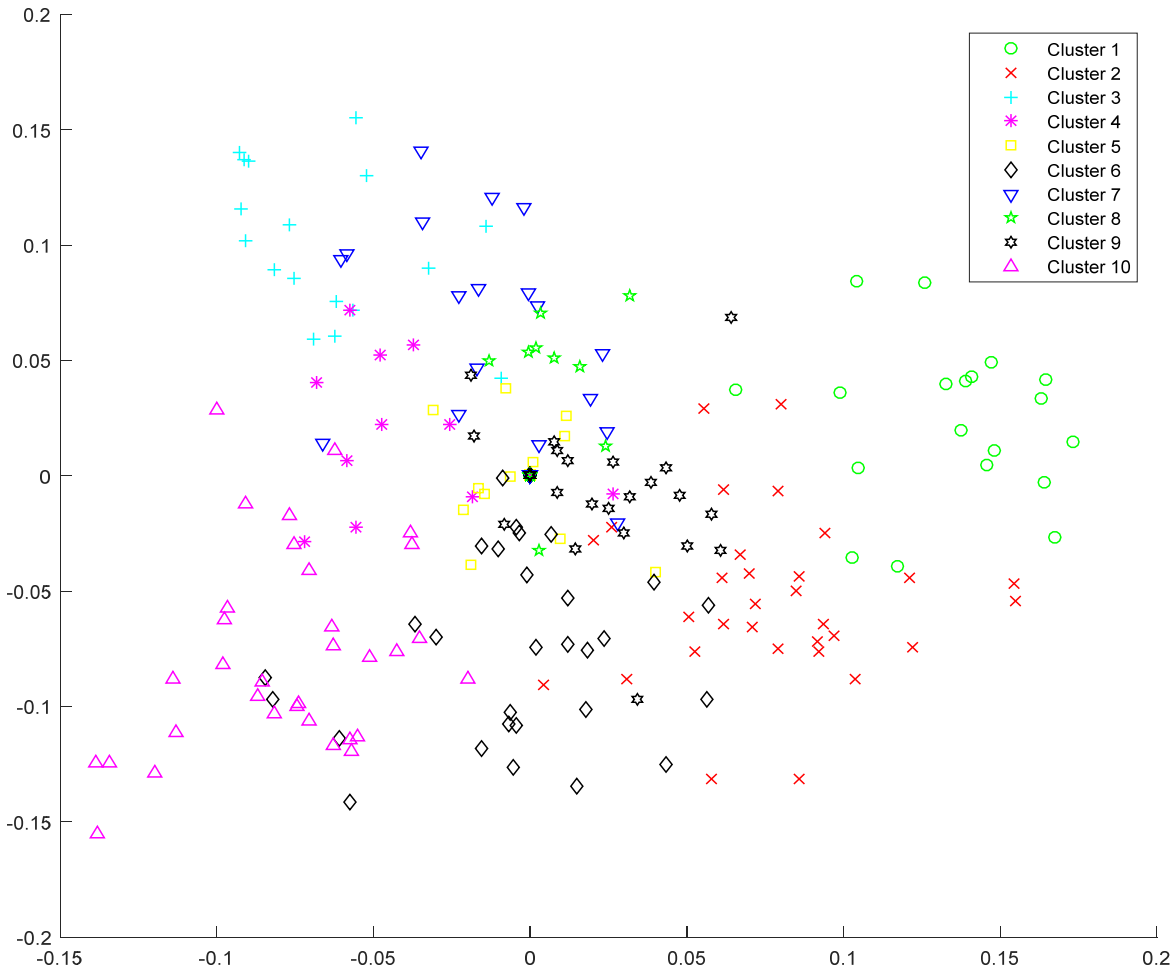

Figure 22: Clustering Kendall dissimilarity matrix

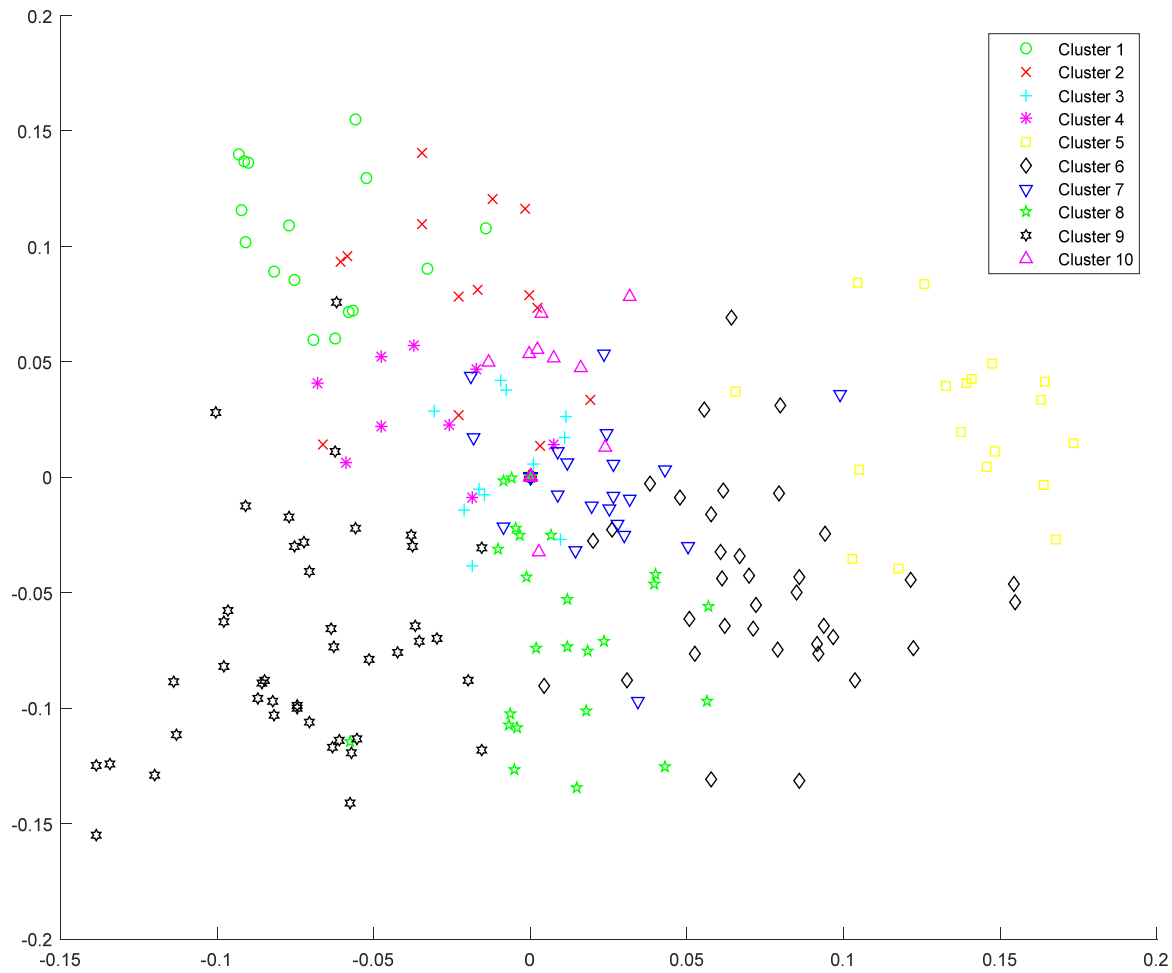

Figure 23: Clustering Partial Kendall dissimilarity matrix

The firms are separated into 10 different sectors: consumer discretionary (CD), financials (F), industrials (I), IT, materials (M), consumer staples (CS), energy (E), healthcare (H), telecomm (T), and utilities (U). A hypothesis was that since the firms in the same industry were exposed to the same environment, it is likely that the majority of the firms in the same sector would fall into the same cluster. The firms are separated into ten clusters, which is the same number with the number of industries used to categorize the firm. Another hypothesis was if the firms in a different sector show a vastly different behavior, then they will fall into different clusters, and each cluster will be dominated by the firms in the same sector. Tables from 1 to 4 were prepared for each dissimilarity matrices. Each row represents a cluster and the size refers to the relative number of 200 firms belonging to the cluster. The contents of the table shows the breakdown of each sector. Each column represents a sector and the sum of each column is 100%. The column shows how firms in the same sector are separated into different clusters. The relative portion of firms in a sector in each cluster is color gradient coded with red. One can easily notice that the two hypotheses proposed do not hold for the most cases. If the hypotheses were true, then each row should have a single cell where a large number of the firms from a sector are grouped together and each column should have a single cell where most of the members of the sector are gathered. However, the result

disagrees and there are handful of cases where more than 75% of the firms from the same sector grouped into the same cluster. One may noticed that such cases were often observed for energy, healthcare, telecomm and utilities sectors. These sectors consist of small numbers of stocks (4, 16, 3, 3) compare to the other sectors (36, 32, 38, 14, 29, 25) which makes it more likely that the few players would dominate the entire market and have significant correlations with each other.

The result has significant implication for practitioners. It suggests that the classification of firms by sector is not a good approach to separate the firms in terms of return correlation. The common practice for a market participant is to assume that the firms in the same sector are exposed to the same risk and/or opportunity, and related to each other; therefore, the stock returns have correlation. When she seeks to diversify her portfolio, she usually does it by picking stocks from different sectors as they are assumed to have smaller correlation than the stocks in the same sector. The tables strongly suggest that it may not be sufficient to solely consider the sector but consider the correlation of returns as well because it seems that the firms in a different sector may be significantly related to each other and the firms in the same sector may have a small correlation which could be a missed opportunity for the market participant.

Table 1: Clustering analysis of the largest 200 firms using Pearson dissimilarity matrix (consumer discretionary (CD), financials (F), industrials (I), IT, materials (M), consumer staples (CS), energy (E), healthcare (H), telecomm (T), utilities (U))

Cluster Size	CD	F	I	IT	M	CS	E	H	T	U
16%	3%		47%	36%	28%					
13%	22%		8%			60%				
13%	6%	9%	18%	50%	17%	4%		6%		
11%	28%	3%	5%			28%				33%
11%		53%							100%	33%
9%	17%	3%	11%	14%	7%	4%	25%			33%
9%					7%	4%		94%		
7%					38%		75%			
7%		28%	8%		3%					
6%	25%	3%	3%							

Table 2: Clustering analysis of the largest 200 firms using Partial dissimilarity matrix
(consumer discretionary (CD), financials (F), industrials (I), IT, materials (M), consumer staples (CS), energy (E), healthcare (H), telecomm (T), utilities (U))

Cluster Size	CD	F	I	IT	M	CS	E	H	T	U
23%	39%	9%	16%			72%			100%	67%
20%	17%	3%	21%	29%	52%	12%	25%	6%		33%
15%			37%	14%	38%		75%			
10%	3%	3%	5%		3%			94%		
9%	6%	13%	11%	50%	3%					
9%	36%		3%			16%				
7%		44%								
5%		28%								
2%			8%		3%					
1%				7%						

Table 3: Clustering analysis of the largest 200 firms using Kendall dissimilarity matrix (consumer discretionary (CD), financials (F), industrials (I), IT, materials (M), consumer staples (CS), energy (E), healthcare (H), telecomm (T), utilities (U))

Cluster Size	CD	F	I	IT	M	CS	E	H	T	U
17%	31%	56%	3%		7%					33%
16%	22%		13%	14%	3%	60%				
15%	33%	6%	8%			28%	25%		100%	33%
11%	3%	9%	16%	36%	17%	4%				
10%					7%	4%		100%		
9%	3%		3%	7%	41%		75%			
9%			39%		7%					
6%	6%		8%		17%	4%				33%
6%	3%		11%	43%						
5%		28%								

Table 4: Clustering analysis of the largest 200 firms using Partial Kendall dissimilarity matrix (consumer discretionary (CD), financials (F), industrials (I), IT, materials (M), consumer staples (CS), energy (E), healthcare (H), telecomm (T), utilities (U))

Cluster Size	CD	F	I	IT	M	CS	E	H	T	U
22%	33%	59%	8%		10%	8%			100%	33%
18%	22%	3%	18%	29%	3%	60%				
12%	31%	3%	8%			24%	25%			67%
10%	6%	6%	13%	29%	17%	4%		6%		
9%					7%	4%		94%		
8%			39%	7%						
7%					38%		75%			
6%	6%		8%		21%					
5%	3%		5%	36%	3%					
5%		28%								

4.2 Market Structure and Portfolio Analysis

The dissimilarity matrices and clusters developed in the previous section were utilized in this section to construct stock portfolios. The dissimilarity matrices serve as a proxy for the market structure and if the market structure revealed by dissimilarity matrices were true, they should be able to cluster the most similar firms, and the portfolio created with firms in different clusters will have minimal correlations with each other and will be able to achieve a superior return-to-risk ratio.

The dissimilarity matrices were estimated using the largest 200 firms in KOSPI and their daily returns of the year 2015. Clustering analysis was then performed on the dissimilarity matrices to group the stocks. Contrary to the previous section where the number of cluster was chosen to be the same with the number of sectors, the silhouette coefficient method described in Chapter 2 was used to determine a cutoff level. Note that cutoff levels may vary from matrix to matrix since the previous sections suggest that the matrices see the market very differently. Once the firms were assigned to clusters, stocks were randomly chosen from each clusters to build a portfolio. It is an equal-weighted portfolio where the weights of each stock are evenly distributed. If a cluster was too small and doesn't have enough firms to choose an equal number of firms, then the cluster is considered an outlier and removed from process of picking

firms. Specifically, if the number of firms in a cluster is less than or equal to the number of firms for portfolio divided by the number of clusters, then the cluster was removed. The process was iterated until all remaining clusters were big enough. After an equal number of firms were picked from each clusters, the remainders, if any, were filled by picking another stock from large clusters in which a member of the largest cluster filling the first remainder, and a member of the second largest filling the next and so on. Small clusters were removed because the small number of firms in the clusters showed different behavior from the vast majority of the firms which makes them a good candidate to be an outlier. Given that the purpose of a portfolio construction is to manage return-to-risk ratio and generate a stable return, it is safer to avoid outliers with erratic behavior. All portfolios and the normalized KOSPI begins with price of 1 at time 0. Except for the normalized KOSPI, the performance of portfolio was random because of randomly chosen stocks. Therefore, the realization is repeated 1000 times and the mean returns and the standard deviation of returns were recorded to compare the performance of portfolio against the benchmark index. A sample realization of mean price series of portfolios using 10 randomly chosen stocks for each dissimilarity matrices and the normalized KOSPI during the same period were plotted in Figure 24. Portfolios were constructed in different sizes to study the

diversification effect. Portfolio sizes of 10, 20 and 30 were constructed 1000 times each for all four dissimilarity matrices. The annual KOSPI return was 0.0181 and the annual standard deviation was 0.0386. The annual returns $\mathrm{r}_{i,T}$ were computed by

$$\mathrm{r}_{i,T} = \frac{S_{i,T}}{S_{i,0}} - 1 \qquad (4-1)$$

where $S_{i,T}$ is the index or stock price at the end of the year and $S_{i,0}$ is the value at the end of the previous year. The annual standard deviation was given by

$$\sigma_{i,T} = \sqrt{T}\sqrt{\langle\left(r_i - \langle r_i \rangle\right)^2\rangle} \qquad (4-2)$$

where $\langle r_i \rangle$ refers to the mean value of r_i over a given period. All portfolios were able to beat the market index. This shouldn't be surprising because the portfolio had the advantage of survivorship bias where the portfolios were created using the stocks fully traded in the period of analysis which effectively eliminates the possibility of adding a firm delisted on a halfway to a portfolio. Therefore, a random portfolio was constructed as an alternative benchmark. The random portfolio was made of stocks randomly chosen from all 200 firms with no constraint such as cluster. Though clustering portfolios picked the stocks randomly as well, they were bound by the clusters and had to pick almost the equal number of stocks from each clusters.

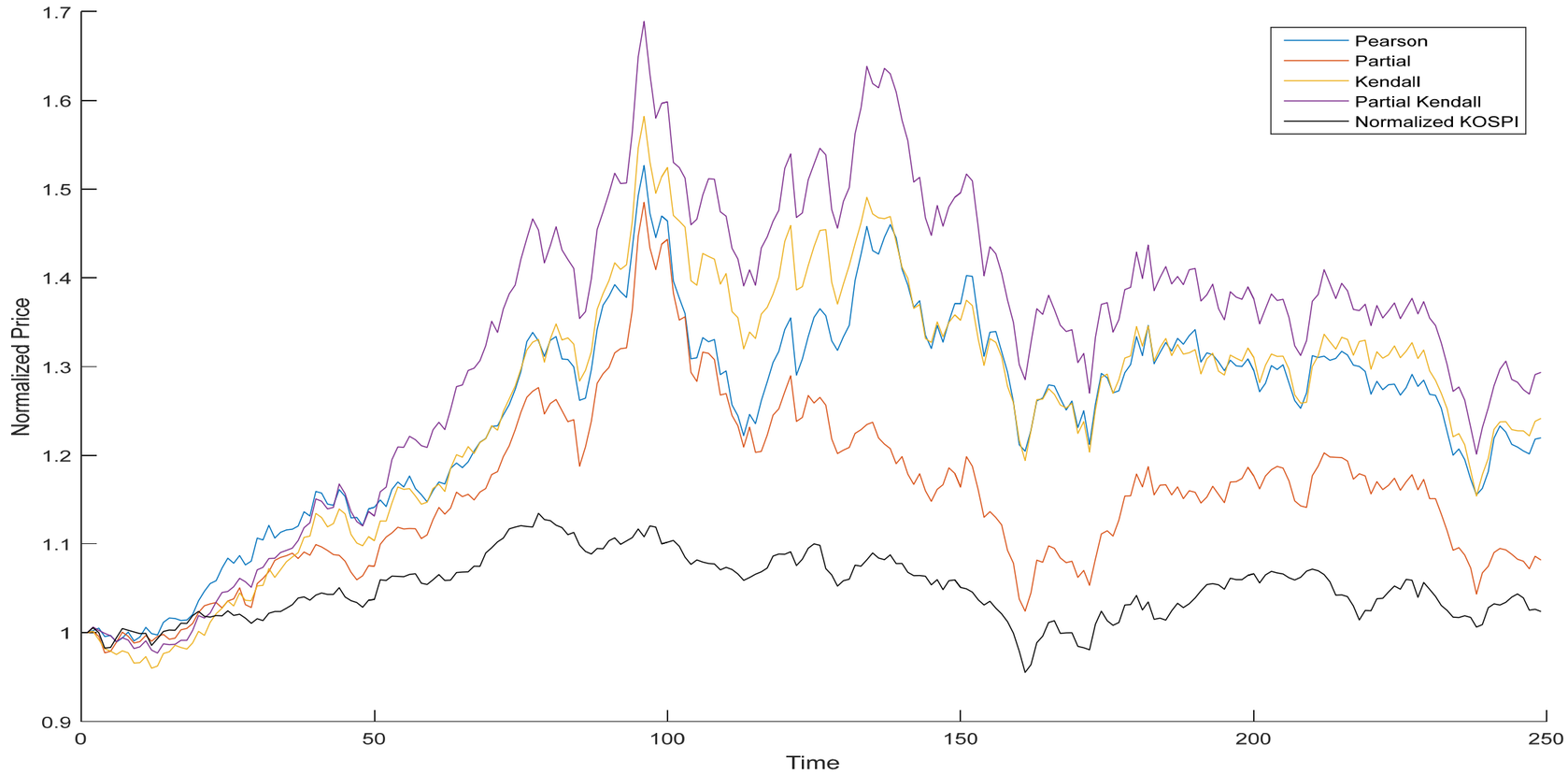

Figure 24: Realizations of portfolios of 10 stocks based on dissimilarity matrices and normalized KOSPI in 2015

Another traditional benchmark Large Cap was added. The Large Cap index is a portfolio of largest firms in KOSPI by market capitalization. The Large Cap index was also created using 10, 20 and 30 largest firms in the market. It represents the relatively stable firms in the market.

Table 5 reports the average annual mean returns and standard deviations of the returns of both benchmarks and four clustering portfolios. The value inside the parentheses of each title of dissimilarity matrix refers to the number of clusters used for analysis which is determined by the silhouette coefficient which is given by equation (2-4). The title of dissimilarity matrices are abbreviated to fit into the table. "P." is an abbreviation for partial. The standard deviation measures the risk of a stock. The values of mean return divided by standard deviation, which is known as return-to-risk ratio, were reported in the table as well. The Large Cap index had a very poor performance and all clustering portfolios were able to beat the index. However, the random portfolio was able to outperform most of the clustering portfolios except for the Pearson correlation with the portfolio size of 20 and 30.

It begs a question why such underperformance was observed for clustering portfolios. Hence, clusters of each clustering portfolio were analyzed. Table 6 provides the details regarding the constituents of clusters. Instead of identifying individual firms, the overall performance of each clusters was analyzed in terms

Table 5: Mean return, standard deviation, and return−to−risk ratio for different size of portfolio

M = 10	Random	Large Cap	Pearson (8)	Partial (10)	Kendall (6)	P. Kendall (7)
Annual Return	0.27	−0.11	0.21	0.14	0.20	0.20
Annual Std	0.22	0.16	0.22	0.21	0.21	0.22
Return / Risk	1.20	−0.68	0.99	0.68	0.94	0.90

M = 20	Random	Large Cap	Pearson (8)	Partial (10)	Kendall (6)	P. Kendall (7)
Annual Return	0.26	0.06	0.28	0.18	0.21	0.16
Annual Std	0.20	0.14	0.20	0.19	0.20	0.18
Return / Risk	1.33	0.40	1.40	0.94	1.08	0.87

M = 30	Random	Large Cap	Pearson (8)	Partial (10)	Kendall (6)	P. Kendall (7)
Annual Return	0.26	0.06	0.29	0.18	0.19	0.17
Annual Std	0.19	0.15	0.19	0.19	0.19	0.18
Return / Risk	1.41	0.42	1.49	0.96	1.01	0.93

of average stock returns. Avg. Change in Stock indicates the mean value of annual returns of all members of the cluster. Number of Stock > 0 counts the number of stock whose price increased over the period and Number of ΔStock < 0 counts the number of stock whose price decreased over the period.

The portfolios were created in such a manner that firms were picked from each clusters and the number of firms picked from each cluster was kept nearly the same. Therefore, with 1000 trials, the average performance should resembles that of the equal-weighted portfolio based on the clusters. The expected return of the equal weighted portfolio should converge to the mean value of cluster returns which is 26%, 19%, 19%, and 15% respectively for each clustering portfolios. The finding is consistent with the performance reported in table 5 with Pearson clustering portfolio showing the highest performance and partial Kendall having the smallest return and return-to-risk ratio especially when the portfolio size is the biggest. It suggests that by removing a cluster with subpar performance, which results in an increased mean value of cluster return, one can achieve a superior expected return of portfolio.

Notice that returns of some of the clustering portfolios were similar to each other or the random portfolio. Given that the mean annual returns are the sample mean of random variable annual return, a hypothesis test is needed to confirm that the values are statistically different.

Table 6: Analysis of constituents of clustering portfolios

Pearson Cluster

Cluster Size	26	26	11	21	39	13	46	18
Avg. Change in Stock	65%	48%	−15%	−6%	6%	−10%	4%	117%
Number of ΔStock > 0	23	18	3	6	18	3	19	18
Number of ΔStock < 0	3	8	8	15	21	10	27	0

Partial Cluster

Cluster Size	14	46	18	40	20	18	4	9	30	1
Avg. Change in Stock	−2%	32%	15%	27%	87%	23%	−33%	1%	8%	−5%
Number of ΔStock > 0	5	31	9	20	17	10	0	3	13	0
Number of ΔStock < 0	9	15	9	20	3	8	4	6	17	1

Kendall Cluster

Cluster Size	9	21	41	50	33	46
Avg. Change in Stock	1%	21%	5%	88%	−8%	7%
Number of ΔStock > 0	3	10	19	46	10	20
Number of ΔStock < 0	6	11	22	4	23	26

Partial Kendall Cluster

Cluster Size	20	20	24	43	30	9	54
Avg. Change in Stock	14%	−11%	13%	−6%	10%	1%	86%
Number of ΔStock > 0	10	4	15	13	14	3	49
Number of ΔStock < 0	10	16	9	30	16	6	5

Table 7: T−test for portfolio's annual returns in 2015
(H_0: $\langle r_{pf_i} \rangle = \langle r_{pf_j} \rangle, i \neq j$)

T test for M = 10	Random	Pearson	Partial	Kendall	Partial Kendall
Random	−	−	−	−	−
Pearson	1.05E−08	−	−	−	−
Partial	3.36E−39	9.45E−18	−	−	−
Kendall	3.02E−13	8.71E−02	1.38E−12	−	−
Partial Kendall	2.75E−15	1.65E−02	1.11E−10	4.88E−01	−

T test for M = 20	Random	Pearson	Partial	Kendall	Partial Kendall
Random	−	−	−	−	−
Pearson	1.07E−02	−	−	−	−
Partial	1.33E−29	4.94E−48	−	−	−
Kendall	2.39E−12	6.71E−25	1.32E−07	−	−
Partial Kendall	3.35E−56	1.47E−86	9.15E−05	8.59E−24	−

T test for M = 30	Random	Pearson	Partial	Kendall	Partial Kendall
Random	−	−	−	−	−
Pearson	2.27E−07	−	−	−	−
Partial	8.58E−53	7.92E−96	−	−	−
Kendall	1.84E−46	1.11E−90	3.59E−02	−	−
Partial Kendall	4.99E−74	8.82E−128	2.33E−02	2.23E−06	−

A simple T-test was performed to compare the sample means and the corresponding p-values were computed. All pairwise results were summarize in table 7. All p-values were nearly zero except for Kendall and Pearson and Kendall and partial Kendall with portfolio size of 10. The p-value suggests Kendall and partial Kendall were very similar when the portfolio size was small. An explanation would be that two clustering portfolios have similar clusters.

Table 8 is prepared to confirm if the constituents of clusters were indeed similar for Kendall and Partial Kendall clusters. For each pair of clusters, the number of constituents found in both clusters and the size of their union set were computed and their ratio in percentage was presented in the table. For example, Cluster 1 from Kendall Clusters and Cluster 6 from Partial Kendall Clusters were identical that the intersection set was the same with both clusters and their union set. Cluster 4 from Kendall Clusters has 50 firms, and Cluster 7 from Partial Kendall Clusters consists of 54 firms. Two clusters has 49 firms in common; therefore, in the table, it was 49 intersecting members divided by size of the union set which is 55, and it returned 89.09%. The values greater than 50% were highlighted in the table which were found in every row, or for every Kendall Clusters implying that two clustering results were similar to each other; hence the two portfolios were more likely to have similar performance until the size of portfolio grows big enough so that

the variance becomes small enough to distinguish the two samples. The tables for other pairs of clustering portfolios were prepared for comparison and can be found in Appendix B. The similarity between Pearson and Kendall clusters weren't as significant but three clusters did have over 50% in common which explains why they may look similar when the variance was large. The correlation matrices used to create clusters for Table 8 and associated appendix were estimated over a year. However, one year may not be sufficient to capture the fundamental differences between stocks, and estimation over longer period may yield a different outcome. If there was a clear difference between stocks and estimation over longer period was able to reveal it, then the clustering results may converge and become much similar to each other.

Table 8: Common constituents of Kendall and Partial Kendall clusters in 2015

		Partial Kendall Clusters					
Kendall Clusters Cluster Number	1	2	3	4	5	6	7
1						100%	
2		58%	3%				7%
3			20%	55%	11%		
4		1%					89%
5				2%	73%		
6		6%	16%		3%	65%	

Hence, correlation matrices were estimated over two years period and three years period to analyze the similarity of

constituents of clusters. Table 9 and 10 provide the common constituents between Pearson and Partial for the two different lengths of estimation, respectively. Table 11 and 12 were prepared for Kendall and Partial Kendall in the same manner. Other cases were placed in Appendix C. As the estimation periods become longer, a correlation matrix and its partial counterpart become increasingly similar to each other which implies that the pair of correlation matrices agrees with the big picture of the market but slightly differ in treating details. In other words, estimation over longer period could be better to capture the fundamental dissimilarity between firms.

Three different sizes of portfolio were constructed to confirm if the diversification effect presents in the clustering portfolios (Statman, 1987). Figure 25 shows the annual standard deviation of portfolios as a function of portfolio size. The diversification effect suggests that a portfolio with greater number of stocks should have a smaller variance because firm specific risks are diversified away as the number of firms in a portfolio grows. The theory was consistent with the case of random portfolio and all clustering portfolios. However, the Large Cap index, which suffered from a poor performance, had a small increase in variance as portfolio grew from 20 to 30.

Table 9: Common constituents of Pearson and Partial clusters estimated over two years

2014 ~ 2015	*Partial Clusters* 1	2	3	4	5	6	7	8	9	10
Pearson Clusters 1		50	2%	9%			1%			
2	13	26	2%	11			5%			
3				3%	92					
4				44			5%			
5		3%	33	8%			4%	18		5%
6									100	
7				4%			70			
8				1%		95				

Table 10: Common constituents of Pearson and Partial clusters estimated over three years

2013 ~ 2015	*Partial Clusters* 1	2	3	4	5	6	7	8	9	10
Pearson Clusters 1					79	3%				
2					7%	44				5%
3	16	32				19	2%			
4	9%	19	57		1%	3%		1%		
5				92						
6		2%				2%	82			
7				3%		2%			89%	
8								90		

Table 11: Common constituents of Kendall and Partial Kendall clusters estimated over two years

2014 ~ 2015	*Partial Kendall Clusters* 1	2	3	4	5	6	7
Kendall Clusters 1			5%	84%	2%		
2			89%			1%	
3					50%	8%	
4	29%	57%		1%	10%		
5							100%
6						86%	

Table 12: Common constituents of Kendall and Partial Kendall clusters estimated over three years

2013 ~ 2015		*Partial Kendall Clusters*						
		1	2	3	4	5	6	7
Kendall Clusters	1			99%	1%			
	2				65%			
	3				2%			91%
	4		3%		16%		52%	
	5					100%		
	6	44%	43%		6%			

The purpose of portfolio diversification is that by doing so, one can achieve a smaller variance which leads to higher return-to-risk ratio. Figure 26 is organized to see if it can be done using clustering portfolios. Though the variances were reduced as the size of portfolios grew, the returns didn't follow for many cases. As it was suggested before, it seems a step to remove inferior cluster is necessary to improve the portfolios' performance.

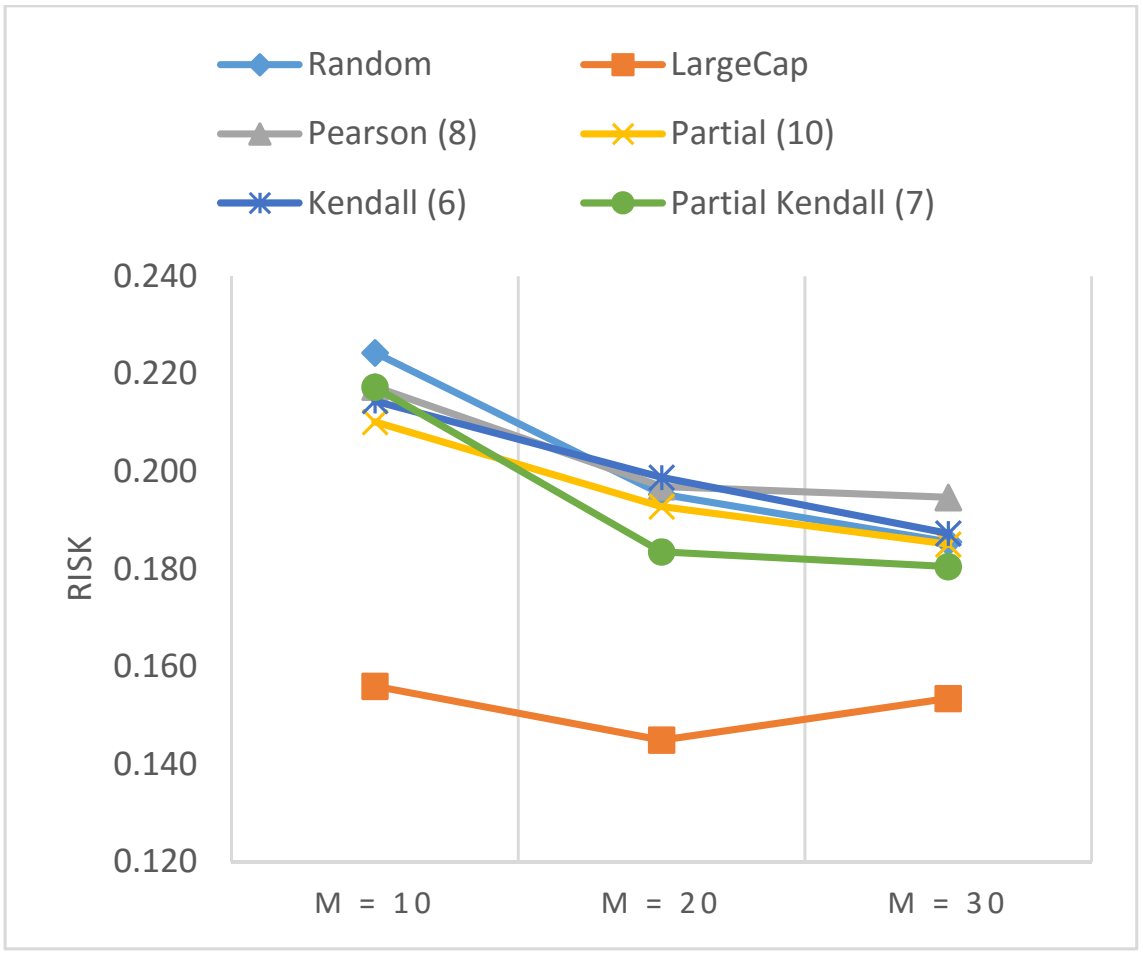

Figure 25: Diversification effect of clustering portfolios

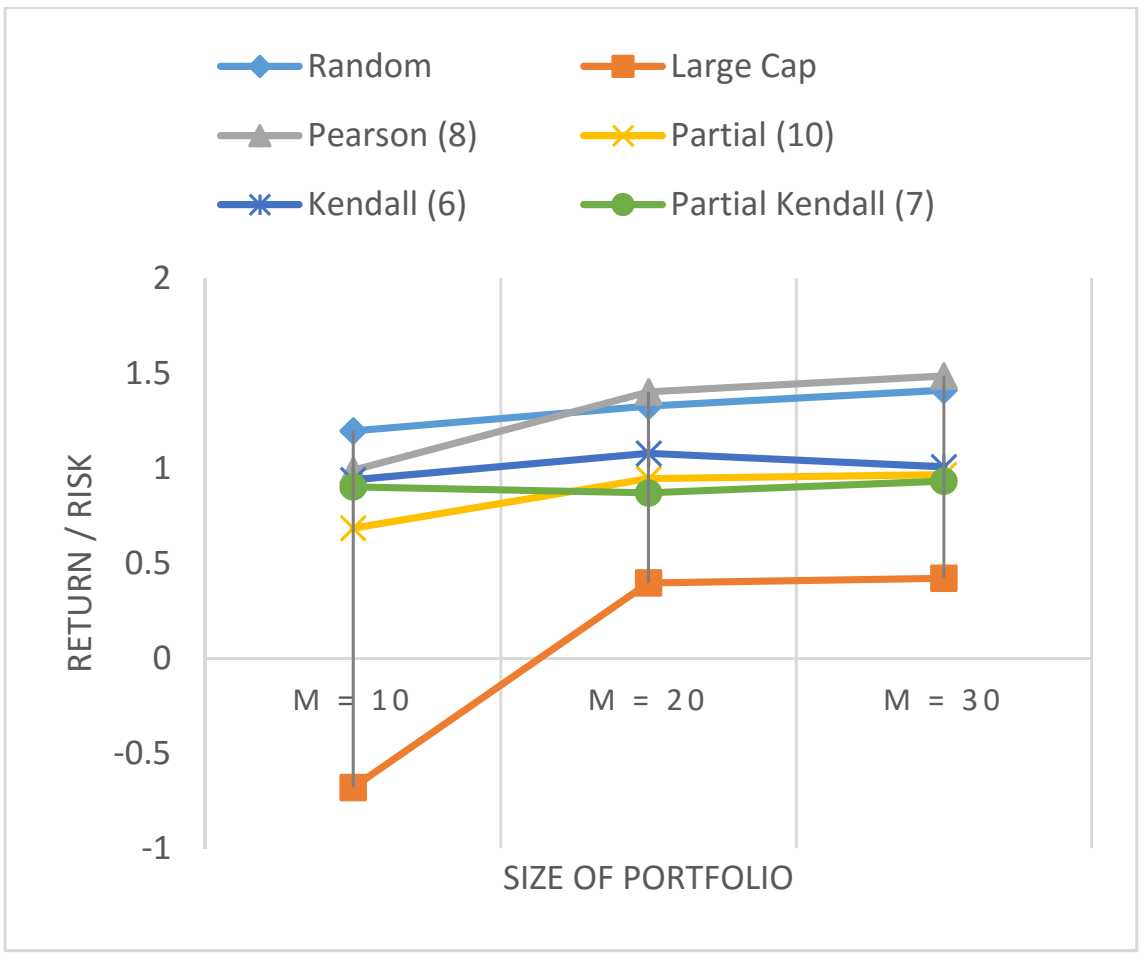

Figure 26: Diversification effect and return-to-risk ratio of clustering portfolios

4.3 Trading Simulation

The previous section was dedicated to the analysis of the market structure and portfolio construction based on correlation coefficients. It was studied in an *ex-post* manner after the market structure was determined. Dissimilarity matrices were compared to study how they interpret the market. In this section, a similar analysis was performed in an *ex-ante* manner. A portfolio construction is important because practitioners do have to buy stocks and build their own portfolio. They do not buy it to study the market structure but to achieve capital gain. In a real world scenario, unlike the previous section, stock returns and their correlations are unknown for the future period. However, one could make an educated guess that if the dissimilarity matrices and clusters were able to separate the firms with fundamental differences, then the fundamental would not change over a short period and one can build a diversified portfolio using the historical data. A trading simulation was performed in an *ex-ante* manner to test this theory. The portfolios were constructed in the same manner described in Section 4.2 except the size of portfolio was fixed to 20. 751 days of observation from January 2012 to December 2014 were used to estimate correlations. The distances were computed and the dissimilarity matrices were constructed. The matrices were used to separate the firms into clusters and the portfolios of randomly chosen stocks from

different clusters were built. The portfolios were then launched after the correlation estimation was completed and observed for the next 250 days until the end of 2015. Sample realizations of mean price series are plotted in figure 27. Similar to the section 4.2, the portfolios were randomly built 1000 times each. The KOSPI return and standard deviation were the same with the previous section since it covered the same period. All portfolios were able to beat the market return of 0.0181 and the return-to-risk ratio of 0.4689.

The return, standard deviation, and return-to-risk ratio of clustering portfolios and benchmarks are presented in table 13. The clustering portfolios were able to outperform the benchmark portfolios except for the random portfolio. Since all portfolios are bounded by the same universe of 200 firms, a possible explanation for a significant underperformance is that the clustering provided an environment where inferior stocks were more likely to be picked into a portfolio. Clustering analysis might be able to provide such condition because it is an algorithm to group similar objects together. If the similar objects happened to be underperformers, then the clustering analysis would put them in the same cluster. Since the portfolio was designed using an equal number of firms from each clusters, the proportion of underperformers could be significantly greater than the proportion in the entire set.

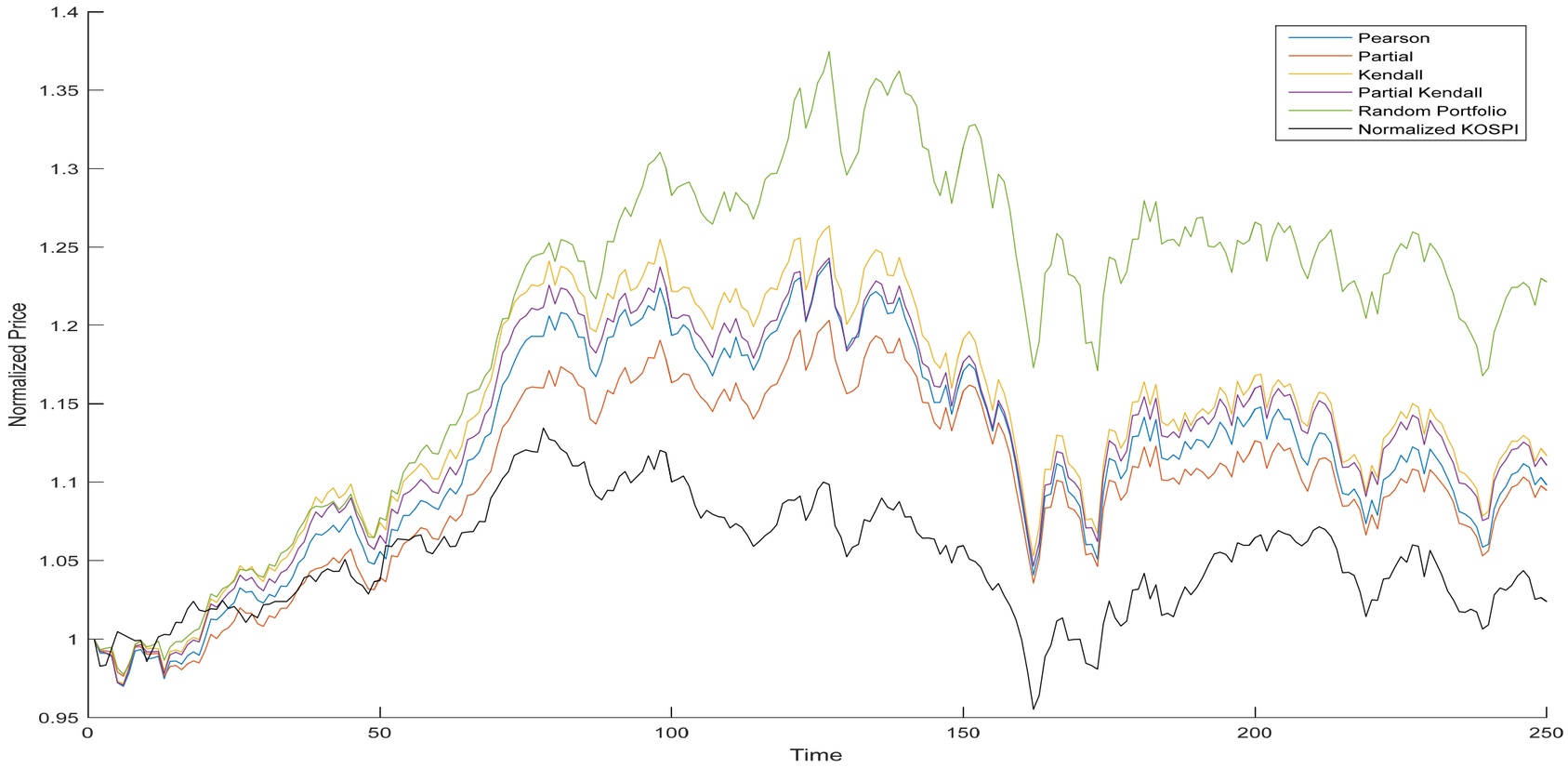

Figure 27: Realizations of trading simulation using dissimilarity matrices in 2015

Table 13: Mean annual return, standard deviation, and return-to-risk ratio for trading simulation

	Random	Large Cap	Pearson	Partial	Kendall	P. Kendall
Annual Return	0.24	0.06	0.11	0.13	0.12	0.12
Annual Std	0.20	0.14	0.18	0.18	0.19	0.18
Return / Risk	1.24	0.40	0.60	0.75	0.65	0.64

Each portfolios and their clusters were analyzed to determine the potential causes of underperformance. Since the firms were clustered based on their stock returns, it would be naïve to determine their future performance using their historical stock returns only. A good historical return doesn't necessarily guarantees the future performance. An additional parameter, earnings, was introduced as a secondary measure for performance of the firms in a cluster. Earnings are one of the items in the income statement that are closely tied to the firm's fundamental. A firm with positive earnings is very likely to be in a good condition though negative earnings doesn't necessarily mean that the firm is struggling.

Table 14 summarizes the three year performance of Pearson Clusters measured by stock returns and earnings. For each cluster, size, or the number of firms in the cluster, was reported. The average stock return was presented in percentage. The number of stock which gained in price and the number of stock whose price dropped were counted separately. Average earnings

is the mean value of firms' three year earnings. Average change in earnings is the percentage difference between the 2014 earnings and 2012 earnings. Since three years' worth of data were used, there are three annual earnings available. Hence, year to year earnings growth was computed twice for each firms and the number of firms with positive growth was counted. Number of Positive YoY, an abbreviation for year to year, = 2 counts the number of firms with two consecutive growth in annual earnings and Number of Positive YoY = 1 refers to the case when a firm saw growth in earnings once and shrink in the other. Number of Positive YoY = 0 is for the one with two consecutive drops in earnings.

The contents of the table provide some insights regarding the firms in the clusters. The average stock return for the firms in Cluster 1 was almost 0 which implies that those 25 firms were relatively losing in the past three years. Also, their earnings suggests they may be struggling in their business operation as well. Cluster 2 only has three firms assigned to it. Given that 200 firms were considered initially, a cluster of three might be a collective outlier. Cluster 3 had good years in terms of their stock return but they weren't doing well when it comes to earnings. Clusters 4 and 5 showed good performances for both stock return and earnings growth. Cluster 6 had a moderate period in terms of both stock return and earnings growth but note that there were more firms that lost its stock value than the ones

Table 14: Analysis of Pearson clusters

Performance in 2012 ~ 2014	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
Size	25	3	15	26	77	9	18	27
Avg. Stock Return %	0%	−19%	69%	40%	93%	9%	15%	−41%
Number of ΔStock > 0	10	1	9	19	66	3	9	2
Number of ΔStock < 0	15	2	6	7	11	6	8	25
Avg. Earnings	4.64E+12	7.70E+11	8.97E+11	9.18E+11	5.55E+11	5.43E+12	1.64E+12	1.37E+12
Avg. Change in Earnings %	−82%	48%	−62%	9%	30%	7%	16%	−52%
Number of Positive YoY = 2	7	1	2	6	25	3	1	3
Number of Positive YoY = 1	12	1	6	12	37	3	11	11
Number of Positive YoY = 0	6	1	7	8	15	3	6	13

gained in value. Cluster 7 had a reasonably good gain in stock but there were more firms doing badly in terms of earnings. Cluster 8 clearly had a bad period and lost both stock value and earnings. It would be interesting to see if the future performance can be improved by removing a historically bad cluster. The return, standard deviation and return-to-risk ratio of the original cluster and portfolios created using the original minus a cluster, which is denoted by the same cluster number from table 14, were summarized in table 15. Clusters 1 was thought to be a bad cluster due to their poor historical performance, but when they were removed, the overall performances of portfolio were damaged. Cluster 2 which was considered an outlier cluster, was removed, and the portfolio's return and risk were improved. Clusters 3, 6, and 7 gave mixed signals with either stock or earnings was good and the other didn't fare as much. Removing Cluster 3 was detrimental to the portfolio's risk and return, but removing Clusters 6 and 7 ended up with an improved return-to-risk ratio. Cluster 8, which thought to be the worst performing cluster, gave a confusing result where removing it resulted in a better solution than the original one.

The analysis so far was done in an *ex ante* manner. The hypothesis of identifying underperforming clusters was made using historical data only and was blindly tested without knowing what actually happened in the investment period.

Table 15: Pearson clusters with a cluster removed

	Pearson	Pearson without Cluster 1	Pearson without Cluster 2	Pearson without Cluster 3	Pearson without Cluster 4	Pearson without Cluster 5	Pearson without Cluster 6	Pearson without Cluster 7	Pearson without Cluster 8
Annual Return	0.11	0.06	0.13	0.08	0.09	0.02	0.12	0.11	0.09
Annual Std	0.18	0.17	0.17	0.17	0.17	0.16	0.17	0.16	0.17
Return / Risk	0.60	0.36	0.73	0.46	0.50	0.15	0.69	0.68	0.55
% Improvement		−39%	22%	−23%	−16%	−74%	16%	13%	−8%

Table 16: Performance of Pearson clusters during investment period

Performance on 2015	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
Avg. Stock Return %	26%	−28%	13%	7%	50%	−14%	−9%	5%
Number of ΔStock > 0	15	0	11	13	53	2	3	13
Number of ΔStock < 0	10	3	4	13	23	7	15	14
Avg. Earnings	1.50E+12	3.02E+11	3.54E+11	3.46E+11	3.41E+11	1.51E+12	6.41E+11	3.27E+11
Avg. Change in Earnings %	21%	8%	64%	28%	13%	−9%	73%	−315%
Number of ΔEarnings > 0	13	1	12	16	50	3	12	10
Number of ΔEarnings < 0	12	2	3	10	27	6	6	17

Unlike the real world, the data for investment period is available to compare the hypothesis to the actual data. Table 16 gives an outline of what really happened in 2015. Table 16 was prepared in the same manner with table 14 but there was only one annual earnings period so the earnings growth only considered if it was positive or negative. Cluster 1 and 8, which were the worst clusters before, had successfully improved their stock return and earnings which explains why removing these two resulted in declined performance of portfolio. The stock price of Cluster 2, the outlier cluster, dropped significantly which is consistent with the finding that removing Cluster 2 brought a higher return for the portfolio.

Clusters 3, 6 and 7 had mixed signals from stock return and earnings, and gave mixed performances. Clusters 4 and 5 had a good historical record and performed well during the investment period. Contrary to the hypothesis made before that it was unlikely for Cluster 8 to turnaround, the average stock return was positive and it was able to regain some of the stock value lost in the previous period, though more than half of the stocks in the cluster lost their value.

To further study the different cases, other clustering portfolios were analyzed in the same manner. Table 17, 18, and 19 are the same tables with the previous three except that they are based on the Partial clusters. The Partial clusters also showed a range of different behaviors. Clusters 1, 2, 8 and 9 showed mediocre

performance both in terms of stock return and earnings. Clusters 3 and 4 were too small to be a part of the portfolio. Clusters 5, 6, and 7 had substantial gain in stock price but Cluster 5 had weak earnings. Cluster 10 had a very bad few years with both stock price and earnings tumbled over the period. When a cluster was removed, some results were predictable but others weren't. As it is stated in table 18, removing the best performing cluster, Cluster 7, resulted in a declined performance while removing the worst performing cluster, Cluster 10, improved the quality of the portfolio.

However, by removing Clusters 5 and 6 which had a good gain in stock price before, the portfolio return was improved which supports the idea that the historical performance of stock alone is not a good predictor of future performance.

Table 19 explains what happened in 2015. Clusters 1 and 10, both of which were underperformers in the previous period, were able to turnaround and generated positive stock returns which is why removing these two resulted in declined performance of portfolio. Clusters 5, 6 and 7 maintained their superior performances while clusters 8 and 9 didn't fare this time.

The tables of Pearson and Partial clusters concludes that the past stock performance may not serve as a good indicator for its future performance but when both stock and earnings were healthy, it remained strong for the next period.

Table 17: Analysis of Partial clusters

Performance in 2012 ~ 2014	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Size	11	14	1	2	15	26	77	9	18	27
Avg. Stock Return %	−4%	4%	−38%	−10%	69%	40%	93%	9%	15%	−41%
Number of ΔStock > 0	4	6	0	1	9	19	66	3	9	2
Number of ΔStock < 0	7	8	1	1	6	7	11	6	8	25
Avg. Earnings	1.66E+11	8.15E+12	4.91E+11	9.10E+11	8.97E+11	9.18E+11	5.55E+11	5.43E+12	1.64E+12	1.37E+12
Avg. Change in Earnings %	−25%	−126%	2%	70%	−62%	9%	30%	7%	16%	−52%
Number of Positive YoY = 2	2	5	0	1	2	6	25	3	1	3
Number of Positive YoY = 1	7	5	1	0	6	12	37	3	11	11
Number of Positive YoY = 0	2	4	0	1	7	8	15	3	6	13

Table 18: Partial clusters with a cluster removed

	Partial	Partial without Cluster 1	Partial without Cluster 2	Partial without Cluster 3	Partial without Cluster 4	Partial without Cluster 5	Partial without Cluster 6	Partial without Cluster 7	Partial without Cluster 8	Partial without Cluster 9	Partial without Cluster 10
Annual Return	0.13	0.10	0.18	0.16	0.16	0.16	0.17	0.11	0.20	0.19	0.18
Annual Std	0.18	0.16	0.17	0.17	0.17	0.17	0.17	0.16	0.17	0.17	0.17
Return / Risk	0.75	0.65	1.08	0.90	0.90	0.96	1.00	0.69	1.13	1.14	1.06
%Improvement		−13%	44%	20%	20%	29%	33%	−8%	52%	52%	42%

Table 19: Performance of Partial Clusters during investment period

Performance on 2015	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Avg. Stock Return %	56%	2%	−8%	−38%	13%	7%	50%	−14%	−9%	5%
Number of ΔStock > 0	10	5	0	0	11	13	53	2	3	13
Number of ΔStock < 0	1	9	1	2	4	13	23	7	15	14
Avg. Earnings	8.82E+10	2.61E+12	2.22E+11	3.42E+11	3.54E+11	3.46E+11	3.41E+11	1.51E+12	6.41E+11	3.27E+11
Avg. Change in Earnings %	106%	−45%	27%	−2%	64%	28%	13%	−9%	73%	−315%
Number of ΔEarnings > 0	5	8	1	0	12	16	50	3	12	10
Number of ΔEarnings < 0	6	6	0	2	3	10	27	6	6	17

Note that the contents of table 14 and table 17 were very similar to each other. As it was mentioned in the previous section, correlation estimation over three years yielded a similar cluster. In this case, Clusters 3 to 8 from Pearson clusters and 5 to 10 from the Partial clusters were identical. Also, Cluster 1 from Pearson clusters was the sum of clusters 1 and 2 from the Partial clusters and Pearson's Cluster 3 was split into Clusters 3 and 4 of the Partial clusters. This finding is consistent with the previous conjecture that a correlation coefficient and its partial counterpart only vary in detail.

Kendall clustering portfolio and the Partial Kendall clustering portfolio were analyzed in the same manner. Tables 20, 21 and 22 were prepared for Kendall clustering portfolio and tables 23, 24 and 25 were for the Partial Kendall clustering portfolio. The overall results were similar that a cluster with good stock and earnings performances did well in the investment period and removing it resulted in declined return and return-to-risk ratio of portfolio. A cluster of inferior stocks were found in both matrices and though some of them were able to make a turnaround, they only showed a mediocre performance. The smallest cluster consisted of six firms for both portfolios and the cluster's performance wasn't good and removing it resulted in an improvement for the portfolios.

Table 20: Analysis of Kendall clusters

Performance in 2012 ~ 2014	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Size	6	28	108	23	21	14
Avg. Stock Return %	−20%	53%	74%	−42%	5%	−3%
Number of ΔStock > 0	2	13	89	1	10	4
Number of ΔStock < 0	4	15	19	22	10	10
Avg. Earnings	1.18E+12	6.58E+11	7.10E+11	1.08E+12	2.14E+12	1.01E+13
Avg. Change in Earnings %	49%	−90%	15%	−56%	35%	−14%
Number of Positive YoY = 2	2	7	29	3	4	3
Number of Positive YoY = 1	3	12	54	8	11	5
Number of Positive YoY = 0	1	9	25	12	6	6

Table 21: Kendall Clusters with a cluster removed

	Kendall	Kendall without Cluster 1	Kendall without Cluster 2	Kendall without Cluster 3	Kendall without Cluster 4	Kendall without Cluster 5	Kendall without Cluster 6
Annual Return	0.12	0.12	0.07	0.04	0.10	0.14	0.13
Annual Std	0.19	0.19	0.19	0.19	0.18	0.18	0.20
Return / Risk	0.65	0.61	0.37	0.23	0.57	0.75	0.66
%Improvement		−6%	−43%	−64%	−13%	15%	0.5%

Table 22: Performance of Kendall Clusters during investment period

Performance on 2015	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Avg. Stock Return %	2%	23%	38%	8%	−10%	−6%
Number of ΔStock > 0	4	18	70	11	3	4
Number of ΔStock < 0	2	10	37	12	18	10
Avg. Earnings	3.04E+11	3.81E+11	3.55E+11	3.10E+11	6.91E+11	2.89E+12
Avg. Change in Earnings %	−8%	79%	17%	−322%	5%	−52%
Number of ΔEarnings > 0	2	19	69	10	11	6
Number of ΔEarnings < 0	4	9	39	13	10	8

Table 23: Analysis of Partial Kendall clusters

Performance in 2012 ~ 2014	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Size	80	27	6	29	22	22	14
Avg. Stock Return %	89%	28%	−20%	54%	−45%	5%	−3%
Number of ΔStock > 0	69	19	2	14	0	11	4
Number of ΔStock < 0	11	8	4	15	22	10	10
Avg. Earnings	3.74E+11	1.70E+12	1.18E+12	6.60E+11	9.72E+11	2.20E+12	1.01E+13
Avg. Change in Earnings %	32%	−36%	49%	−86%	−58%	33%	−14%
Number of Positive YoY = 2	25	3	2	8	3	4	3
Number of Positive YoY = 1	39	15	3	12	8	11	5
Number of Positive YoY = 0	16	9	1	9	11	7	6

Table 24: Partial Kendall clusters with a cluster removed

	P. Kendall	P. Kendall without Cluster 1	P. Kendall without Cluster 2	P. Kendall without Cluster 3	P. Kendall without Cluster 4	P. Kendall without Cluster 5	P. Kendall without Cluster 6	P. Kendall without Cluster 7
Annual Return	0.12	0.05	0.14	0.14	0.11	0.14	0.16	0.16
Annual Std	0.18	0.18	0.19	0.18	0.19	0.18	0.18	0.19
Return / Risk	0.64	0.30	0.70	0.77	0.61	0.80	0.92	0.85
%Improvement		−54%	9%	21%	−5%	24%	44.4%	33.5%

Table 25: Performance of Partial Kendall clusters during investment period

Performance on 2015	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Avg. Stock Return %	49%	5%	2%	22%	8%	−9%	−6%
Number of ΔStock > 0	56	14	4	18	10	4	4
Number of ΔStock < 0	23	13	2	11	12	18	10
Avg. Earnings	1.54E+11	9.51E+11	3.04E+11	3.78E+11	2.72E+11	7.11E+11	2.89E+12
Avg. Change in Earnings %	21%	8%	−8%	77%	−337%	5%	−52%
Number of ΔEarnings > 0	54	14	2	20	9	12	6
Number of ΔEarnings < 0	26	13	4	9	13	10	8

4.4 Long-Term Portfolio Management and Portfolio Strategy

A detailed analysis of a single period from the previous section suggests that it is possible to outperform the market using clustering portfolios and the performance can be improved with the trimming process in which underperforming clusters were removed from the portfolio construction step. In this section, the trimming process was formalized and portfolios were managed over two decades. So far, the analyses were done using Korean data but due to limited amount of data available, it is impossible to study the long term performance of clustering portfolio. Therefore, the stocks listed in the New York Stock Exchange and NASDAQ were used. The largest 200 firms in the US market from 1989 to 2015 were used to form clusters and their earnings per share data, along with stock returns, were used to determine the performance of clusters. 27 years' worth of data were available but the first year was used to compute the earning's growth rate; therefore, 26 years' worth of data were used for portfolio management. The Standard & Poor's 500, which consists of the largest 500 companies listed in the New York Stock Exchange and NASDAQ, is the benchmark index in this case. Unlike the previous section where each portfolios and clusters were studied in detail, it is impossible to track down all

clusters. Clusters were created every period along a moving window and the portfolio was rebalanced every period. The size of moving window is equivalent to the number of years used to estimate correlation coefficients. An increment of moving window is equivalent to the investment period and the rebalancing period. In this study, rebalancing period, the period in which the available cash was redistributed, was varied and its effect on the portfolio's performance was studied. Three rebalancing periods, 3 month, 6 month, and 12 month were considered. The number of year used to compute correlation was varied as well. Three periods of 1 year, 3 year and 5 year of data were used for correlation estimation. Thus, total of 9 portfolios were constructed for each correlation coefficients. Figure 28 plots the sample realization of Pearson portfolio with 5 year of estimation period and 6 month of rebalancing period.

Unlike the previous result where the market index was easy to beat, the market's performance was very close to the clustering portfolio and outperformed the clustering portfolio in the long run. Three correlation estimation periods implies varying periods of S&P500 were used as the benchmark. The detailed performances during each periods are shown in table 26. The S&P500 gained about 8% annually and the return-to-risk ratio was around 0.41 for the entire period of study.

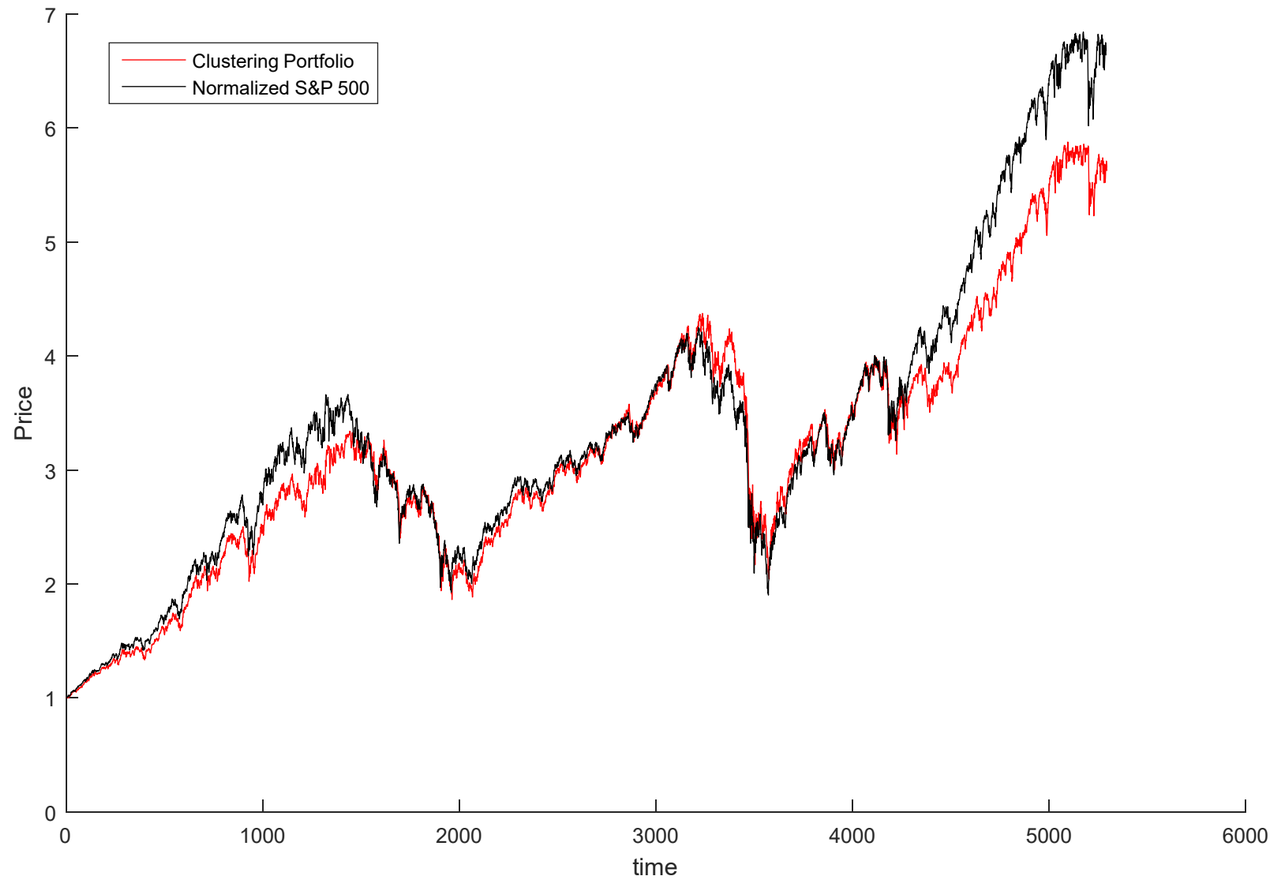

Figure 28: Realizations of clustering portfolio in the US market

The large cap benchmark was reconstructed for this section as well. Unlike a single period study of previous section, the constituents of large cap index inevitably change over time. In this study, a large cap index was made of the largest 20 firms in the NYSE and NASDAQ, and the reconstitution was done every year. Table 27 provides returns and return-to-risk ratio of the large cap index.

Table 26: Performance of S&P 500 index

	1991 ~ 2015	1993 ~ 2015	1995 ~ 2015
Annual Return	0.098	0.083	0.079
Annual Std	0.178	0.182	0.188
Return / Risk	0.552	0.458	0.417

Table 27: Performance of the large cap index

	1991 ~ 2015	1993 ~ 2015	1995 ~ 2015
Annual Return	0.054	0.044	0.054
Annual Std	0.184	0.187	0.192
Return / Risk	0.294	0.235	0.283

The large cap index fared worse than the S&P index because especially after the 2008 financial crisis, the index didn't gain much. The annual return was about 5.4% with the return-to-risk ratio of 0.28. Most of the clustering portfolios without any trimming were able to beat the large cap index but failed to beat

S&P500. The results are shown in table 28. Each column is for the different period of correlation estimation and rebalancing period. The first number with yr, which is an abbreviation of year, marks the number of year used to estimate correlation coefficients. The second number gives the rebalancing period which is either 3, 6 or 12 mo, an abbreviation of month. Note that the period is in trading days, so a month is made of 21 days and a year means 252 days. Though the performance varied widely, all portfolios had return greater than 7% and return-to-risk ratio higher than 0.38 which means many clustering portfolios were unable to beat the benchmark index. The result is similar to what happened in the previous section against the random portfolio. A regular clustering portfolio is likely to have a larger proportion of assets allocated to bad clusters so the trimming step is required to improve its performance.

Here, the portfolios were constructed using two different parameters: correlation estimation period and rebalancing period. A hypothesis was that these two variables may play a role for the portfolio's performance. However, when the return-to-risk ratio was analyzed along those two axes, no clear trend arise. There is no upward or downward trend along those two axes, and the maximum and minimum return-to-risk ratio were found in different periods for different portfolios which imply that two variables may not have a significant influence on correlation coefficients. It is interesting to note that the portfolios with 3

year correlation estimation period were always the worst for given correlation portfolios. Between two periods, rebalancing period has a smaller influence on the performance where the average standard deviation for varying rebalancing period was about 0.034 while the same value for the correlation estimation period was about 0.0691.

The purpose of the trading analysis is to show that by using information embedded in clusters one is able to not only outperform the market but able to improve the performance of portfolio using the trimming process. The previous section showed that a cluster of good firms remained strong for the next period while a small cluster or a cluster with bad historical performance gave poor return. In this section, portfolios were trimmed down under a clear rule so that it can be done with an algorithm. Each rule was applied separately to study their individual effect.

Five rules were established based on the previous result. The first rule was to remove a cluster with the worst stock performance over the estimation period. The second rule was to remove clusters with a greater number of firms with negative stock return than the number of positive stock return. For the third approach, a cluster with the largest relative number of firms with more period of negative earnings per share was removed. The fourth rule was to remove clusters with a relatively large number of firms whose earning dropped.

Table 28: Performance of clustering portfolios

Pearson	1 yr / 3 mo	1 yr / 6 mo	1 yr / 12 mo	3 yr / 3 mo	3 yr / 6 mo	3 yr / 12 mo	5 yr / 3 mo	5 yr / 6 mo	5 yr / 12 mo
Return	0.079	0.074	0.080	0.079	0.080	0.080	0.086	0.085	0.079
Std	0.178	0.178	0.174	0.187	0.184	0.180	0.189	0.186	0.186
r/s	0.442	0.417	0.459	0.422	0.434	0.447	0.454	0.455	0.426

Partial	1 yr / 3 mo	1 yr / 6 mo	1 yr / 12 mo	3 yr / 3 mo	3 yr / 6 mo	3 yr / 12 mo	5 yr / 3 mo	5 yr / 6 mo	5 yr / 12 mo
Return	0.083	0.080	0.082	0.071	0.075	0.073	0.088	0.089	0.087
Std	0.185	0.178	0.176	0.177	0.177	0.175	0.183	0.181	0.180
r/s	0.452	0.450	0.463	0.402	0.424	0.419	0.481	0.492	0.483

Kendall	1 yr / 3 mo	1 yr / 6 mo	1 yr / 12 mo	3 yr / 3 mo	3 yr / 6 mo	3 yr / 12 mo	5 yr / 3 mo	5 yr / 6 mo	5 yr / 12 mo
Return	0.082	0.088	0.078	0.072	0.076	0.074	0.085	0.082	0.077
Std	0.178	0.176	0.179	0.186	0.186	0.184	0.189	0.187	0.189
r/s	0.462	0.499	0.436	0.388	0.407	0.401	0.448	0.440	0.407

P. Kendall	1 yr / 3 mo	1 yr / 6 mo	1 yr / 12 mo	3 yr / 3 mo	3 yr / 6 mo	3 yr / 12 mo	5 yr / 3 mo	5 yr / 6 mo	5 yr / 12 mo
Return	0.092	0.090	0.081	0.072	0.071	0.072	0.080	0.075	0.076
Std	0.180	0.178	0.180	0.187	0.186	0.182	0.192	0.191	0.188
r/s	0.511	0.507	0.450	0.385	0.383	0.395	0.418	0.393	0.402

Though similar to the third rule, the fourth rule considers the growth rate of earnings. The fifth rule was to remove outliers. Similar to the previous section, the portfolio construction process did remove a small cluster with the number of constituents smaller than the size of portfolio, which is 20, divided by the number of cluster. This rule is designed to remove all clusters with the number of constituents smaller than a predetermined value. Portfolios were constructed under one rule at a time and they were studied for all correlation estimation periods and rebalancing periods, and compared against the original portfolios with no rule applied. Table 29 reports the result of using the first rule, where a cluster of the worst average stock return was removed. The reported numbers are return-to-risk ratio in which a greater value implies a better performance. Regular portfolio refers to the clustering portfolio with no rule applied and trimmed portfolio is the one with a cluster trimmed down by the rule. Detailed values of returns and standard deviations are available in the Appendix D. When a cluster of previously worst stock performance was removed, there was no clear evidence of improved performance. If the rule was able to consistently identify an inferior cluster then the resulting portfolio's return-to-risk ratio should be greater than that of the regular one. However, each correlation estimation periods reacted differently and there is no observable trend along the rebalancing period. All correlation coefficients gave varying results.

Table 29: Performance of portfolios with the worst stock performance cluster removed

	1 yr / 3 mo	1 yr / 6 mo	1 yr / 12 mo	3 yr / 3 mo	3 yr / 6 mo	3 yr / 12 mo	5 yr / 3 mo	5 yr / 6 mo	5 yr / 12 mo
Pearson									
Regular	0.442	0.417	0.459	0.422	0.434	0.447	0.454	0.455	0.426
Trimmed	0.424	0.412	0.436	0.380	0.469	0.436	0.434	0.452	0.430
Partial									
Regular	0.452	0.450	0.463	0.402	0.424	0.419	0.481	0.492	0.483
Trimmed	0.434	0.472	0.467	0.421	0.398	0.429	0.454	0.498	0.479
Kendall									
Regular	0.462	0.499	0.436	0.388	0.407	0.401	0.448	0.440	0.407
Trimmed	0.416	0.480	0.406	0.366	0.473	0.393	0.415	0.456	0.415
P. Kendall									
Regular	0.511	0.507	0.450	0.385	0.383	0.395	0.418	0.393	0.402
Trimmed	0.451	0.524	0.420	0.351	0.441	0.402	0.439	0.401	0.429

It shouldn't come as a surprise that there is no observable trend here because there is no reason a previous performance of stock should guarantee its future return. The second rule was applied to the portfolios where clusters with a relative large number of firms whose stock price dropped is removed. The first rule only removed a single cluster with the worst performance. In this case, however, all clusters which has more stocks whose price dropped than stocks whose price gained were removed. It is difficult for the first rule to determine the number of bad clusters to be removed because it is unknown how many clusters were created and how bad the second worst cluster is. For the second rule, on the other hand, one can systematically remove more than one cluster by removing clusters in which more than half of the constituents were stock losers. A problem may arise during a financial crisis because it is highly likely that all clusters would have more firms whose value dropped in this period. Then the algorithm may proceed and remove all clusters which defeat the purpose of trimming. Therefore, a condition was set that if all clusters satisfied the second rule, then the trimming step was skipped and no cluster was removed. Given that the period of study includes the 2008 financial crisis this constraint was necessary. Also, the average stock return of the first rule is vulnerable to an outlier and a firm with large drop in its stock price may have a significant influence on the outlook of a cluster. Instead of looking into the value of stock return, the second rule

counts the number of firms with negative stock return which should be more robust against an extreme value of negative stock return.

The performance of portfolios trimmed by the second rule is provided in table 30. The Appendix E provides the returns and standard deviations. Since it was based on the stock returns, the result was similar to that of the first rule where no clear pattern was observed. There were few cases where the clustering portfolio's performance improved but there were larger number of cases where the performance declined. Apparently, stock return alone is not a good predictor for the future performance.

Earnings were adopted as a secondary measure of performance in the previous section. A similar item, earnings per share was adopted for the same purpose. Similar to the earnings, EPS is considered a good predictor of health and longevity of a firm. Using EPS, the third rule was constructed in a similar manner to compute the number of positive YoY in the previous section. For each firm, the EPS data were available annually so depending on the correlation estimation period, up to five EPSs were available. The third rule counts the number of firms whose earnings were negative for more than half of the correlation estimation period, and a cluster with the largest proportion of such firms was removed. The third rule suffers from a similar problem to that of the first rule. It is unclear that how many clusters should be removed. One can arbitrarily set a threshold level but seeing how

Table 30: Performance of portfolios with clusters of a larger number of firms whose price dropped removed

	1 yr / 3 mo	1 yr / 6 mo	1 yr / 12 mo	3 yr / 3 mo	3 yr / 6 mo	3 yr / 12 mo	5 yr / 3 mo	5 yr / 6 mo	5 yr / 12 mo
Pearson									
Regular	0.442	0.417	0.459	0.422	0.434	0.447	0.454	0.455	0.426
Trimmed	0.423	0.487	0.440	0.247	0.414	0.423	0.481	0.385	0.435
Partial									
Regular	0.452	0.450	0.463	0.402	0.424	0.419	0.481	0.492	0.483
Trimmed	0.451	0.468	0.455	0.391	0.400	0.454	0.483	0.424	0.501
Kendall									
Regular	0.462	0.499	0.436	0.388	0.407	0.401	0.448	0.440	0.407
Trimmed	0.523	0.554	0.421	0.254	0.457	0.381	0.440	0.411	0.448
P. Kendall									
Regular	0.511	0.507	0.450	0.385	0.383	0.395	0.418	0.393	0.402
Trimmed	0.535	0.581	0.380	0.306	0.409	0.355	0.440	0.351	0.437

earnings can suffer from one time event such as global crisis or national tragedy, a fixed level cannot accommodate all time. Hence, a single cluster with relatively the biggest earnings losers was removed from portfolio construction step. The resulting portfolios' performance is recorded in table 31 with details summarized in the Appendix F. Unlike to the previous results where clusters were trimmed using stock return, majority of portfolios, except for the ones created using Partial dissimilarity matrix, saw improvement in performance. One can also notice that 2 out of 3 failures in improvement was found in 1 year correlation estimation period. Given that a short-term earnings is not a suitable measure for long-term performance it is not surprising that those two cases didn't perform so well. On the contrary, except for the Partial correlation cases, all clustering portfolios showed improved performance with 5 year correlation estimation period which supports the idea that a long-term earnings are useful to predict the long-term performance of the firm.

The next rule is another trimming process based on EPS. It is an EPS equivalent of the second rule in which all clusters with more than half of the firms which saw shrank in earnings were removed. Note that contrary to the third rule where it considered the earnings itself, the fourth rule considers the growth rate of earnings. A significant difference is that the earnings growth rate can capture the direction of recent changes.

Table 31: Performance of portfolios with a cluster of the worst EPS performance removed

	1 yr / 3 mo	1 yr / 6 mo	1 yr / 12 mo	3 yr / 3 mo	3 yr / 6 mo	3 yr / 12 mo	5 yr / 3 mo	5 yr / 6 mo	5 yr / 12 mo
Pearson									
Regular	0.442	0.417	0.459	0.422	0.434	0.447	0.454	0.455	0.426
Trimmed	0.472	0.474	0.478	0.441	0.497	0.475	0.505	0.486	0.436
Partial									
Regular	0.452	0.450	0.463	0.402	0.424	0.419	0.481	0.492	0.483
Trimmed	0.440	0.459	0.441	0.413	0.441	0.451	0.452	0.468	0.478
Kendall									
Regular	0.462	0.499	0.436	0.388	0.407	0.401	0.448	0.440	0.407
Trimmed	0.500	0.549	0.439	0.441	0.431	0.380	0.524	0.496	0.443
P. Kendall									
Regular	0.511	0.507	0.450	0.385	0.383	0.395	0.418	0.393	0.402
Trimmed	0.501	0.550	0.422	0.416	0.421	0.428	0.493	0.467	0.444

A firm may have a negative earnings during a period of time but it may gain back and reduce the magnitude of loss in the next period. The third rule would see both periods as the negative earnings and count up a bad firm when in fact the firm was recovering. Similarly, a historical good firm with positive earnings may be losing its ground and realized smaller, albeit positive, earnings in the following period. Regardless of the direction the firm was going, the third rule would consider the firm in a good status which could be misleading. The fourth rule is designed to capture this shortcoming of the third rule. It counts how many times each firm saw a growth in earnings and if it's less than the half of the correlation estimation period, then the firm is considered bad. If the number of bad firms was greater than half of the size of cluster, then the cluster is removed. All clusters which satisfied this condition were removed and the portfolios were constructed. The result is summarized in table 32 and the details are provided in the Appendix G. Regardless of rebalancing period, the performance of portfolio was improved when the correlation estimation period was long; that is, except for the case when the estimation period was 1 year, the performance improved. This is in line with the result of the third rule where the long-term estimation of earnings gave a positive outcome. With the estimation period of 1 year, half of the portfolios saw improvement while the other half didn't, which implies that a short-term earnings growth is not a good predictor

Table 32: Performance of portfolios with clusters of a larger number of firms whose earnings shrank removed

	1 yr / 3 mo	1 yr / 6 mo	1 yr / 12 mo	3 yr / 3 mo	3 yr / 6 mo	3 yr / 12 mo	5 yr / 3 mo	5 yr / 6 mo	5 yr / 12 mo
Pearson									
Regular	0.442	0.417	0.459	0.422	0.434	0.447	0.454	0.455	0.426
Trimmed	0.504	0.474	0.526	0.456	0.464	0.463	0.486	0.467	0.431
Partial									
Regular	0.452	0.450	0.463	0.402	0.424	0.419	0.481	0.492	0.483
Trimmed	0.444	0.427	0.422	0.439	0.484	0.471	0.508	0.533	0.490
Kendall									
Regular	0.462	0.499	0.436	0.388	0.407	0.401	0.448	0.440	0.407
Trimmed	0.472	0.493	0.465	0.447	0.455	0.452	0.488	0.454	0.414
P. Kendall									
Regular	0.511	0.507	0.450	0.385	0.383	0.395	0.418	0.393	0.402
Trimmed	0.491	0.512	0.488	0.443	0.446	0.443	0.459	0.410	0.416

of the future performance. The four rules studied so far strongly suggests that the historical performance of stock return is not a good parameter for the trimming process, but the earnings, when combined with a long enough estimation period, are a good candidate to trim down clusters.

The fifth rule which looked promising in the previous section was the removal of outlier clusters. A cluster with a small number of firms can be said that these firms were different from others and failed to create a major group. Since the clustering is based on stock returns, it is highly likely that the returns of these firms are on the either end of spectrum. In the previous analysis of Korean data, removal of such cluster resulted in improved performance but it is still unknown whether such result can be delivered consistently. In order to proceed and test this rule, a threshold size to remove clusters needs to be predetermined. Since there is no clear line between outlier cluster and non-outlier cluster, two different size of 5 and 10 were chosen. For example, if the threshold was 5, all clusters with the number of constituents fewer than or equal to 5 are removed. Table 33 summarizes the both cases of outlier clusters removed with details provided in the Appendix H. Contrary to the previous result where the return-to-risk ratio was improved upon removing outlier clusters, results were mixed but there were more portfolios with declined performance. Outliers, though they may be a group of underperformers sometime, were not always

Table 33: Performance of portfolios with outlier cluster removed

	1 yr / 3 mo	1 yr / 6 mo	1 yr / 12 mo	3 yr / 3 mo	3 yr / 6 mo	3 yr / 12 mo	5 yr / 3 mo	5 yr / 6 mo	5 yr / 12 mo
Pearson									
Regular	0.442	0.417	0.459	0.422	0.434	0.447	0.454	0.455	0.426
Trimmed (5)	0.447	0.431	0.463	0.437	0.435	0.432	0.449	0.459	0.431
Trimmed (10)	0.404	0.393	0.418	0.410	0.449	0.412	0.414	0.435	0.393
Partial									
Regular	0.452	0.450	0.463	0.402	0.424	0.419	0.481	0.492	0.483
Trimmed (5)	0.462	0.451	0.450	0.410	0.411	0.426	0.462	0.472	0.478
Trimmed (10)	0.433	0.449	0.371	0.448	0.471	0.434	0.463	0.445	0.430
Kendall									
Regular	0.462	0.499	0.436	0.388	0.407	0.401	0.448	0.440	0.407
Trimmed (5)	0.470	0.500	0.436	0.392	0.410	0.401	0.442	0.435	0.410
Trimmed (10)	0.456	0.501	0.419	0.400	0.421	0.387	0.435	0.439	0.397
P. Kendall									
Regular	0.511	0.507	0.450	0.385	0.383	0.395	0.418	0.393	0.402
Trimmed (5)	0.521	0.516	0.449	0.384	0.383	0.395	0.415	0.390	0.416
Trimmed (10)	0.426	0.489	0.390	0.307	0.432	0.378	0.391	0.437	0.417

bad clusters and their removal was often detrimental to portfolio's performance.

The five rules identified were studied so far and it was found that the rules based on stock returns were unreliable to determine a weak cluster. Two rules based on earnings per share, on the other hand, were able to improve the performance consistently. One can hypothesize that combination of these two rules, the third rule and the fourth rule, may result in even better performance. As a final step of clustering portfolio trimming analysis, both rules were applied to the clustering portfolios and their performances were analyzed. Since those two are not always overlapping combining two strategies may deliver a better result.

Table 34 reports the result of combined strategy. The returns and standard deviations were reported in the Appendix I. One may notice that except for two occasions found in 1 year correlation estimation period, the trimmed portfolios were able to outperform all the regular clustering portfolios. When compared against the individual strategies, except for the Partial clusters, the combined strategy outperformed most of them especially when the estimation period was 3 or 5 years. The result signifies that the clustering analysis of stock returns were able to separate the firms in terms of their fundamental differences, and using the long term analysis of earnings, one can systematically identify a group of weak firms. After eliminating

such clusters, a stock portfolio made of firms from different clusters were able to benefit from diversification effect and outperformed the market index. An example of the combined results are summarized and plotted in figure 29. Figure 29 shows the plots of Pearson portfolios created using correlation estimation period of 5 years and rebalancing period of 3 months. The regular clustering portfolio, portfolio trimmed with the third rule, portfolio trimmed with the fourth rule, and portfolio trimmed with the combined strategy were plotted with the benchmark index, S&P500. It is visually clear that the trimmed clustering portfolio did outperform the benchmark.

Table 34: Performance of portfolios with clusters of the worst EPS and weak EPS growth removed

	1 yr / 3 mo	1 yr / 6 mo	1 yr / 12 mo	3 yr / 3 mo	3 yr / 6 mo	3 yr / 12 mo	5 yr / 3 mo	5 yr / 6 mo	5 yr / 12 mo
Pearson									
Regular	0.442	0.417	0.459	0.422	0.434	0.447	0.454	0.455	0.426
Trimmed	0.532	0.540	0.557	0.522	0.570	0.515	0.563	0.516	0.442
Partial									
Regular	0.452	0.450	0.463	0.402	0.424	0.419	0.481	0.492	0.483
Trimmed	0.481	0.452	0.428	0.467	0.507	0.486	0.489	0.516	0.489
Kendall									
Regular	0.462	0.499	0.436	0.388	0.407	0.401	0.448	0.440	0.407
Trimmed	0.514	0.555	0.490	0.524	0.488	0.449	0.585	0.520	0.462
P. Kendall									
Regular	0.511	0.507	0.450	0.385	0.383	0.395	0.418	0.393	0.402
Trimmed	0.514	0.544	0.433	0.483	0.490	0.489	0.566	0.524	0.466

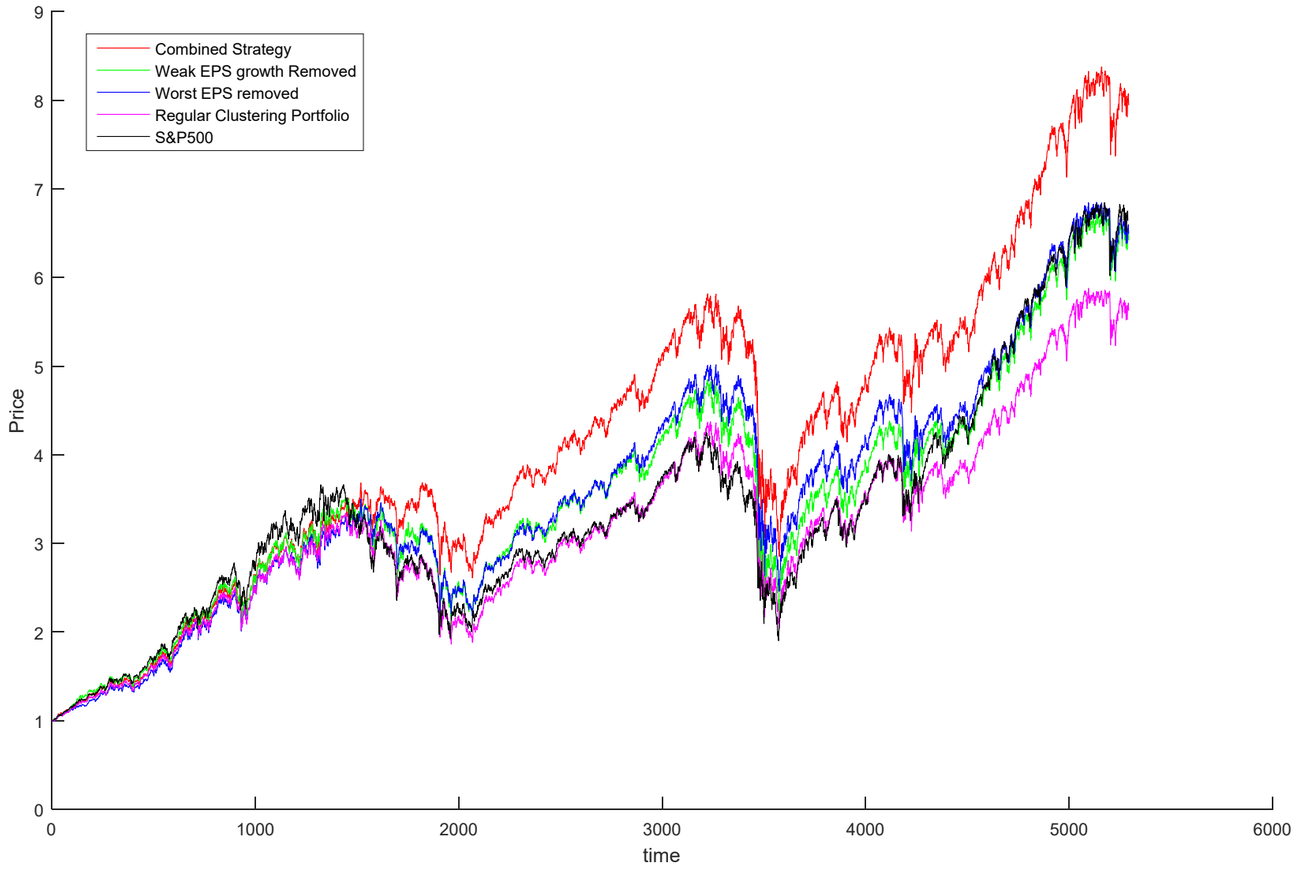

Figure 29: Realizations of 5 yr / 3 mo Pearson portfolios using different strategies

Chapter 5

Conclusion

5.1 Summary and Implications

For the past several decades the financial market was a subject of study from many different fields. Classical Finance and Financial Economics began the analysis of financial data using the tools found in Statistics and Econometrics. Subsequently it drew attention from Mathematics, Signal Processing and Physics and made contributions. For example, mathematical and time series models were developed to describe the financial market, and frequently observed phenomena were established as the stylized facts. However, the financial crises and unexpected events did occur even when the market was continuously under scrutiny. This fact along suggests a further effort is required from a different point of view.

This research is dedicated to better understand the financial market. Previous approaches were studied and it was found that

though studies did provide insights and frameworks, an empirical aspect of the market was relatively under-developed.

Data mining methods gained popularity in the past few decades and successfully delivered results in many fields including biology, web search and social media. The strength of this method comes from the fact that it can process large amount of data but it doesn't always ask for a strict assumption as it is the case for mathematical or statistical models. It can be used for exploratory analysis with no assumption at all. Given that the conventional wisdom regarding financial market often fail to protect the market participants from crises, it may be better to start from the scratch without any assumption to gain a new perspective.

Therefore, in this study, a set of daily closing price of stocks was used to study the financial market. Annual earnings data was also used to refine the analysis in the later chapter, but the primary analyses were all conducted using daily returns generated from daily closing prices.

Daily returns were used to estimate correlation matrices. Given that the financial market is essentially a complex system of firms and their interactions, correlation matrix of firms are likely to contain information regarding the market structure. Pearson correlation coefficients and Kendall correlation coefficients were estimated for study. These correlations were then used to generate a partial version of themselves. A partial correlation is

a correlation between two random variables when a third random variable, which thought to have an influence on the random variables of interest, is removed. The financial market is exposed to countless external factors other than firms themselves such as a growth rate of market, price of commodities, and foreign exchange rates. Though it is unlikely to be able to identify and control all external factors, one can choose a few that might have the most effect on all firms in the market. In this study, the market index, which is a good proxy for overall economy of the country, was taken as the common factor and the partial correlations were removed of its effect.

The curse of dimensionality and spurious correlation were already extensively studied, so a filtering procedure was adopted to ensure the correlation coefficients estimated were non-random. Random matrix theory was used to decompose the correlation matrices into their eigenvalues and eigenvectors, and the noise part of the matrix, which is determined by the RMT, was filtered.

The filtered correlation coefficients were transformed using the distance function for analysis and the dissimilarity matrix was created. The dissimilarity matrices were used to build hierarchical trees to take a snapshot of the financial market. Dissimilarities between a small numbers of large firms were calculated and trees were constructed to study the relative position of the firms. The dissimilarity matrices agreed in many

points such as putting financial firms together. However, they weren't entirely the same which leaves the room for a major difference when the data set becomes larger.

Clustering analysis was then performed to study a larger data set. The first step was to see if the clusters do agree with classification by industrial sector. Classification by industrial sector is a de facto standard for the market participants to categorize firms. It intuitively makes sense because the firms in the same sector are engaged in a similar business and they are exposed to the same risk factors; therefore, it is likely that their performances are correlated to each other. The clustering analysis, however, didn't group the firms by their industrial sector. Most clusters were made of firms from different sector and firms in the same sector were often assigned to different clusters.

The clustering analysis suggests a different picture for the market structure from the traditional notion. It is important to put the theory into test and portfolio analysis, where the categorization of firms by sector is a norm, is chosen as a test bed. Portfolios based on the dissimilarity matrices and clustering were constructed and their performance was compared against benchmarks. All portfolios were able to outperform the market index but some were unable to beat a random portfolio. This analysis was done in an *ex-post* manner where the market structure was studied using the historical data.

The reason market participants build a portfolio is to generate stable returns in the future. It means that the information of the future is, obviously, unavailable at the time of portfolio construction and they have to build a portfolio using historical data only. Therefore, second analysis was performed in an *ex-ante* manner. Correlation coefficients were estimated using the first 3 years' worth of data. Dissimilarity matrices were built based on these correlation coefficients and clustering analysis was performed to separate the stocks into few groups. Then, portfolios were constructed and launched for a year. The performance of the portfolios were again compared with the market index and a random portfolio. Because the market was bad the entire year, it was easy for the portfolios to beat the market. However, all clustering portfolios failed to beat the random portfolio measured by return-to-risk ratio. Given that the clustering portfolios were constructed using the exact same data set which were used to create the random portfolio, a possible explanation for the clustering portfolios to have significantly worse performance is by picking a disproportionately larger number of underperforming stocks. Since the clustering algorithm is designed to put similar objects together, it is likely that the algorithm would put the underperforming stocks in the same cluster. The portfolio construction was done by picking equal number of stocks from each clusters so if there is a cluster of inferior stocks, it is likely

that relatively large portion of inferior stocks compared to that of the entire market were added to the portfolios. The clustering portfolio and associated clusters were studied in detail and it was found that many clusters were a group of underperforming stocks and by choosing an equal number of stocks from each clusters, a clustering portfolio had a bigger proportion of underperforming stocks compared to the market index. By trimming down such clusters and removing them from the portfolio construction step, the performance of clustering portfolio was vastly improved.

Using the framework of trimmed clustering portfolio, portfolios were created in the US market and managed from 1990 to 2015. The portfolios were rebalanced periodically but the clustering portfolios were unable to beat the market index. The trimming process was formalized and when a cluster with a large number of firms with bad earnings and clusters with more firms with shrinking earnings were removed, the performance of clustering portfolios was improved. The result has significant implication for market participants, especially institutional investors. The framework provide a view of the market and firms, and separate them by fundamental difference in return, which investors seek to maximize. Also, it is able to identify groups of bad prospects and by removing them, investors can allocate their asset more efficiently to winning stocks. There are only few theory and framework available to investors that can actually help with stock

picking process so the framework proposed here can be a valuable asset for them.

5.2 Contributions

The contributions of this dissertation is as follows.

First, the Pearson and Kendall correlation coefficients and their partial version were studied with random matrix theory applied. The empirical study of the correlation matrix was usually only done for the Pearson correlation and mostly using the US data. This study utilized both Pearson and Kendall correlations and the RMT filtering was performed for a Korean data set.

Second, this research is one of the earliest attempt to create a portfolio based on clustering and to run a trading simulation. Based on the result of trading simulation, it was found that the clustering analysis was able to identify groups of inferior stocks and by removing the groups from the candidates of investment, a marked improvement in portfolio's performance was observed. Given that this result was obtained using the data only available at the time of portfolio construction, this is one of few studies which are done in the real world set up and can be quickly adopted by the market participants.

Third, the research formalized the steps to identify clusters of inferior stocks and proposed a framework to construct a clustering portfolio. Clustering portfolios were empirically tested and proved they can outperform the market index.

5.3 Limitations and Future Research

The dissertation attempted to use novel approaches and proposed a framework to construct a portfolio. In other words, it left some room for details.

One of them was methods adopted in this study. Random Matrix Theory was used for filtering procedure because previous studies suggest it is one of the best tools to filter a financial correlation matrix. However, these studies didn't consider trading simulation and there is no extensive study to determine the optimal filtering method for trading simulation. The same argument goes for the clustering analysis. The agglomerative hierarchical clustering is one of the simplest form of clustering algorithm. Though the financial data are usually not extremely big to be considered a Big Data or has many missing values and require extensive pre-processing, there could be a refined method more suitable to cluster financial data.

Four correlation coefficients were used for the analyses and portfolio construction, and their results and performances varied. A further study is required to analyze the possible cause of difference.

The research is mainly an empirical study but the data set used in this study was limited. Only the US market was studied extensively, and among the variables available in the three financial statements, only the earnings were considered. A study

of other market may provide further insight and it is highly likely that an emerging market may have very different picture from the developed markets which were studied in this research. Also, though earnings are arguably the single most important value in the three statements, many more variables are available, and some of them hold valuable information which doesn't overlap with earnings. Adding those variable may further strengthen the trimming process.

Bibliography

Ang, A., & Chen, J. (2002). Asymmetric correlations of equity portfolios. *Journal of Financial Economics, 63*(3), 443-494

Anscombe, F. J. (1973). Graphs in Statistical-Analysis. *American Statistician, 27*(1), 17-21

Arestis, P., Caporale, G. M., Cipollini, A., & Spagnolo, N. (2005). Testing for financial contagion between developed and emerging markets during the 1997 East Asian crisis. *International Journal of Finance & Economics, 10*(4), 359-367

Baillie, R. T. (1996). Long memory processes and fractional integration in econometrics. *Journal of Econometrics, 73*(1), 5-59

Barkoulas, J. T., Baum, C. F., & Travlos, N. (2000). Long memory in the Greek stock market. *Applied Financial Economics, 10*(2), 177-184

Basalto, N., Bellotti, R., De Carlo, F., Facchi, P., & Pascazio, S. (2005). Clustering stock market companies via chaotic map synchronization. *Physica a-Statistical Mechanics and Its Applications, 345*(1-2), 196-206

Bastos, J. A., & Caiado, J. (2014). Clustering financial time series with variance ratio statistics. *Quantitative Finance, 14*(12), 2121-2133

Bentes, S. R., Menezes, R., & Mendes, D. A. (2008). Long memory and volatility clustering: Is the empirical evidence consistent across stock markets? *Physica A: Statistical Mechanics and its Applications, 387*(15), 3826-3830

Bernaschi, M., Grilli, L., & Vergni, D. (2002). Statistical analysis of fixed income market. *Physica a-Statistical Mechanics and Its Applications, 308*(1-4), 381-390

Bertero, E., & Mayer, C. (1990). Structure and performance: Global interdependence of stock markets around the crash of October 1987*. *European Economic Review, 34*(6), 1155-1180

Best, M. J., & Grauer, R. R. (1991). On the Sensitivity of Mean-Variance-Efficient Portfolios to Changes in Asset Means - Some Analytical and Computational Results. *Review of Financial Studies, 4*(2), 315-342

Billio, M., Getmansky, M., Lo, A. W., & Pelizzon, L. (2012). Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of Financial Economics, 104*(3), 535-559

Black, F. (1976). The pricing of commodity contracts. *Journal of Financial Economics, 3*(1-2), 167-179

Black, F., & Litterman, R. (1992). Global portfolio optimization. *Financial Analysts Journal, 48*(5), 28-43

Black, F., & Scholes, M. (1973). The Pricing of Options and Corporate Liabilities. *Journal of Political Economy, 81*(3), 637-654

Bollerslev, T. (1986). Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics, 31*(3), 307-327

Bollerslev, T., & Mikkelsen, H. O. (1996). Modeling and pricing long memory in stock market volatility. *Journal of Econometrics, 73*(1), 151-184

Bonanno, G., Caldarelli, G., Lillo, F., & Mantegna, R. N. (2003). Topology of correlation-based minimal spanning trees in real and model markets. *Phys Rev E Stat Nonlin Soft Matter Phys, 68*(4 Pt 2), 046130

Bonanno, G., Caldarelli, G., Lillo, F., Micciche, S., Vandewalle, N., & Mantegna, R. N. (2004). Networks of equities in financial markets. *European Physical Journal B, 38*(2), 363-371

Bonanno, G., Lillo, F., & Mantegna, R. N. (2001). High-frequency cross-correlation in a set of stocks. *Quantitative Finance, 1*(1), 96-104

Brennan, M. J., Schwartz, E. S., & Lagnado, R. (1997). Strategic asset allocation. *Journal of Economic Dynamics & Control, 21*(8-9), 1377-1403

Campbell, J. Y., Lo, A. W., MacKinlay, A. C., & Whitelaw, R. F. (1998). The Econometrics of Financial Markets. *Macroeconomic Dynamics, 2*(04), 559-562

Caporale, G. M., Cipollini, A., & Spagnolo, N. (2005). Testing for contagion: a conditional correlation analysis. *Journal of Empirical Finance, 12*(3), 476-489

Carhart, M. M. (1997). On persistence in mutual fund performance. *Journal of Finance, 52*(1), 57-82

Carr, P., Geman, H., Madan, D. B., & Yor, M. (2002). The fine structure of asset returns: An empirical investigation. *Journal of Business, 75*(2), 305-332

Cheung, Y. W., & Lai, K. S. (1995). A Search for Long Memory in International Stock-Market Returns. *Journal of International Money and Finance, 14*(4), 597-615

Chiang, T. C., Jeon, B. N., & Li, H. M. (2007). Dynamic correlation analysis of financial contagion: Evidence from Asian markets. *Journal of International Money and Finance, 26*(7), 1206-1228

Chunhachinda, P., Dandapani, K., Hamid, S., & Prakash, A. J. (1997). Portfolio selection and skewness: Evidence from international stock markets. *Journal of Banking & Finance, 21*(2), 143-167

Coelho, R., Gilmore, C. G., Lucey, B., Richmond, P., & Hutzler, S. (2007). The evolution of interdependence in world equity markets—Evidence from minimum spanning trees. *Physica A: Statistical Mechanics and its Applications, 376*, 455-466

Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance, 1*(2), 223-236

De Luca, G., & Zuccolotto, P. (2011). A tail dependence-based dissimilarity measure for financial time series clustering. *Advances in Data Analysis and Classification, 5*(4), 323-340

De Luca, G., & Zuccolotto, P. (2014). Time Series Clustering on Lower Tail Dependence for Portfolio Selection. In M. Corazza & C. Pizzi (Eds.), *Mathematical and Statistical Methods for Actuarial Sciences and Finance* (pp. 131-140). Cham: Springer International Publishing.

Di Matteo, T. (2007). Multi-scaling in finance. *Quantitative Finance, 7*(1), 21-36

Di Matteo, T., Pozzi, F., & Aste, T. (2009). The use of dynamical networks to detect the hierarchical organization of financial market sectors. *The European Physical Journal B, 73*(1), 3-11

Ding, Z., Granger, C. W. J., & Engle, R. F. (1993). A long memory property of stock market returns and a new model. *Journal of Empirical Finance, 1*(1), 83-106

Durland, J. M., & McCurdy, T. H. (1994). Duration-Dependent Transitions in a Markov Model of U.S. GNP Growth. *Journal of Business & Economic Statistics, 12*(3), 279

Engle, R. F. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United-Kingdom Inflation. *Econometrica, 50*(4), 987-1007

Engle, R. F., & Ng, V. K. (1993). Measuring and Testing the Impact of News on Volatility. *Journal of Finance, 48*(5), 1749-1778

Eraker, B., Johannes, M., & Polson, N. (2003). The impact of jumps in volatility and returns. *Journal of Finance, 58*(3), 1269-1300

Eun, C. S., & Shim, S. (1989). International Transmission of Stock-Market Movements. *Journal of Financial and Quantitative Analysis, 24*(2), 241-256

Fama, E. F., & French, K. R. (1993). Common Risk-Factors in the Returns on Stocks and Bonds. *Journal of Financial Economics, 33*(1), 3-56

Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics, 116*(1), 1-22

Fama, E. F., & Macbeth, J. D. (1973). Risk, Return, and Equilibrium - Empirical Tests. *Journal of Political Economy, 81*(3), 607-636

Fernholz, R., & Shay, B. (1982). Stochastic Portfolio Theory and Stock-Market Equilibrium. *Journal of Finance, 37*(2), 615-624

Gabaix, X., Gopikrishnan, P., Plerou, V., & Stanley, H. E. (2003). A theory of power-law distributions in financial market fluctuations. *Nature, 423*(6937), 267-270

Glosten, L. R., Jagannathan, R., & Runkle, D. E. (1993). On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks. *Journal of Finance, 48*(5), 1779-1801

Gopikrishnan, P., Plerou, V., Nunes Amaral, L. A., Meyer, M., & Stanley, H. E. (1999). Scaling of the distribution of fluctuations of financial market indices. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics, 60*(5 Pt A), 5305-5316

Gopikrishnan, P., Rosenow, B., Plerou, V., & Stanley, H. E. (2001). Quantifying and interpreting collective behavior in financial markets. *Phys Rev E Stat Nonlin Soft Matter Phys, 64*(3 Pt 2), 035106

Granger, C. W. J., & Hyung, N. (2004). Occasional structural breaks and long memory with an application to the S&P

500 absolute stock returns. *Journal of Empirical Finance, 11*(3), 399-421
Hamilton, J. D. (1989). A New Approach to the Economic-Analysis of Nonstationary Time-Series and the Business-Cycle. *Econometrica, 57*(2), 357-384
Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (T. Edition Ed.): Elsevier.
He-Shan, G., & Qing-Shan, J. (2007, 2-4 Nov. 2007). *Cluster financial time series for portfolio.* Paper presented at the Wavelet Analysis and Pattern Recognition, 2007. ICWAPR '07. International Conference on.
Henry, Ó. T. (2002). Long memory in stock returns: some international evidence. *Applied Financial Economics, 12*(10), 725-729
Jacobsen, B., & Dannenburg, D. (2003). Volatility clustering in monthly stock returns. *Journal of Empirical Finance, 10*(4), 479-503
Jiang, Z. Q., & Zhou, W. X. (2008). Multifractality in stock indexes: Fact or fiction? *Physica a-Statistical Mechanics and Its Applications, 387*(14), 3605-3614
Jizba, P., Kleinert, H., & Shefaat, M. (2012). Renyi's information transfer between financial time series. *Physica a-Statistical Mechanics and Its Applications, 391*(10), 2971-2989
Jorion, P. (1992). Portfolio Optimization in Practice. *Financial Analysts Journal, 48*(1), 68-74
Kalay, A. (1982). The Ex-Dividend Day Behavior of Stock Prices: A Re-Examination of the Clientele Effect. *The Journal of Finance, 37*(4), 1059-1070
Kantelhardt, J. W., Zschiegner, S. A., Koscielny-Bunde, E., Havlin, S., Bunde, A., & Stanley, H. E. (2002). Multifractal detrended fluctuation analysis of nonstationary time series. *Physica a-Statistical Mechanics and Its Applications, 316*(1-4), 87-114
Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika, 30*(1/2), 81-93
Kenett, D. Y., Huang, X. Q., Vodenska, I., Havlin, S., & Stanley, H. E. (2015). Partial correlation analysis: applications for financial markets. *Quantitative Finance, 15*(4), 569-578
Kenett, D. Y., Preis, T., Gur-Gershgoren, G., & Ben-Jacob, E. (2012). Dependency Network and Node Influence: Application to the Study of Financial Markets. *International Journal of bifurcation and chaos, 22*(7), 14

Kenett, D. Y., Shapira, Y., & Ben-Jacob, E. (2009). RMT Assessments of the Market Latent Information Embedded in the Stocks' Raw, Normalized, and Partial Correlations. *Journal of Probability and Statistics, 2009*, 1-13

Kenett, D. Y., Shapira, Y., Madi, A., Bransburg-Zabary, S., Gur-Gershgoren, G., & Ben-Jacob, E. (2011). Index cohesive force analysis reveals that the US market became prone to systemic collapses since 2002. *PLoS One, 6*(4), e19378

Kenett, D. Y., Tumminello, M., Madi, A., Gur-Gershgoren, G., Mantegna, R. N., & Ben-Jacob, E. (2010). Dominating clasp of the financial sector revealed by partial correlation analysis of the stock market. *PLoS One, 5*(12), e15032

Kim, D. H., & Jeong, H. (2005). Systematic analysis of group identification in stock markets. *Phys Rev E Stat Nonlin Soft Matter Phys, 72*(4 Pt 2), 046133

Konno, H., Shirakawa, H., & Yamazaki, H. (1993). A mean-absolute deviation-skewness portfolio optimization model. *Annals of Operations Research, 45*(1), 205-220

Konno, H., & Yamazaki, H. (1991). Mean-Absolute Deviation Portfolio Optimization Model and Its Applications to Tokyo Stock-Market. *Management Science, 37*(5), 519-531

Laloux, L., Cizeau, P., Bouchaud, J. P., & Potters, M. (1999). Noise dressing of financial correlation matrices. *Physical Review Letters, 83*(7), 1467-1470

Lintner, J. (1965a). Security Prices, Risk, and Maximal Gains from Diversification. *Journal of Finance, 20*(4), 587-616

Lintner, J. (1965b). The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets. *Review of Economics and Statistics, 47*(1), 13-37

Lobato, I. N., & Savin, N. E. (1998). Real and Spurious Long-Memory Properties of Stock-Market Data. *Journal of Business & Economic Statistics, 16*(3), 261-268

Longin, F., & Solnik, B. (1995). Is the correlation in international equity returns constant: 1960-1990? *Journal of International Money and Finance, 14*(1), 3-26

Maasoumi, E., & Racine, J. (2002). Entropy and predictability of stock market returns. *Journal of Econometrics, 107*(1-2), 291-312

Maheu, J. M., & McCurdy, T. H. (2000). Identifying bull and bear markets in stock returns. *Journal of Business & Economic Statistics, 18*(1), 100-112

Mandelbrot, B. B. (1997). *The variation of certain speculative prices*: Springer.
Mandelbrot, B. B. (2013). *Fractals and Scaling in Finance: Discontinuity, Concentration, Risk. Selecta Volume E*: Springer Science & Business Media.
Mantegna, R. N. (1999). Hierarchical structure in financial markets. *European Physical Journal B, 11*(1), 193-197
Mantegna, R. N., & Stanley, H. E. (1995). Scaling Behavior in the Dynamics of an Economic Index. *Nature, 376*(6535), 46-49
Markowitz, H. (1952). Portfolio Selection. *Journal of Finance, 7*(1), 77-91
Marschinski, R., & Kantz, H. (2002). Analysing the information flow between financial time series - An improved estimator for transfer entropy. *European Physical Journal B, 30*(2), 275-281
Mel. (2015). Random Matrix Theory (RMT) Filtering of Financial Time Series for Community Detection. MathWorks File Exchange.
Merton, R. C. (1973a). An Intertemporal Capital Asset Pricing Model. *Econometrica, 41*(5), 867
Merton, R. C. (1973b). Theory of Rational Option Pricing. *Bell Journal of Economics, 4*(1), 141-183
Micciche, S., Bonanno, G., Lillo, F., & Mantegna, R. N. (2003). Degree stability of a minimum spanning tree of price return and volatility. *Physica a-Statistical Mechanics and Its Applications, 324*(1-2), 66-73
Nanda, S. R., Mahanty, B., & Tiwari, M. K. (2010). Clustering Indian stock market data for portfolio management. *Expert Systems with Applications, 37*(12), 8793-8798
Nelson, D. B. (1991). Conditional Heteroskedasticity in Asset Returns - a New Approach. *Econometrica, 59*(2), 347-370
Nelson, P. I., & Yang, S. (1988). Some properties of Kendall's partial rank correlation coefficient. *Statistics & Probability Letters, 6*(3), 147-150
Onnela, J. P., Chakraborti, A., Kaski, K., & Kertesz, J. (2002). Dynamic asset trees and portfolio analysis. *European Physical Journal B, 30*(3), 285-288
Onnela, J. P., Chakraborti, A., Kaski, K., & Kertesz, J. (2003). Dynamic asset trees and Black Monday. *Physica a-Statistical Mechanics and Its Applications, 324*(1-2), 247-252

Onnela, J. P., Chakraborti, A., Kaski, K., Kertesz, J., & Kanto, A. (2003). Dynamics of market correlations: taxonomy and portfolio analysis. *Phys Rev E Stat Nonlin Soft Matter Phys, 68*(5 Pt 2), 056110

Onnela, J. P., Kaski, K., & Kertesz, J. (2004). Clustering and information in correlation based financial networks. *European Physical Journal B, 38*(2), 353-362

Pafka, S., & Kondor, I. (2004). Estimated correlation matrices and portfolio optimization. *Physica a-Statistical Mechanics and Its Applications, 343*, 623-634

Plerou, V., Gopikrishnan, P., Nunes Amaral, L. A., Meyer, M., & Stanley, H. E. (1999). Scaling of the distribution of price fluctuations of individual companies. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics, 60*(6 Pt A), 6519-6529

Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L. A. N., Guhr, T., & Stanley, H. E. (2002). Random matrix approach to cross correlations in financial data. *Physical Review E, 65*(6), 066126

Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L. A. N., & Stanley, H. E. (1999). Universal and nonuniversal properties of cross correlations in financial time series. *Physical Review Letters, 83*(7), 1471-1474

Pojarliev, M., & Polasek, W. (2003). Portfolio construction by volatility forecasts: Does the covariance structure matter? *Financial Markets and Portfolio Management, 17*(1), 103-116

Pozzi, F., Di Matteo, T., & Aste, T. (2013). Spread of risk across financial markets: better to invest in the peripheries. *Sci Rep, 3*, 1665

Ramchand, L., & Susmel, R. (1998). Volatility and cross correlation across major stock markets. *Journal of Empirical Finance, 5*(4), 397-416

Roll, R. (1977). A critique of the asset pricing theory's tests Part 1: On past and potential testability of the theory. *Journal of Financial Economics, 4*, 129-176

Sandoval, L., & Franca, I. (2012). Correlation of financial markets in times of crisis. *Physica A: Statistical Mechanics and its Applications, 391*(1-2), 187-208

Shapira, Y., Kenett, D. Y., & Ben-Jacob, E. (2009). The Index cohesive effect on stock market correlations. *European Physical Journal B, 72*(4), 657-669

Sharpe, W. F. (1964). Capital-Asset Prices - a Theory of Market Equilibrium under Conditions of Risk. *Journal of Finance, 19*(3), 425-442

Skjeltorp, J. A. (2000). Scaling in the Norwegian stock market. *Physica a-Statistical Mechanics and Its Applications, 283*(3-4), 486-528

Stanley, H. E., Amaral, L. A. N., Gopikrishnan, P., & Plerou, V. (2000). Scale invariance and universality of economic fluctuations. *Physica a-Statistical Mechanics and Its Applications, 283*(1-2), 31-41

Statman, M. (1987). How Many Stocks Make a Diversified Portfolio? *The Journal of Financial and Quantitative Analysis, 22*(3), 353

Tola, V., Lillo, F., Gallegati, M., & Mantegna, R. N. (2008). Cluster analysis for portfolio optimization. *Journal of Economic Dynamics & Control, 32*(1), 235-258

Tseng, J. J., & Li, S. P. (2011). Asset returns and volatility clustering in financial time series. *Physica a-Statistical Mechanics and Its Applications, 390*(7), 1300-1314

Tumminello, M., Di Matteo, T., Aste, T., & Mantegna, R. N. (2007). Correlation based networks of equity returns sampled at different time horizons. *European Physical Journal B, 55*(2), 209-217

Tumminello, M., Lillo, F., & Mantegna, R. N. (2010). Correlation, hierarchies, and networks in financial markets. *Journal of Economic Behavior & Organization, 75*(1), 40-58

Uechi, L., Akutsu, T., Stanley, H. E., Marcus, A. J., & Kenett, D. Y. (2015). Sector dominance ratio analysis of financial markets. *Physica a-Statistical Mechanics and Its Applications, 421*, 488-509

Utsugi, A., Ino, K., & Oshikawa, M. (2004). Random matrix theory analysis of cross correlations in financial markets. *Phys Rev E Stat Nonlin Soft Matter Phys, 70*(2 Pt 2), 026110

Zhou, W. X. (2009). The components of empirical multifractality in financial returns. *EPL (Europhysics Letters), 88*(2), 28004

Zunino, L., Zanin, M., Tabak, B. M., Perez, D. G., & Rosso, O. A. (2009). Forbidden patterns, permutation entropy and stock market inefficiency. *Physica a-Statistical Mechanics and Its Applications, 388*(14), 2854-2864

Appendix

A. List of firms in each cluster

List of firms for Pearson clusters

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
SamsungC&T	KEPCO	LGChem	SamsungElec	SK	Naver	HyundaiMtr	SamsungLife	SamsungSecu	HanmiPharm
SamsungSDS	KT&G	SK Innovation	Posco	LGH&H	SK hynix	Mobis	ShinhanGroup	NHIS	Yuhan
Coway	Kangwonland	LotteChemical	LG	CJ	KoreaAerospace	KiaMtr	SKTelecom	MiraeAssetSec	GC Corp
NCsoft	LotteShopping	S-Oil	LGElectronics	ORION	Hyundai Dvp	HyundaiGlovis	SamsungF&MIns	DaewooSecu	SKChem
KCC	Emart	GS	KorZinc	CJ CheilJedang	KEPCO KPS	HankookTire	KBFinancialGr	KIH	GCH Corp
HtlShilla	LotteConf	Hyosung	LG Display	BGF Retail	CheilWorldwide	HanonSystems	KT	KAL	SamyangHoldings
GKL	S-1	HanwhaChem	HHI	CJ korea express	DWS	HyundaiWia	HanaFinancialGr	MeritzSecu	Donga ST
SKNetworks	HyundaiDepSt	Hanwha	HyundaySteel	Hanssem	Meritz Financial	HankookTireWor	Woori Bank	HyundaiSecu	JeilPharm
NexenTire	Nongshim	OCI	SamsungSDI	GS Retail	SsangyongCemen	Mando	IBK	Kiwoom	LGLS
KEPCO E&C	Shinsegae	KumhoPetrochem	HyundaiEng&Con	Ottogi	LGInt	HyundaiHomeSho	HanwhaLife	Hanjinkal	BukwangPharm
TaekwangInd	Youngone Corp	KolonIND	SamsungElecMec	LotteChilsung	HyundaiElev	KumhoTire	DongbuIns	Asiana Airlines	CHONGKUNDAN
SsangyongMtr	HiteJinro	KPIC	DaelimInd	CJ CGV	YungjinPharm		LG Uplus	AK Holdings	DaewoongPharm
Muhak	HyundaiGreenFo	SKC	DWEC	Hansae	LGHausys		SamsungCard	Yuanta Securities	IlyangPharm
KDHC	Korean Re	LotteFineChem	DHICO	Cuckoo	IS Dongseo		Kogas		Donga Socio
Huchems	LotteHimart		PoscoDaewoo	SamlipGenFood	NHN Ent		BNKFinancialGr		HanallBiopharma
KT Skylife	DongwonInd		SamsungHvyInd	KoreaKolmar	DaouTech		Hyundai M&F		JWPharma
SKGas	S&T MOTIV		SamsungEng	Cosmax	SYC		LIG Insurance		Kolon
Hanwha General	KorElecTerm		Doosan	FilaKorea	TCK		Meritz Insurance		IldongPharm
	YoungoneHolding		SamsungTechwin	DongwonF&B	Tongyang		DGBFinancialGr		
	Daekyo		GS E&C	LotteFood	HanilCement		TongyangLife		
	LF		Youngpoong	HanaTour	Dongbu HiTek		JB Financial		
			LG Innotek	Handsome	Hanchem				
			LS	Daesang	KISWire				
			DS Infra	HansaeYes24Hol	Nice				
			HyundaiRotem	KoreaKolmarHol	DuzonBizon				
			LSIIndustiralSyst	CrownConf	MiraeAssetLife				
			DSME						
			HyundaiMipoDoc						
			SBC						
			DongkukStlMill						
			POONGSAN						

L&L

List of firms for Partial clusters

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
SamsungLife	KEPCO	SamsungC&T	Posco	HanmiPharm	HyundaiMtr	KAL	SamsungSecu	LGChem	SamsungElec
ShinhanGroup	SKTelecom	Naver	LG	Yuhan	Mobis	Hanjinkal	NHIS	SK Innovation	
SamsungF&MIns	KT&G	SK hynix	LGElectronics	Doosan	KiaMtr	Asiana Airlines	MiraeAssetSec	LotteChemical	
KBFinancialGr	SK	SamsungSDS	KorZinc	GC Corp	LGH&H	AK Holdings	DaewooSecu	S-Oil	
HanaFinancialGr	Kangwonland	NCsoft	LG Display	SKChem	HyundaiGlovis		KIH	HHI	
Woori Bank	LotteShopping	KCC	HyundaySteel	SKNetworks	HanonSystems		MeritzSecu	SamsungSDI	
IBK	KT	KEPCO KPS	HankookTire	GCH Corp	Hanssem		HyundaiSecu	GS	
DongbuIns	Coway	CheilWorldwide	CJ CheilJedang	Korean Re	HtlShilla		Kiwoom	HyundaiEng&Con	
BNKFinancialGr	KoreaAerospace	Meritz Financial	Hyosung	Donga ST	HyundaiWia		Yuanta Securities	HanwhaChem	
Hyundai M&F	CJ	LGHausys	SamsungElecMec	JeilPharm	CJ CGV			DaelimInd	
LIG Insurance	HanwhaLife	NHN Ent	Hyundai Dvp	LGLS	Youngone Corp			Hanwha	
Meritz Insurance	ORION	DaouTech	PoscoDaewoo	BukwangPharm	Cuckoo			DWEC	
DGBFinancialGr	Emart	TCK	HankookTireWor	CHONGKUNDAN	KoreaKolmar			DHICO	
JB Financial	LG Uplus	HanilCement	Youngpoong	DaewoongPharm	GKL			OCI	
	SamsungCard	Nice	DWS	IlyangPharm	Mando			SamsungHvyInd	
	BGF Retail	Hanwha General	KolonIND	Donga Socio	Cosmax			SamsungEng	
	CJ korea express	DuzonBizon	LSIIndustiralSyst	HanallBiopharma	YoungoneHolding			SamsungTechwin	
	GS Retail	MiraeAssetLife	SsangyongCemen	KT Skylife	KoreaKolmarHol			KumhoPetrochem	
	Kogas		LGInt	JWPharma				GS E&C	
	LotteConf		HyundaiElev	IldongPharm				LG Innotek	
	S-1		YungjinPharm					LS	
	HyundaiDepSt		SamyangHoldings					DS Infra	
	Ottogi		NexenTire					HyundaiRotem	
	LotteChilsung		KumhoTire					KPIC	
	Nongshim		IS Dongseo					DSME	
	Hansae		KEPCO E&C					HyundaiMipoDoc	
	Shinsegae		TaekwangInd					SKC	
	HiteJinro		SsangyongMtr					LotteFineChemic	
	HyundaiGreenFo		SBC					DongkukStlMill	
	SamlipGenFood		SYC					L&L	
	HyundaiHomeSho		POONGSAN						
	LotteHimart		Tongyang						
	FilaKorea		Dongbu HiTek						
	TongyangLife		KDHC						
	DongwonF&B		Huchems						
	LotteFood		Hanchem						
	HanaTour		KISWire						
	Handsome		LF						
	Daesang		SKGas						
	DongwonInd		Kolon						
	S&T MOTIV								
	KorElecTerm								
	Muhak								
	HansaeYes24Hol								
	Daekyo								
	CrownConf								

List of firms for Kendall clusters

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
HanmiPharm	LGH&H	HHI	SamsungElec	Hyundai Dvp	KEPCO	LGChem	SamsungSecu	SK hynix	HyundaiMtr
Yuhan	KoreaAerospace	HyundaiEng&Con	SamsungC&T	Youngpoong	SKTelecom	SK Innovation	NHIS	NCsoft	Mobis
GC Corp	CJ	DaelimInd	Naver	GKL	KT&G	LotteChemical	MiraeAssetSec	KCC	SamsungLife
SKChem	ORION	DWEC	SamsungSDS	SKNetworks	SK	S-Oil	DaewooSecu	KAL	Posco
GCH Corp	CJ CheilJedang	DHICO	LG	NexenTire	Kangwonland	KorZinc	KIH	CheilWorldwide	ShinhanGroup
YungjinPharm	BGF Retail	PoscoDaewoo	LGElectronics	KEPCO E&C	KT	GS	MeritzSecu	Meritz Financial	KiaMtr
SamyangHoldings	Hanssem	SamsungHvyInd	LG Display	TaekwangInd	Coway	Hyosung	HyundaiSecu	SsangyongCemen	SamsungF&MIns
Donga ST	GS Retail	SamsungEng	SamsungSDI	SBC	HanwhaLife	HanwhaChem	Kiwoom	HyundaiElev	KBFinancialGr
JeilPharm	S-1	SamsungTechwin	SamsungElecMec	Muhak	Emart	Hanwha	Yuanta Securities	LGHausys	LotteShopping
LGLS	KEPCO KPS	GS E&C	Doosan	POONGSAN	LG Uplus	OCI		NHN Ent	HyundaySteel
BukwangPharm	HtlShilla	LS	LSIIndustiralSyst	KDHC	CJ korea express	KumhoPetrochem		Hanjinkal	HanaFinancialGr
CHONGKUNDAN	Ottogi	DS Infra		Huchems	LotteConf	KolonIND		DaouTech	Woori Bank
DaewoongPharm	CJ CGV	HyundaiRotem			HyundaiDepSt	LG Innotek		SYC	HyundaiGlovis
IlyangPharm	Hansae	DSME			LotteChilsung	KPIC		Asiana Airlines	IBK
Donga Socio	Cuckoo	HyundaiMipoDoc			Nongshim	LGInt		Tongyang	HankookTire
HanallBiopharma	HyundaiGreenFo	DongkukStlMill			Shinsegae	SKC		HanilCement	HanonSystems
JWPharma	DWS	L&L			Youngone Corp	LotteFineChemic		Dongbu HiTek	DongbuIns
Kolon	SamlipGenFood				HiteJinro	TCK		AK Holdings	SamsungCard
IldongPharm	KoreaKolmar				Korean Re			KISWire	Kogas
IS Dongseo					LotteHimart			Nice	BNKFinancialGr
Cosmax					Handsome			MiraeAssetLife	Hyundai M&F
FilaKorea					DongwonInd				HyundaiWia
DongwonF&B					S&T MOTIV				HankookTireWor
LotteFood					KorElecTerm				LIG Insurance
HanaTour					YoungoneHolding				Meritz Insurance
Daesang					Daekyo				Mando
HansaeYes24Hol					KT Skylife				HyundaiHomeSho
KoreaKolmarHol					LF				DGBFinancialGr
Hanchem					SKGas				KumhoTire
DuzonBizon									TongyangLife
CrownConf									SsangyongMtr
									JB Financial
									Hanwha General

List of firms for Partial Kendall clusters

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
HHI	LGChem	Hyundai Dvp	SamsungElec	HanmiPharm	LGH&H	Naver	KEPCO	HyundaiMtr	SamsungSecu
SamsungSDI	SK Innovation	Youngpoong	SamsungC&T	Yuhan	KoreaAerospace	SK hynix	SK	Mobis	NHIS
HyundaiEng&Con	LotteChemical	GKL	SamsungSDS	GC Corp	CJ	NCsoft	Kangwonland	SamsungLife	MiraeAssetSec
DaelimInd	S-Oil	SKNetworks	LGElectronics	SKChem	ORION	KCC	CJ korea express	Posco	DaewooSecu
DWEC	KorZinc	NexenTire	LG Display	GCH Corp	CJ CheilJedang	KAL	LotteConf	ShinhanGroup	KIH
DHICO	GS	KEPCO E&C	SamsungElecMec	YungjinPharm	BGF Retail	CheilWorldwide	HyundaiDepSt	KiaMtr	MeritzSecu
PoscoDaewoo	Hyosung	TaekwangInd	Doosan	SamyangHoldings	Hanssem	Meritz Financial	LotteChilsung	SKTelecom	HyundaiSecu
SamsungHvyInd	HanwhaChem	SBC	LG Innotek	Donga ST	GS Retail	SsangyongCemen	Nongshim	KT&G	Kiwoom
SamsungEng	Hanwha	POONGSAN	KISWire	LGLS	S-1	LGInt	Shinsegae	SamsungF&MIns	Yuanta Securities
SamsungTechwin	OCI	Huchems		BukwangPharm	KEPCO KPS	JeilPharm	Youngone Corp	KBFinancialGr	
GS E&C	KumhoPetrochem	L&L		CHONGKUNDAN	HtlShilla	NHN Ent	HiteJinro	LG	
LS	KolonIND			DaewoongPharm	Ottogi	Hanjinkal	HyundaiHomeSho	LotteShopping	
DS Infra	KPIC			IlyangPharm	CJ CGV	SKC	Korean Re	HyundaySteel	
HyundaiRotem	LotteFineChemic			Donga Socio	Hansae	SYC	LotteHimart	KT	
DSME				HanallBiopharma	Cuckoo	Asiana Airlines	Handsome	HanaFinancialGr	
HyundaiMipoDoc				JWPharma	HyundaiGreenFo	TCK	DongwonInd	Coway	
				Kolon	DWS	Tongyang	S&T MOTIV	Woori Bank	
				IldongPharm	SamlipGenFood	HanilCement	KorElecTerm	HyundaiGlovis	
					KoreaKolmar	AK Holdings	Muhak	IBK	
					HyundaiElev	MiraeAssetLife	YoungoneHolding	HankookTire	
					LGHausys		KDHC	HanwhaLife	
					IS Dongseo		KT Skylife	HanonSystems	
					Cosmax		LF	Emart	
					FilaKorea		SKGas	DongbuIns	
					DongwonF&B			LG Uplus	
					LotteFood			SamsungCard	
					DaouTech			Kogas	
					HanaTour			BNKFinancialGr	
					Daesang			Hyundai M&F	
					HansaeYes24Hol			HyundaiWia	
					KoreaKolmarHol			HankookTireWor	
					Dongbu HiTek			LIG Insurance	
					Hanchem			Meritz Insurance	
					Nice			Mando	
					DuzonBizon			DGBFinancialGr	
					CrownConf			LSIIndustiralSyst	
								KumhoTire	
								TongyangLife	
								SsangyongMtr	
								JB Financial	
								DongkukStlMill	
								Daekyo	
								Hanwha General	

B. Tables of Common Constituents of Clusters

Common constituents between Pearson and Partial

Pearson Clusters \ Cluster Number	*Partial Clusters* 1	2	3	4	5	6	7	8	9	10
1		33%		2%		19%				
2		1%	42%	22%						
3		2%		6%		32%				
4	67%	12%								
5		29%	10%	11%	5%	8%				
6					31%	69%				
7			19%	2%				65%	2%	
8			4%	73%						

Common constituents between Pearson and Kendall

Pearson Clusters \ Kendall Clusters	1	2	3	4	5	6
1			6%	41%		
2		47%	2%	10%		4%
3					33%	
4			7%		46%	
5		3%	57%	3%	4%	2%
6	69%	13%				
7			4%		3%	80%
8				36%		

Common constituents between Pearson and Partial Kendall

Pearson Clusters \ Partial Kendall Clusters	1	2	3	4	5	6	7
1			9%				38%
2	35%	5%					18%
3			3%	23%			
4				49%			
5	4%	16%	43%	9%			3%
6	14%					69%	
7	2%	18%		6%	65%		
8	3%						31%

Common constituents between Partial and Kendall

Partial Clusters \ Kendall Clusters	1	2	3	4	5	6
1					42%	
2			38%	22%	7%	
3		39%		3%	2%	7%
4		11%	16%	8%	9%	13%
5			5%	30%		2%
6			5%	13%	16%	
7		19%				
8	100%					
9						65%
10						2%

Common constituents between Partial and Partial Kendall

Partial Clusters \ Partial Kendall Clusters	1	2	3	4	5	6	7
1				33%			
2			32%	16%			20%
3	36%	6%		2%			7%
4	7%	25%	5%	11%	6%		11%
5	3%	5%	5%				25%
6		3%	5%	13%			13%
7	20%						
8						100%	
9	2%	4%		1%	76%		
10		5%					

Common constituents between Kendall and Partial Kendall

Kendall Clusters \ Partial Kendall Clusters	1	2	3	4	5	6	7
1						100%	
2	58%	3%					7%
3		20%	55%	11%			
4	1%						89%
5			2%	73%			
6	6%	16%		3%	65%		

C. Tables of Common Constituents of Clusters Estimated over Long Periods

Common constituents between Pearson and Partial estimated over two years

2014 – 2015		*Partial Clusters*									
		1	2	3	4	5	6	7	8	9	10
Pearson Clusters	1		50%	2%	9%			1%			
	2	13%	26%	2%	11%			5%			
	3				3%	92%					
	4				44%			5%			
	5		3%	33%	8%			4%	18%		5 %
	6									100%	
	7				4%			70%			
	8				1%		95%				

Common constituents between Pearson and Kendall estimated over two years

2014 – 2015		*Kendall Clusters*					
		1	2	3	4	5	6
Pearson Clusters	1	5%		4%	37%	2%	
	2	2%		22%	26%		
	3	14%	43%				
	4	46%	3%	7%	3%		
	5	2%	36%	5%			8%
	6				1%	80%	
	7	2%		4%			80%
	8				25%		

Common constituents between Pearson and Partial Kendall estimated over two years

2014 – 2015		*Partial Kendall Clusters*						
		1	2	3	4	5	6	7
Pearson Clusters	1	2%	58%		5%	3%		2%
	2	4%	19%		2%	44%		
	3			48%	10%			
	4		3%	5%	46%	4%	3%	
	5			31%	2%	5%	9%	
	6					3%		80%
	7				2%		79%	
	8	86%						

Common constituents between Partial and Kendall estimated over two years

2014 – 2015		*Kendall Clusters*					
		1	2	3	4	5	6
Partial Clusters	1				5%		
	2	4%	1%	13%	40%		
	3	3%	10%	8%	1%		5%
	4	36%	11%	3%	15%		3%
	5	13%	41%				
	6				24%		
	7			15%	1%	2%	68%
	8		11%				
	9				1%	80%	
	10		3%				

Common constituents between Partial and Partial Kendall estimated over two years

2014 – 2015	*Partial Kendall Clusters*						
	1	2	3	4	5	6	7
Partial Clusters 1		4%			8%		
2		52%	1%	4%	19%		
3			9%	3%	10%	4%	
4	4%	14%	13%	33%	6%	2%	
5			43%	11%			
6	82%						
7	2%				5%	76%	2%
8			11%				
9					3%		80%
10						3%	

Common constituents between Kendall and Partial Kendall estimated over two years

2014 – 2015	*Partial Kendall Clusters*						
	1	2	3	4	5	6	7
Kendall Clusters 1			5%	84%	2%		
2			89%			1%	
3					50%	8%	
4	29%	57%		1%	10%		
5							100%
6						86%	

Common constituents between Pearson and Partial estimated over three years

2013 – 2015	*Partial Clusters*									
	1	2	3	4	5	6	7	8	9	10
Pearson Clusters 1					79%	3%				
2					7%	44%				5 %
3	16%	32%				19%	2%			
4	9%	19%	57%		1%	3%		1%		
5				92%						
6		2%				2%	82%			
7				3%		2%			89%	
8								90%		

Common constituents between Pearson and Kendall estimated over three years

2013 – 2015		*Kendall Clusters*					
		1	2	3	4	5	6
Pearson Clusters	1					89%	2%
	2				36%	4%	15%
	3	2%	66%		9%		4%
	4	62%	3%	3%	8%		
	5						31%
	6		2%				38%
	7	26%					
	8			82%			

Common constituents between Pearson and Partial Kendall estimated over three years

2013 – 2015		*Partial Kendall Clusters*						
		1	2	3	4	5	6	7
Pearson Clusters	1	2%				89%		
	2	5%	18%		3%	4%	42%	
	3	2%		2%	60%		4%	
	4			60%	11%			1%
	5		67%					
	6	65%			5%			
	7			26%				
	8							90%

Common constituents between Partial and Kendall estimated over three years

2013 – 2015	*Kendall Clusters*					
	1	2	3	4	5	6
Partial Clusters 1	6%	19%				2%
2	18%	21%		6%		4%
3	38%	2%		9%		
4	1%					30%
5			3%	4%	83%	
6	3%	17%		28%	3%	13%
7		5%				32%
8			91%			
9	23%					
10						3%

Common constituents between Partial and Partial Kendall estimated over three years

2013 – 2015	*Partial Kendall Clusters*						
	1	2	3	4	5	6	7
Partial Clusters 1			6%	15%			
2	2%		17%	27%			
3			39%	8%			
4		63%	1%				
5					83%	5%	3%
6	6%	10%	3%	20%	3%	26%	
7	68%			3%			
8				2%			82%
9			23%				
10		6%					

Common constituents between Kendall and Partial Kendall estimated over three years

2013 – 2015	*Partial Kendall Clusters*						
	1	2	3	4	5	6	7
Kendall Clusters 1			99%	1%			
2				65%			
3				2%			91%
4		3%		16%		52%	
5					100%		
6	44%	43%		6%			

D. Return and standard deviation of clustering portfolios when a cluster of the worst stock performance was removed

Pearson	5 yr / 3 mo	5 yr / 6 mo	5 yr / 12 mo	3 yr / 3 mo	3 yr / 6 mo	3 yr / 12 mo	1 yr / 3 mo	1 yr / 6 mo	1 yr / 12 mo
Regular Return	0.086	0.085	0.079	0.079	0.080	0.080	0.079	0.074	0.080
Regular Std	0.189	0.186	0.186	0.187	0.184	0.180	0.178	0.178	0.174
Trimmed Return	0.084	0.085	0.082	0.070	0.086	0.080	0.078	0.074	0.078
Trimmed Std	0.193	0.187	0.191	0.185	0.182	0.183	0.183	0.178	0.179
Partial									
Regular Return	0.088	0.089	0.087	0.071	0.075	0.073	0.083	0.080	0.082
Regular Std	0.183	0.181	0.180	0.177	0.177	0.175	0.185	0.178	0.176
Trimmed Return	0.085	0.090	0.090	0.074	0.070	0.076	0.082	0.084	0.084
Trimmed Std	0.187	0.180	0.188	0.176	0.177	0.178	0.189	0.179	0.179
Kendall									
Regular Return	0.085	0.082	0.077	0.072	0.076	0.074	0.082	0.088	0.078
Regular Std	0.189	0.187	0.189	0.186	0.186	0.184	0.178	0.176	0.179
Trimmed Return	0.080	0.086	0.082	0.068	0.086	0.075	0.076	0.085	0.075
Trimmed Std	0.194	0.189	0.197	0.187	0.182	0.190	0.184	0.177	0.185
P. Kendall									
Regular Return	0.080	0.075	0.076	0.072	0.071	0.072	0.092	0.090	0.081
Regular Std	0.192	0.191	0.188	0.187	0.186	0.182	0.180	0.178	0.180
Trimmed Return	0.084	0.076	0.080	0.066	0.080	0.077	0.084	0.094	0.078
Trimmed Std	0.190	0.190	0.187	0.187	0.181	0.192	0.186	0.179	0.185

E. Return and standard deviation of clustering portfolios with clusters of a larger number of firms whose price dropped removed

Pearson	5 yr / 3 mo	5 yr / 6 mo	5 yr / 12 mo	3 yr / 3 mo	3 yr / 6 mo	3 yr / 12 mo	1 yr / 3 mo	1 yr / 6 mo	1 yr / 12 mo
Regular Return	0.086	0.085	0.079	0.079	0.080	0.080	0.079	0.074	0.080
Regular Std	0.189	0.186	0.186	0.187	0.184	0.180	0.178	0.178	0.174
Trimmed Return	0.092	0.073	0.080	0.049	0.078	0.077	0.075	0.087	0.076
Trimmed Std	0.190	0.191	0.183	0.197	0.188	0.182	0.177	0.179	0.173
Partial									
Regular Return	0.088	0.089	0.087	0.071	0.075	0.073	0.083	0.080	0.082
Regular Std	0.183	0.181	0.180	0.177	0.177	0.175	0.185	0.178	0.176
Trimmed Return	0.087	0.080	0.092	0.070	0.069	0.081	0.083	0.083	0.078
Trimmed Std	0.180	0.188	0.183	0.178	0.174	0.179	0.184	0.178	0.171
Kendall									
Regular Return	0.085	0.082	0.077	0.072	0.076	0.074	0.082	0.088	0.078
Regular Std	0.189	0.187	0.189	0.186	0.186	0.184	0.178	0.176	0.179
Trimmed Return	0.084	0.079	0.085	0.049	0.080	0.072	0.095	0.102	0.075
Trimmed Std	0.191	0.193	0.189	0.193	0.176	0.190	0.182	0.185	0.178
P. Kendall									
Regular Return	0.080	0.075	0.076	0.072	0.071	0.072	0.092	0.090	0.081
Regular Std	0.192	0.191	0.188	0.187	0.186	0.182	0.180	0.178	0.180
Trimmed Return	0.085	0.069	0.082	0.058	0.073	0.066	0.097	0.108	0.069
Trimmed Std	0.193	0.198	0.189	0.189	0.177	0.187	0.182	0.186	0.181

F. Return and standard deviation of clustering portfolios with a cluster of the worst EPS performance removed

Pearson	5 yr / 3 mo	5 yr / 6 mo	5 yr / 12 mo	3 yr / 3 mo	3 yr / 6 mo	3 yr / 12 mo	1 yr / 3 mo	1 yr / 6 mo	1 yr / 12 mo
Regular Return	0.086	0.085	0.079	0.079	0.080	0.080	0.079	0.074	0.080
Regular Std	0.189	0.186	0.186	0.187	0.184	0.180	0.178	0.178	0.174
Trimmed Return	0.093	0.088	0.081	0.080	0.090	0.082	0.083	0.083	0.082
Trimmed Std	0.184	0.181	0.185	0.183	0.180	0.172	0.176	0.175	0.172
Partial									
Regular Return	0.088	0.089	0.087	0.071	0.075	0.073	0.083	0.080	0.082
Regular Std	0.183	0.181	0.180	0.177	0.177	0.175	0.185	0.178	0.176
Trimmed Return	0.082	0.083	0.085	0.073	0.078	0.079	0.082	0.081	0.078
Trimmed Std	0.181	0.177	0.178	0.176	0.176	0.175	0.187	0.177	0.177
Kendall									
Regular Return	0.085	0.082	0.077	0.072	0.076	0.074	0.082	0.088	0.078
Regular Std	0.189	0.187	0.189	0.186	0.186	0.184	0.178	0.176	0.179
Trimmed Return	0.097	0.090	0.083	0.079	0.077	0.068	0.089	0.096	0.079
Trimmed Std	0.185	0.182	0.188	0.179	0.179	0.178	0.177	0.175	0.180
P. Kendall									
Regular Return	0.080	0.075	0.076	0.072	0.071	0.072	0.092	0.090	0.081
Regular Std	0.192	0.191	0.188	0.187	0.186	0.182	0.180	0.178	0.180
Trimmed Return	0.093	0.084	0.082	0.075	0.075	0.076	0.090	0.097	0.078
Trimmed Std	0.190	0.180	0.185	0.180	0.179	0.177	0.180	0.177	0.185

G. Return and standard deviation of clustering portfolios with clusters of a larger number of firms whose earnings shrank removed

Pearson	5 yr / 3 mo	5 yr / 6 mo	5 yr / 12 mo	3 yr / 3 mo	3 yr / 6 mo	3 yr / 12 mo	1 yr / 3 mo	1 yr / 6 mo	1 yr / 12 mo
Regular Return	0.086	0.085	0.079	0.079	0.080	0.080	0.079	0.074	0.080
Regular Std	0.189	0.186	0.186	0.187	0.184	0.180	0.178	0.178	0.174
Trimmed Return	0.093	0.087	0.080	0.086	0.086	0.084	0.089	0.083	0.091
Trimmed Std	0.191	0.187	0.186	0.188	0.185	0.181	0.177	0.176	0.173
Partial									
Regular Return	0.088	0.089	0.087	0.071	0.075	0.073	0.083	0.080	0.082
Regular Std	0.183	0.181	0.180	0.177	0.177	0.175	0.185	0.178	0.176
Trimmed Return	0.094	0.097	0.089	0.078	0.085	0.082	0.080	0.075	0.073
Trimmed Std	0.185	0.181	0.181	0.178	0.176	0.175	0.181	0.176	0.174
Kendall									
Regular Return	0.085	0.082	0.077	0.072	0.076	0.074	0.082	0.088	0.078
Regular Std	0.189	0.187	0.189	0.186	0.186	0.184	0.178	0.176	0.179
Trimmed Return	0.094	0.085	0.079	0.084	0.085	0.084	0.084	0.088	0.082
Trimmed Std	0.192	0.188	0.190	0.187	0.187	0.187	0.178	0.177	0.177
P. Kendall									
Regular Return	0.080	0.075	0.076	0.072	0.071	0.072	0.092	0.090	0.081
Regular Std	0.192	0.191	0.188	0.187	0.186	0.182	0.180	0.178	0.180
Trimmed Return	0.089	0.078	0.078	0.083	0.083	0.081	0.088	0.091	0.086
Trimmed Std	0.194	0.191	0.188	0.187	0.186	0.183	0.179	0.177	0.176

H. Return and standard deviation of Pearson and Partial clustering portfolios with outlier clusters removed

Pearson	5 yr / 3 mo	5 yr / 6 mo	5 yr / 12 mo	3 yr / 3 mo	3 yr / 6 mo	3 yr / 12 mo	1 yr / 3 mo	1 yr / 6 mo	1 yr / 12 mo
Regular Return	0.086	0.085	0.079	0.079	0.080	0.080	0.079	0.074	0.080
Regular Std	0.189	0.186	0.186	0.187	0.184	0.180	0.178	0.178	0.174
Trimmed Return	0.093	0.087	0.080	0.086	0.086	0.084	0.089	0.083	0.091
Trimmed Std	0.191	0.187	0.186	0.188	0.185	0.181	0.177	0.176	0.173
Partial									
Regular Return	0.088	0.089	0.087	0.071	0.075	0.073	0.083	0.080	0.082
Regular Std	0.183	0.181	0.180	0.177	0.177	0.175	0.185	0.178	0.176
Trimmed Return	0.094	0.097	0.089	0.078	0.085	0.082	0.080	0.075	0.073
Trimmed Std	0.185	0.181	0.181	0.178	0.176	0.175	0.181	0.176	0.174

Return and standard deviation of Kendall and Partial Kendall clustering portfolios with outlier clusters removed

Kendall	5 yr / 3 mo	5 yr / 6 mo	5 yr / 12 mo	3 yr / 3 mo	3 yr / 6 mo	3 yr / 12 mo	1 yr / 3 mo	1 yr / 6 mo	1 yr / 12 mo
Regular Return	0.085	0.082	0.077	0.072	0.076	0.074	0.082	0.088	0.078
Regular Std	0.189	0.187	0.189	0.186	0.186	0.184	0.178	0.176	0.179
Trimmed Return	0.094	0.085	0.079	0.084	0.085	0.084	0.084	0.088	0.082
Trimmed Std	0.192	0.188	0.190	0.187	0.187	0.187	0.178	0.177	0.177
P. Kendall									
Regular Return	0.080	0.075	0.076	0.072	0.071	0.072	0.092	0.090	0.081
Regular Std	0.192	0.191	0.188	0.187	0.186	0.182	0.180	0.178	0.180
Trimmed Return	0.089	0.078	0.078	0.083	0.083	0.081	0.088	0.091	0.086
Trimmed Std	0.194	0.191	0.188	0.187	0.186	0.183	0.179	0.177	0.176

I. Return and standard deviation of clustering portfolios with clusters of the worst EPS and weak EPS growth removed

Pearson	5 yr / 3 mo	5 yr / 6 mo	5 yr / 12 mo	3 yr / 3 mo	3 yr / 6 mo	3 yr / 12 mo	1 yr / 3 mo	1 yr / 6 mo	1 yr / 12 mo
Regular Return	0.086	0.085	0.079	0.079	0.080	0.080	0.079	0.074	0.080
Regular Std	0.189	0.186	0.186	0.187	0.184	0.180	0.178	0.178	0.174
Trimmed Return	0.104	0.093	0.082	0.095	0.103	0.089	0.094	0.094	0.095
Trimmed Std	0.184	0.180	0.186	0.183	0.180	0.173	0.176	0.173	0.171
Partial									
Regular Return	0.088	0.089	0.087	0.071	0.075	0.073	0.083	0.080	0.082
Regular Std	0.183	0.181	0.180	0.177	0.177	0.175	0.185	0.178	0.176
Trimmed Return	0.090	0.091	0.087	0.083	0.090	0.086	0.087	0.078	0.074
Trimmed Std	0.183	0.177	0.179	0.178	0.176	0.177	0.182	0.173	0.173
Kendall									
Regular Return	0.085	0.082	0.077	0.072	0.076	0.074	0.082	0.088	0.078
Regular Std	0.189	0.187	0.189	0.186	0.186	0.184	0.178	0.176	0.179
Trimmed Return	0.109	0.095	0.087	0.094	0.088	0.080	0.092	0.098	0.087
Trimmed Std	0.186	0.182	0.189	0.179	0.179	0.178	0.178	0.176	0.177
P. Kendall									
Regular Return	0.080	0.075	0.076	0.072	0.071	0.072	0.092	0.090	0.081
Regular Std	0.192	0.191	0.188	0.187	0.186	0.182	0.180	0.178	0.180
Trimmed Return	0.107	0.094	0.086	0.087	0.088	0.086	0.093	0.097	0.080
Trimmed Std	0.189	0.179	0.185	0.180	0.179	0.177	0.180	0.178	0.184

국문 초록

금융 시장은 다양한 분야에서 연구 주제로써 다뤄지고 있다. 예컨대 금융경제학은 경제학의 주요 분야로 자리 잡았으며, 계량경제학은 금융 시장을 비롯한 여러경제 지표와 자료를 분석하기 위한 방법론들을 제시하며 발전되어왔다. 이와 더불어 수리통계적 모형을 활용한 금융 자산의 가치 평가 또는 리스크 측정이 활발하게 이루어지고 있으며, 이를 위한 방법론도 끊임없이 개발되고 있다. 경제학에 물리학 기법을 도입하여 연구하는 경제물리학 (Econophysics) 분야에서는 여러 금융 자산에서 공통적으로 발견되는 몇가지 현상들을 발견하였고, 이로부터 정형화된 사실 (Stylized fact) 을 도출하는 데 기여하였다. 이러한 연구들은 금융 시장에 대한 정량적인 이해를 높여왔기도 했지만, 지금까지의 연구로는 금융 위기와 같이 금융시장에서 발생하는 대규모의 변동을 예측하거나 설명하는 데에 한계를 지니고 있기도 하다. 이는 금융 시장을 체계적으로 이해하기 위해 끊임없는 연구가 필요하다는 방증이다.

한편 금융시장을 설명하는 좋은 방법 중 하나로 네트워크 이론 (Network theory) 을 들 수 있다. 네트워크 이론에서 자주 사용되는 계층 트리 (Hierarchical tree) 와 최소 신장 트리 (Minimum Spanning Tree) 와 같은 방법론은 자산들간의 관계를 2 차원 평면 상의 거리 척도로 나타낼 수 있는 장점을 지니고 있다. 이러한 방법론을 적용하기 위해 먼저 비유사도 측도 (Dissimilarity measure) 를 구하게 되는데, 금융 시장에서는 이 측도로 상관계수 (Correlation coefficient) 를 주로 사용한다. 상관계수는 자산의 수익률을 통해 산출되는데, 이 측도는 자산들간의 동조화 (Co-movement) 경향을 효과적으로 나타낼 수 있어서 금융 자산의 상관 관계 실증 분석에서 빈번하게 사용된다. 그런데 상관계수 자체로는 거리측도로써 사용할 수 없다. 따라서 네트워크 이론에서는 이를 거리 함수 (Distance function) 에 적용하여 거리를 측정할 수 있는 값으로 바꾸어 사용하게 된다. 이러한 거리 함수는 본래 상관계수 간의 토폴로지 (Topology) 를 보존하는 성질을 가지며, 이를 통해 자산 간의 관계를 있는

그대로 볼 수 있다. 이러한 방법론은 여러 자산을 이용하여 포트폴리오를 구성하는 현대 포트폴리오 이론 (Modern portfolio theory) 과 같은 방향에서 접근하는 것이라 할 수 있다.

본 논문에서는 대한민국 코스피 (Korea Composite Stock Price Index) 지수에 포함되어 있는 기업 중 시가 총액이 가장 큰 30 개의 기업을 이용하여 이들의 비유사도를 상관계수에 기반하여 측정하였다. 측정된 비유사도를 기반으로 계통수 (Dendrogram) 분석을 수행하였고, 이를 통해 기업들간의 상대적인 거리와 위치를 확인해 보았다. 그 결과 은행업 또는 자동차 업계 (Industrial sector) 의 기업들은 상대적으로 가까운 것으로 나타났으나, 그 외에 중공업, 정보 통신, 소비재 관련 기업 등은 업계와 관련없이 기업들이 모이는 것으로 나타났다. 이는 업계의 다양화 (Diversification)와 포트폴리오의 다양화 경향이 다른 방향으로 나타날 수 있음을 의미한다.

다음 단계로 코스피에 상장되어 있는 시가 총액 기준 상위 200 개의 주식들의 비유사도 행렬 (Dissimilarity matrix) 을 만들어 클러스터 분석 (Clustering analysis) 을 수행하였다. 현재 시장에서는 업계를 기준 삼아 개별 종목들의 주식을 클러스터링하는 방법을 널리 취하고 있는데, 이는 여러 업계로부터 주식을 선택하여 포트폴리오를 구성함으로서 포트폴리오의 다양화를 꾀하기 위함이다. 본 연구에서는 주식의 수익률에 기반한 클러스터 분석을 통해 업계의 다양화를 통한 포트폴리오 다양화가 절대적인 수익률 및 위험 대비 수익률의 극대화에 적합한지 알아보고자 한다. 분석 결과 같은 업계에 있는 주식들이 다른 클러스터에 할당되는 경우나 다른 업계에 있는 주식들이 같은 클러스터에 소속되는 경우가 빈번하게 발견되었다. 이는 포트폴리오의 다양화를 얻기 위한 수단으로써 업계에 의한 다양화를 꾀하기엔 다소 그 효과가 미흡할 수 있다는 것을 함의한다.

이와 더불어 클러스터 분석을 바탕으로 주식 포트폴리오 구성을 시도해 보았다. 클러스터 분석 결과 서로 다른 클러스터에 속하게 된 주식으로 구성한 포트폴리오는 다양화된 포트폴리오 (Diversified portfolio) 로, 위험 대비 수익률 측면에서 기준 지수 (Benchmark index) 에 비하여 더 나은 결과를 낼 수 있다라는 가설을 세울 수 있다. 이를 실제로 검증해 보기 위해 위에서 사용한 것과 동일한 200 개의 주식으로

포트폴리오를 구성해 보았다. 각각 주식 10 개, 20 개, 30 개로 이루어진 세 종류의 포트폴리오를 구성하였고, 이들의 수익률 및 변동성의 척도인 표준편차를 도출하였다. 이를 위해 주식을 클러스터로 나누어 각각의 클러스터에서 최대한 동수의 주식을 선택하였고, 나머지는 크기가 큰 클러스터로부터 추가적으로 선택하는 방법으로 구성하였다. 각각의 클러스터로부터 임의의 주식을 선택하여 포트폴리오를 구성하였고, 이러한 포트폴리오를 1000 번 생성하여 이들의 평균 수익률과 표준편차를 기준으로 performance 를 비교하였다. 그 결과 기준 지수인 코스피 지수 및 상위 10 개 기업을 지수화한 대기업 지수 (Large cap index)보다 클러스터에 기반한 포트폴리오가 절대적인 수익률과 위험 대비 수익를 모든 측면에서 더 나은 결과를 보여주었다. 그러나 이는 자료 기간인 2015 년 중 상위 기업들의 실적 및 코스피 지수가 좋지 못하였던 점을 고려하였을 때 적절한 기준이되기 어렵다라고 판단될 여지가 있다. 따라서 랜덤 포트폴리오 (Random Portfolio) 를 구성해 또 하나의 기준 지수로 활용하였다. 랜덤 포트폴리오는 클러스터 분석을 통해 도출된 클러스터가 아니라 200 개의 전체 주식 중에서 각각 10 개, 20 개, 30 개를 임의로 선택하여 구성한 포트폴리오이다. 이를 통해 클러스터 포트폴리오가 랜덤 포트폴리오와 통계적으로 유의하게 다른지 검정할 수 있다. 실제로 대부분의 클러스터 포트폴리오가 랜덤 포트폴리오에 비해 더 좋은 성적을 내는데 실패하였고, t 검정 (t test) 을 통해 두 포트폴리오의 수익률이 통계적으로 유의하게 같지 않음을 검정 결과 확인할 수 있었다. 클러스터 포트폴리오의 수익률이 낮은 이유를 확인하기 위해 각각의 포트폴리오와 클러스터들을 분석해본 결과 몇몇 클러스터가 실적이 저조한 주식들로 이루어진 클러스터임을 확인할 수 있었다. 클러스터 포트폴리오가 모든 클러스터에서 동일한 수의 주식을 선택한 점을 고려할때, 클러스터에 포트폴리오는 필연적으로 실적이 좋지 못한 주식을 특정 빈도 이상 포함하게 됨을 알 수 있다. 이러한 클러스터를 포트폴리오 구성 단계에서 제외하면 클러스터 포트폴리오의 기대 수익률이 올라갈 수 있다.

본 연구를 통해 클러스터 포트폴리오가 시장의 구조를 사후적으로 분석하여 특성이 서로 다른 주식을 찾아내는데 이용될 수 있다는 점을 확인하였다. 그러나 현실에서는 미래시점에 관한 정보는 존재하지 않으며, 어떤 주식으로 포트폴리오를 구성할지에 대한 의사 결정은 사전적으로 내려져야 한다. 이를 위해 포트폴리오 구성

시점을 기준으로 과거의 자료만을 이용하여 포트폴리오 구성을 시도하였다. 2012 년부터 2014 년까지 코스피 지수에 편입되어 있던 종목들 중 시가 총액이 큰 200 개의 기업을 이용하여 이전과 동일한 방법으로 포트폴리오를 구성하였다. 그리고 2015 년 1 월 1 일부터 2015 년 12 월 31 일까지의 포트폴리오 수익률을 기록하였다. 그 결과 클러스터 포트폴리오들이 기준 지수들 보다 좋은 실적을 내는 데에는 성공하였으나, 랜덤 포트폴리오 보다 더 좋은 성적을 내는데에는 실패하였다. 앞선 연구 결과로부터 성적이 좋지 않은 주식들로 이루어진 클러스터가 존재할 수 있는 것을 확인하였는데, 이를 재차 확인해 보기 위해 클러스터를 하나씩 제거하면서 다시 포트폴리오를 구성해본 결과 클러스터의 실적이 대부분 향상됨을 발견하였다. 특히 2012 년부터 2014 년까지의 주식의 수익률 및 기업의 영업이익률이 낮은 주식들로 이루어진 클러스터의 경우 2015 년 까지 좋은 성적을 내지 못한 경우가 많음을 관찰하였고, 이러한 클러스터를 제거하면 상대적으로 실적이 향상됨을 확인하였다.

앞선 연구에서는 하나의 시점에 대해 클러스터 포트폴리오를 구성한 후 각각의 클러스터의 특성을 분석하였는데, 이를 일반화하여 보다 장기간에 걸쳐 포트폴리오를 구성 및 운용하고자 한다. 이 경우 보다 긴 기간의 자료가 필요하므로 뉴욕 증권 거래소 (New York Stock Exchange) 와 나스닥 (NASDAQ) 에 상장되어 있는 각 시점별 시가 총액 기준 상위 200 개 기업으로 포트폴리오를 구성하였다. 따라서 이 때의 기준 지수 또한 자연스럽게 S&P500 지수와 대기업 지수로 변경되었다. 1990 년부터 일정 기간동안 상관계수를 측정하였고, 측정이 완료된 시점부터 포트폴리오를 구성하여 일정 기간 간격으로 리밸런싱을 하며 포트폴리오를 운용하였다. 그 결과 클러스터 포트폴리오가 S&P500 지수에 비해 좋은 실적을 내는데 실패하였다. 또한 성적이 좋지 않은 주식들로 이루어진 클러스터를 제거하기 위해 앞선 연구 결과를 기반으로 클러스터 제거를 위한 규칙을 생성하였고, 이를 통해 구성된 포트폴리오들의 성적을 확인해 보았다. 그 결과 영업 손실이 자주 발생한 주식이 가장 많이 속한 클러스터 또는 영업 이익의 성장률이 음의 값을 가지는 주식이 주로 속한 클러스터를 제거하였을때 포트폴리오의 수익률이 상승하였고, 이 두 가지 규칙을 동시에 적용한 포트폴리오가 다른 클러스터 포트폴리오 또는 기준 지수보다 수익률이 높음을 확인하였다.

본 연구에서는 주식의 수익률에 기반한 상관관계 분석과 클러스터 분석을 통하여 시장 구조를 분석하였고, 이를 통해 시장 구조가 업계 기준으로 파악한 구조와 일치하지 않음을 확인하였다. 아울러 클러스터 분석을 통한 클러스터 포트폴리오 구성 방안을 제시하였고, 클러스터 분석의 강점을 이용하여 저조한 실적의 주식으로 구성된 클러스터를 제거함으로서 클러스터 포트폴리오의 실적을 개선시킬 수 있음을 확인하였다.

Keywords : 상관 관계 분석, 클러스터 분석, 포트폴리오 관리, 운용 시뮬레이션
Student ID : 2012-21072