Physica A 462 (2016) 410–420

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Clustering stocks using partial correlation coefficients

Sean S. Jung, Woojin Chang [*]

Department of Industrial Engineering, Seoul National University, Republic of Korea

HIGHLIGHTS

- Correlation analyses are conducted on Korean stock market.
- Agglomerative hierarchical clustering is performed based on correlation matrices.
- Each cluster consists of firms from multiple business sectors.

ARTICLE INFO

Article history:
Received 15 October 2015
Received in revised form 14 May 2016
Available online 23 June 2016

Keywords:
Financial market
Stylized facts
Pearson correlation
Partial correlation
Agglomerative clustering

ABSTRACT

A partial correlation analysis is performed on the Korean stock market (KOSPI). The difference between Pearson correlation and the partial correlation is analyzed and it is found that when conditioned on the market return, Pearson correlation coefficients are generally greater than those of the partial correlation, which implies that the market return tends to drive up the correlation between stock returns. A clustering analysis is then performed to study the market structure given by the partial correlation analysis and the members of the clusters are compared with the Global Industry Classification Standard (GICS). The initial hypothesis is that the firms in the same GICS sector are clustered together since they are in a similar business and environment. However, the result is inconsistent with the hypothesis and most clusters are a mix of multiple sectors suggesting that the traditional approach of using sectors to determine the proximity between stocks may not be sufficient enough to diversify a portfolio.

1. Introduction

For decades, the financial market has received an enormous amount of attention from academia. Yet its complex nature still remains elusive and recent financial crises support the need to better understand it. Statistical Physics is one of the popular tools to analyze financial data and many important discoveries have been made by using it. For example, though a financial asset's returns show no serial correlation, the absolute values of returns have positive correlation over long lags, a phenomenon well known as the long memory property [1]. Also, a financial return series has heavier tails than Gaussian distribution and the absolute value of the return series tends to follow a power law [2,3]. It also shows nonlinearity in behavior as observed in phenomena such as volatility clustering and regime switching behavior [4–7]. Another interesting finding is that when multiple assets are considered together, the return series are often correlated and the correlation tends to become larger during a financial crisis [8,9]. Correlation plays a critical role in analysis of financial time series. For example, it has been known for a while that the correlation in equity returns is time varying [10]. Also, co-movement of international markets was widely studied and methods such as conditional correlation and dynamic correlation were some of the most

[*] Corresponding author.
E-mail address: changw@snu.ac.kr (W. Chang).

http://dx.doi.org/10.1016/j.physa.2016.06.094
0378-4371/© 2016 Elsevier B.V. All rights reserved.

popular tools to approach this subject [11–13]. Network theory can be applied to study a correlation matrix as well, and previously it was demonstrated that a correlation matrix can be associated with a hierarchical tree and when tested in a small set of stocks, the stocks in the same sectors were clustered together [14]. When it comes to the construction of portfolio of assets, correlation is one of the key variables an investor needs to consider [15,16].

This study employs the correlation to study how firms are related to others over different periods of time including the recent financial crises. Pearson correlation is widely accepted as a standard measure of co-movement between two financial return series but it has few weaknesses. For example, as it was mentioned before, the market dynamics are nonlinear in nature and so are the correlations between assets, but Pearson correlation measures a linear co-movement between two random variables. Also, when it comes to analysis of firms in a similar environment, one should consider the common factors affecting the firms such as the growth rate of market or foreign exchange rates. Those common factors may have an influence on the stocks, resulting in a bias in analysis. Such bias may skew the results to favor one side to the other, giving it a 'false correlation'. Though it is impossible to completely remove all external factors from a financial time series as some of the factors are unobservable or difficult to assign numerical values such as the corporate governance, it is likely that removing some of the obvious and prevalent factors may provide different insights which have not discovered.

Recently, Kenett et al. applied the concept of partial (residual) correlation analysis to study the financial markets [17]. They successfully demonstrated the index cohesive effect and identified a dominating sector within a market using the partial correlation analysis [18,19]. The network approach was also used to construct node–node correlation matrices and it was shown that a node with insignificant influence does not disrupt the network even when the node is removed [20]. Another study proposed a measure called Sector Dominance Ratio using Pearson and partial correlations to study the market structure and it has empirically shown that the financial sector exhibits strong dominance for US and UK markets [21]. The studies mentioned above utilize the partial correlation analysis and the strength of this method comes from its ability to remove a common factor in correlation between two variables. A partial correlation measures how a random variable '*i*' correlates with another random variable '*j*' when a common factor '*k*' is removed from both of them. In this sense, it can be said that it is correlation between residuals of *i* and *j*.

One may notice that if a common factor is chosen to be the market return, the returns used to compute correlation resemble the return from Capital Asset Pricing Model [22–24]. CAPM was developed to describe how a risky asset should be valued and if the model was correct, the correlation between the residuals, or the partial correlation coefficient of two risky assets, should be statistically zero. Section 3 discusses this point and shows that the correlation coefficients are in fact non-zero. This should not come as a surprise as the CAPM was known for its relatively weak explanatory power [25].

This study is another application of the partial correlation analysis. A market index such as Korea Composite Stock Price Index (KOSPI) reflects the overall size of Korean economy. A growth of the index means that the size of economy is growing as well which will positively affect every firm in the market. By removing the effect of the market index, it may be possible to shed a new light regarding the individual performance of the stocks and their correlation with others. Specifically, it addresses following two questions:

i. How much difference exists between the partial correlation and Pearson correlation? Does the difference persist through different period of time?
ii. In terms of the partial correlation, how are firms related to each other?

Both the partial correlation and Pearson correlation are computed every 30 months to minimize the effect of change in correlation coefficients which occurs due to the nonlinearity of the market. The length of 30 months also ensures that the computed correlation coefficients are statistically significant. It is then visualized to address the first question and a simple agglomerative clustering approach is used to explore the second question.

The remaining sections of this paper are organized as follows: Section 2 describes the data set used for analysis and method of how to compute the correlations used in this paper. Section 3 reports the results of the partial correlation analysis and Pearson correlation analysis. In Section 4, clustering analyses are performed. Section 5 provides the conclusion.

2. Data and methods

The monthly adjusted closing price series of KOSPI and the stocks listed in the index from December of 2004 to December of 2014 are used for this study. The data are provided by DataGuide from FnGuide (http://www.fnguide.com/), a professional financial analytics service for the Korean market. All computation and analytics are done using MATLAB. Total of 732 firms existed throughout this period and trade volume data are used to filter out the ones that did not trade during a period of analysis. Note that not all the firms existed for the entire period of time resulting in different number of firms and different size of correlation matrices for each period.

For each firm, a log return $r_i(t)$ was computed by

$$r_i(t) = \ln(P_i(t)) - \ln(P_i(t-1)) \qquad (1)$$

where $P_i(t)$ denotes the adjusted closing price of firm i at time t. The log return series are used for analyses which covered ten years from January of 2005 to December of 2014.

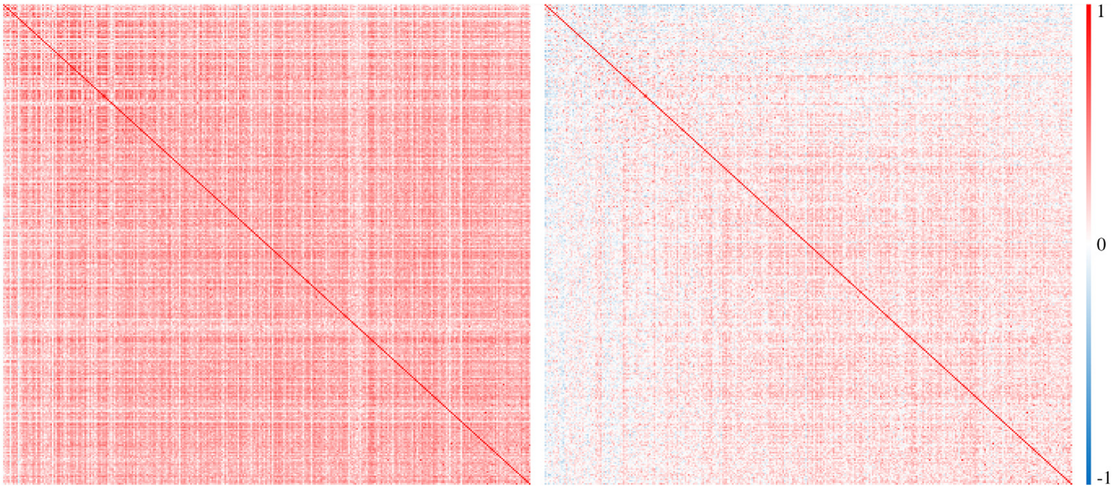

Fig. 1. Pearson correlation matrix (left) and partial correlation matrix (right). A matrix was made using 324 firms which were listed in the KOSPI index for the entire period from 2005 to 2014. Firms were sorted by market capitalization and are in descending order. The coefficients were color coded for visualization where negative values are in blue, zero is in white and positive values are in red. The main diagonal for the both matrices is autocorrelation which is always equal to one. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The correlation matrix is then computed using the return series. Pearson correlation $\rho(r_i, r_j)$ is defined by

$$\rho\left(r_i, r_j\right) = \frac{\langle (r_i - \langle r_i \rangle)(r_j - \langle r_j \rangle)\rangle}{\sigma_i \cdot \sigma_j} \tag{2}$$

where $\langle r_i \rangle$ represents the mean value of r_i over a given period, which is 30 months in this study, and σ_i denotes the standard deviation of r_i over the same period.

The partial correlation with a common factor is then computed by

$$\rho\left(r_i, r_j : r_m\right) = \frac{\rho\left(r_i, r_j\right) - \rho\left(r_i, r_m\right) \cdot \rho\left(r_j, r_m\right)}{\sqrt{\left(1 - \rho^2\left(r_i, r_m\right)\right) \cdot \left(1 - \rho^2\left(r_j, r_m\right)\right)}} \tag{3}$$

where r_m represents a common factor which in this case is the KOSPI index return series [17]. Pearson correlation and the partial correlation are computed for every 30 months; that is, the time series are split into four subperiods, each of which is 30 months long.

3. Pearson correlation vs. partial correlation

Fig. 1 shows color maps of Pearson correlation matrix and the partial correlation matrix for entire period from 2005 to 2014. The firms are sorted by market capitalization.

The values are color gradient coded so that blue to white represents negative value ($-1 \leq \rho < 0$), white for zero ($\rho = 0$), and white to red corresponds to positive value ($0 < \rho \leq 1$). It is visibly clear that the Pearson correlation matrix has much more red which implies it mostly consists of positive values. However, the partial correlation matrix shows some blue cells which correspond to negative coefficient value. One can easily deduce that the overall coefficient values of Pearson correlations are greater than that of the partial correlation. Given that the mean value of correlation between firms and the KOSPI index is 0.4606, it is highly likely that when this factor is removed, the overall value of correlation shrinks. If the market return is the only factor, then the correlation between two stocks becomes zero after the market factor is removed. However, the result shows that the correlation remains non-zero suggesting that the market is not the only factor affecting the correlation between stocks.

Table 1 summarizes the data of both correlation matrices. For each subperiod, the number of firms used to pair up and compute the correlation coefficients are reported. The mean values of the lower triangle of the matrix are also shown for each correlation coefficient matrix. 'Number of $\rho > 0$' and 'Number of $\rho \leq 0$' are the number of correlation coefficients which have positive and negative values, respectively. For each subperiod and the entire period of the sample, the mean value of the Pearson correlation was always greater than that of Partial correlation which is consistent with the hypothesis that the index, which serves as a proxy for the systematic risk of the market, drives up the correlations between firms. Note that 'Number of $\rho > 0$' and 'Number of $\rho \leq 0$' for each correlation coefficients provide insight regarding the first question from the introduction section. The larger number in 'Number of $\rho > 0$' for Pearson correlation implies that many firms have a positive correlation with each other, but when stripped of the common driving force of the index, the partial correlation suggests

Table 1
Summary of the Pearson correlation matrix and the partial correlation matrix. The number of firms used in the analysis changed over time as the index listed more firms and some firms were delisted. Mean Pearson and Mean Partial are the mean values of the lower triangle of Pearson and the partial correlation matrices, respectively. The first Number of $\rho > 0$ and Number of $\rho < 0$, respectively, count the number of correlation coefficients that are positive and negative in the lower triangle of Pearson correlation matrix. The second Number of $\rho > 0$ and Number of $\rho < 0$ count the same for the partial correlation matrix.

Year	Number of firms	Pearson correlation			Partial correlation		
		Mean	Number of $\rho > 0$	Number of $\rho \leq 0$	Mean	Number of $\rho > 0$	Number of $\rho \leq 0$
Period 1	475	0.28	103,564	9,011	0.08	72,052	40,523
Period 2	497	0.38	118,871	4,385	0.12	87,615	35,641
Period 3	551	0.23	131,717	19,808	0.07	96,733	54,792
Period 4	553	0.09	104,467	48,161	0.05	92,464	60,164
Entire	324	0.28	51,632	694	0.08	39,654	12,672

Table 2
Confidence intervals at 5% significance level for the coefficients of Pearson and the partial correlation matrices. Only the lower triangle part of the coefficients are used. P-values were 0 for all cases.

	Pearson					Partial				
	Period 1	Period 2	Period 3	Period 4	Entire	Period 1	Period 2	Period 3	Period 4	Entire
Lower	0.2801	0.3783	0.2267	0.0940	0.2758	0.0779	0.1163	0.0725	0.0536	0.0829
Upper	0.2824	0.3804	0.2287	0.0960	0.2780	0.0804	0.1187	0.0746	0.0556	0.0850

that the correlation without the influence of the market, which can be loosely interpreted as an individual performance, is more likely to have smaller or negative correlation.

One may notice that the partial correlation without the influence of the market conflicts with an assumption of the Capital Asset Pricing Model (CAPM) in which the model assumes that the residuals are independent and identically distributed [24]. If the assumption is true, then the residuals would effectively be white noise and they should not be correlated, having correlation coefficients of zero, and clustering analysis performed on a group of white noise would yield no meaningful results. Therefore, it may be necessary to take a precaution and ensure that the correlations between residuals are in fact non-zero. A simple t-test is performed to determine whether the coefficients of the lower triangle part of each of the partial and Pearson correlation matrices are zero. For all cases, the null hypothesis of mean zero is rejected at 5% significance level and P-values are zero. Table 2 provides the confidence interval at 5% significance level.

4. Clustering analysis

The partial correlation analysis from the previous section proposes that the interconnection between firms is different from what it normally seems. This section is dedicated to study the proximity between firms using partial correlation analysis to see which firms are close to each other. Agglomerative clustering analysis is performed with the Euclidean distance between correlation coefficients as a distance measure to determine the proximity between objects, which are firms in this case. To illustrate the agglomerative method, each firm with n features, which is equal to the number of firms in each year for this study and having the value of the partial correlation, is treated as an object. The distance, measured by Euclidean distance, is computed for every pair of objects. Then, two objects with the shortest distance are merged and become the first cluster. The distance between other objects and the cluster is then recalculated. Two objects, or clusters, with the shortest distance are merged again to form a cluster. The process is repeated until all objects are in one cluster.

The distance between two clusters is determined by the average distance between the members of two clusters. That is, the distance $dist_{i,j}$ between cluster C_i and C_j is

$$dist_{i,j} = \frac{1}{n_i n_j} \sum_{o_i \in C_i,\ o_j \in C_j} \delta(o_i - o_j) \tag{4}$$

where n_i is the number of objects in C_i, o_i is an object belongs to C_i and $\delta(o_i - o_j)$ is the distance between o_i and o_j.

Once the pairwise distance is computed, a dendrogram is constructed to visualize the relative positions among the objects. A dendrogram is a tree diagram and it shows how objects are formed into cluster at a given level of distance. It is a hierarchical clustering tree which can be used to construct clusters at varying level of distance. From the bottom to the top, beginning with two objects with the shortest distance merging to form the first cluster, objects are merged together to form a bigger cluster at each step and at the top of a tree, all objects are grouped into a single cluster. A researcher can freely choose a distance to divide data into clusters or choose the number of cluster and divide the data. For both cases, a researcher practically draws a horizontal line on a dendrogram.

A dendrogram for the Korean stock market is prepared but due to the number of firms in the market, it is impossible to present the figure in a legible way. Therefore, a dendrogram for the 30 largest firms using the Pearson correlation is shown in Fig. 2. When the clustering analysis is performed, all firms in the Korean market are used.

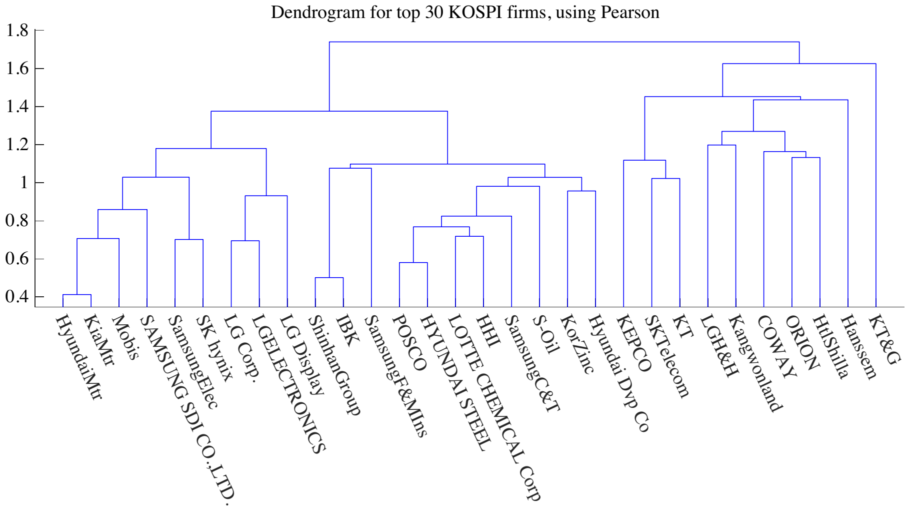

Fig. 2. Dendrogram for 30 largest firms in the Korean market. The Pearson correlation matrix was used to compute the distance between firms. Instead of presenting the entire market, a representative diagram is presented here for illustrative purposes. The analyses performed were based on the similar diagram using the entire Korean market.

The dendrogram is more useful for illustrative purposes; therefore, clustering analysis has to follow to draw a concrete line which then can be used to provide insights regarding the closeness among objects. The number of clusters is usually determined using heuristic methods to minimize the sum of within-cluster variance. However, such study is useful when extensive research has already performed on the subject, which is not the case for clustering partial correlation matrix of stock return of the Korean market. At this point, it is unclear whether the partial correlation coefficients are capable of clustering the Korean stocks into distinct groups and even if the clustering analysis is successful, the resulting clusters may not be consistent with the conventional wisdom, making it harder to interpret the result. Hence, in this study, the firms are divided into ten clusters. Ten clusters are chosen because they can then be directly compared with MSCI's Global Industry Classification Standard (GICS). GICS is an industry classification system developed by MSCI and Standard & Poor's. The firms are reviewed annually to determine its principal business activity and classified into one of ten sectors. Given that the firms in the same GICS sector are involved in a similar business and share similar relationship with other firms in a different sector, one can expect that these firms can have a similar 'correlation profile' and belong to the same cluster.

In Fig. 3, the result of clustering is shown. Ten clusters which are discretized along the vertical axis are shown for each subperiod. The firms are sorted and marked with different colors and symbols based on their GICS sector. If the hypothesis is true, the firms in the same sector could be grouped into the same cluster. However, as it is shown in Fig. 3, the same symbols are found in almost every cluster for every period. It clearly indicates that the GICS sector classification is not the best one to divide the firms if one seeks to minimize the correlation in stock returns.

Table 3 provides the details regarding each cluster found in Fig. 3. The Industry titles given by GICS are shortened to fit into the tables (CD—consumer discretionary, F—financials, I—industrials, IT—information technology, M—materials, CS—consumer staples, E—energy, H—healthcare, T—telecommunication services, U—utilities). The size of a cluster, which was computed by the number of objects in the cluster divided by the total number of objects in the period, is reported and color coded using blue. Remaining portion of the table shows the breakdown of each sector into clusters. Each row represents a sector and the number shown in cells gives a relative portion of sector assigned to a cluster.

Contrary to the hypothesis of the same sector converging into the same cluster, in most cases, the firms in the same sector are not grouped together into a single cluster. There are few cases where more than 50% of the firms from the same sector grouped into the same cluster. One should be cautious when interpreting these results because many of those cases are for the energy, telecomm and utility sectors. These sectors are made of small number of firms (8, 3, and 15 respectively as of 2014) so that the firms from the same sector are likely to be classified into the same cluster by chance.

The overall result implies that whether a firm belongs to a particular sector is not a good predictor of its proximity to other firms. The implication is significant for many practitioners because they usually assume that the firms in the same sector are subject to a similar environment and show similar behavior which makes them close to each other. When a practitioner attempts to diversify her portfolio, she usually does it by picking stocks from different sectors to minimize the correlation between stocks in the portfolio. This study strongly suggests that it may not be sufficient to solely consider sector but consider the correlation of returns because even if the firms were not in the same sector, they may belong to the same correlation cluster and actually be close to each other in terms of stock returns, which prevents diversifying the risk away from the portfolio.

The results in Fig. 3 and Table 3 suggest that the market structure changes over time along with proximity between firms. One may wonder if there are firms that are always clustered together or never close to each other. Due to the size of data it

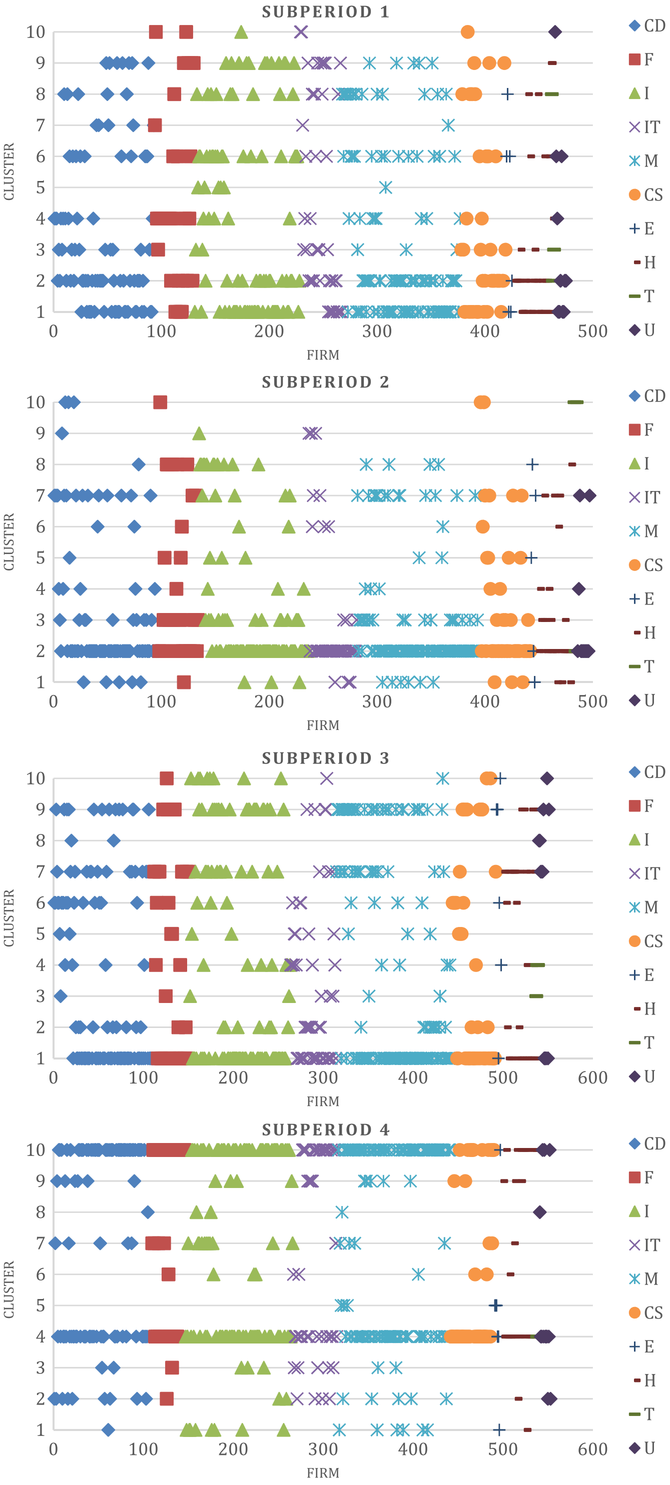

Fig. 3. Clustering results for each period. The horizontal axis represents firms, and it takes discrete value from 1 to the number of firms in the period. The vertical axis is for clusters and it takes values from 1 to 10. Firms are sorted by their sectors, and each sectors are marked with different colors and symbols which are shown in plot legend. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

is infeasible to report the results for all firms used in this study. However, the analysis performed on the 475 firms available since the beginning period of the study show that only a small portion (65 firms to be exact) have more than 10 firms that are always clustered together while 264 firms do not have a single firm that are clustered together for all four subperiods.

Table 3
Members of partial correlation clusters in terms of GICS. The table shows how each sector is divided among the clusters. Each column represents a cluster. The second row, Size, represents the relative size of a cluster against the entire data set for the period of analysis. It is color gradient coded using white (0%) to blue (100%) to represent its size. The remainder of the table, which is color coded in red, shows the relative size of each sector in a cluster to the sector as a whole; hence, each row sums up to 100%. A cell is left blank if it was zero. If the hypothesis of same sector-same cluster was true, each row should have a single dominating cell with most of the sector's constituents grouped into a single cluster. However, except for few cases, the firms in the same sector are divided into several clusters.

Cluster	1	2	3	4	5	6	7	8	9	10
CD	23%	27%	11%	11%		11%	5%	5%	8%	
F	14%	14%	3%	41%		8%	3%	3%	11%	5%
I	32%	16%	2%	5%	5%	17%		9%	12%	1%
IT	18%	21%	13%	5%		8%	3%	10%	18%	5%
M	33%	27%	3%	7%	1%	12%	1%	12%	5%	
CS	26%	29%	12%	5%		10%		10%	7%	2%
E	33%	17%				33%		17%		
H	31%	42%	6%	3%		8%		6%	6%	
T		33%	33%					33%		
U	36%	27%		9%		18%				9%
Size	27%	24%	6%	9%	1%	12%	2%	8%	8%	1%

Subperiod 1

Cluster	1	2	3	4	5	6	7	8	9	10
CD	5%	57%	10%	5%	1%	2%	14%	1%	1%	3%
F	3%	35%	30%	3%	5%	3%	5%	14%		3%
I	3%	58%	15%	3%	3%	2%	5%	11%	1%	
IT	7%	69%	7%			7%	4%		7%	
M	5%	60%	15%	3%	2%	1%	11%	4%		
CS	6%	55%	11%	4%	9%	2%	9%			4%
E	20%	20%			20%		20%	20%		
H	9%	57%	14%	6%		3%	9%	3%		
T		33%								67%
U		75%		8%			17%			
Size	5%	57%	13%	3%	3%	2%	9%	5%	1%	2%

Subperiod 2

Cluster	1	2	3	4	5	6	7	8	9	10
CD	45%	9%	1%	4%	2%	14%	13%	2%	12%	
F	45%	5%	3%	5%	5%	5%	20%		10%	3%
I	44%	6%	2%	6%	2%	3%	12%		19%	7%
IT	39%	14%	6%	12%	8%	6%	6%		8%	2%
M	50%	7%	2%	3%	2%	3%	12%		21%	1%
CS	53%	6%		2%	6%	12%	4%		12%	4%
E	17%			17%		17%			33%	17%
H	42%	5%		5%		5%	32%		11%	
T			33%	67%						
U	33%						25%	17%	17%	8%
Size	45%	7%	2%	5%	3%	7%	13%	1%	15%	3%

Subperiod 3

(continued on next page)

On the other hand, 265 firms turn out to be never clustered with at least 100 firms, suggesting that it is much more likely to find firms that are never clustered together.

Table 3 (*continued*)

Cluster	1	2	3	4	5	6	7	8	9	10
CD	1%	8%	2%	31%			5%	1%	6%	47%
F		3%	3%	27%		3%	11%			54%
I	7%	2%	3%	34%		3%	9%	2%	3%	38%
IT		8%	10%	26%		4%	2%		10%	40%
M	5%	4%	2%	32%	2%	1%	3%	1%	4%	46%
CS				65%		4%	4%		4%	22%
E	14%			29%	43%					14%
H	3%	3%		50%		3%	3%		8%	33%
T				100%						
U		15%		54%				8%		23%
Size	3%	4%	2%	37%	1%	2%	5%	1%	5%	40%

Subperiod 4

Entire	1	2	3	4	5	6	7	8	9	10
CD	57%	17%		12%			7%		8%	
F	27%	4%	15%	4%	27%	8%	12%			4%
I	53%	6%	3%	5%		27%	3%		3%	
IT	39%	25%		7%	4%	4%		4%	18%	
M	65%	11%	1%	7%	3%	13%				
CS	53%	7%		20%	3%		17%			
E		25%		25%		50%				
H	50%	13%	4%	21%		8%			4%	
T				33%			67%			
U	63%						38%			
Size	52%	11%	2%	10%	3%	11%	6%	0%	4%	0%

Entire Period

One may ask a question whether a similar result can be obtained using Pearson correlation. Table 4 is organized in the same manner used to create Table 3, but Pearson correlation coefficients are used for clustering. The most significant difference is that in most cases, the majority of the firms are clustered into one or two clusters together suggesting that a large number of firms are similar to each other. In other words, the Pearson correlation is unable to distinguish firms even when they are involved in a vastly different business, and the clustering analyses result in poorly divided groups. Seeing how the Pearson correlation is driven by a common factor which has a powerful influence on every firm, this result does not come as a surprise.

It could be interesting to see how the color maps used in Fig. 1 look different if the firms are sorted by clusters. Fig. 4 shows the color maps of partial correlation matrix for the entire period of analysis. The left panel is identical to the one used in Fig. 1, sorted by market capitalization. The right panel is sorted by clusters and solid lines were drawn between each cluster. One can easily notice that the firms in the same cluster are positively correlated with each other and form a small red squared box but not necessarily with firms in the other clusters.

T-tests are performed to statistically confirm whether the red squared box is in fact different from the other part of the matrix. Seven clusters are big enough to have a lower triangle part of the red square box greater than 30 entries so the test is performed for these seven clusters. The null hypothesis (H_0) is that the mean value of correlation coefficients in the same cluster is the same with the mean value of the entire matrix minus the cluster part. Table 5 summarizes the test performed on the seven clusters. Except for one case, the mean values of the correlation coefficients of a cluster are different from the rest of the matrix. The confidence interval (CI) gives the upper and lower bounds of the difference of means at 5% significance level.

5. Conclusion

In this paper, the partial correlation analysis is performed on the Korean stock market. The partial correlation analysis is an analysis of co-movement of two random variables when common factors are controlled. This is an important aspect to

Table 4
Members of Pearson correlation clusters in terms of GICS. The table is identical to Table 3 except the Pearson correlation is used for clustering. Compared to the result using the partial correlation, most of the firms clustered into one or two clusters.

Cluster	1	2	3	4	5	6	7	8	9	10
CD			33%	48%	6%	2%	3%	4%	1%	1%
F			41%	51%	3%		3%			3%
I		2%	26%	57%	2%	3%	2%	5%	3%	
IT			31%	54%		3%	3%	5%		5%
M			27%	54%	2%	1%	12%	2%	1%	2%
CS	2%		17%	52%	2%	2%	7%	7%	5%	5%
E			33%	33%			33%			
H			22%	67%			8%	3%		
T			33%	33%		33%				
U			18%	73%	9%					
Size	0%	0%	28%	54%	3%	2%	6%	4%	1%	2%

Subperiod 1

Cluster	1	2	3	4	5	6	7	8	9	10
CD	3%	5%		2%	49%	36%	2%			2%
F	3%	19%	3%		57%	16%			3%	
I	3%	2%		1%	72%	23%				
IT	2%	2%		2%	49%	38%	7%			
M	2%	2%		2%	64%	30%		1%		
CS	4%	4%	4%	4%	53%	19%		2%	9%	
E	40%				20%	40%				
H	6%	3%	3%		51%	37%				
T		33%						33%	33%	
U		8%			58%	33%				
Size	3%	4%	1%	2%	58%	29%	1%	1%	1%	0%

Subperiod 2

Cluster	1	2	3	4	5	6	7	8	9	10
CD	1%	1%	25%	23%		2%	41%	3%	2%	4%
F			13%	30%			48%	5%	5%	
I			19%	21%	2%	3%	50%	1%	2%	4%
IT		2%	37%	24%	2%		27%		8%	
M	1%		22%	12%			63%		3%	
CS		2%	31%	18%		2%	29%	8%	6%	4%
E				17%		17%	67%			
H			21%	18%			53%	3%	3%	3%
T						100%				
U			8%	42%			25%		25%	
Size	0%	1%	23%	20%	1%	2%	46%	2%	4%	2%

Subperiod 3

(continued on next page)

consider in Econophysics as many financial time series are subjected to common factors which may skew the results of an analysis.

In the Korean market, the market index, KOSPI, has a strong influence on stocks, driving correlation between them higher. By removing this common factor, this study successfully demonstrates that many pairs of firms have a negative value of correlation compared to Pearson correlation. A clustering analysis is then performed to group the firms by its proximity measured by Euclidean distance of the partial correlation coefficients. It is hypothesized that firms in the same

Table 4 (*continued*)

Cluster	1	2	3	4	5	6	7	8	9	10
CD	3%	6%	3%	5%	16%	24%	6%	35%	4%	
F		8%		3%	35%	22%		32%		
I	1%	4%	3%	4%	23%	22%	3%	37%	1%	2%
IT	6%	6%	6%	2%	10%	26%	12%	30%	2%	
M	2%	6%	2%	2%	26%	22%	5%	33%	1%	2%
CS	2%	16%	4%		8%	12%	6%	51%		
E					57%			43%		
H		13%	3%	8%	13%	15%	8%	40%		3%
T				33%				67%		
U	8%	8%	8%		23%			46%	8%	
Size	2%	7%	3%	3%	20%	20%	5%	37%	1%	1%

Subperiod 4

Cluster	1	2	3	4	5	6	7	8	9	10
CD	2%	7%	2%	48%	3%	3%	2%	2%	25%	7%
F		12%		4%	12%				73%	
I		5%	6%	26%	5%	3%			55%	2%
IT		11%		36%					39%	14%
M		1%	4%	28%	4%				63%	
CS	3%	20%	7%	33%		10%		3%	23%	
E				25%	50%				25%	
H	4%	4%		58%	4%				29%	
T				33%				67%		
U				25%		25%			50%	
Size	1%	6%	3%	33%	4%	3%	0%	1%	45%	3%

Entire Period

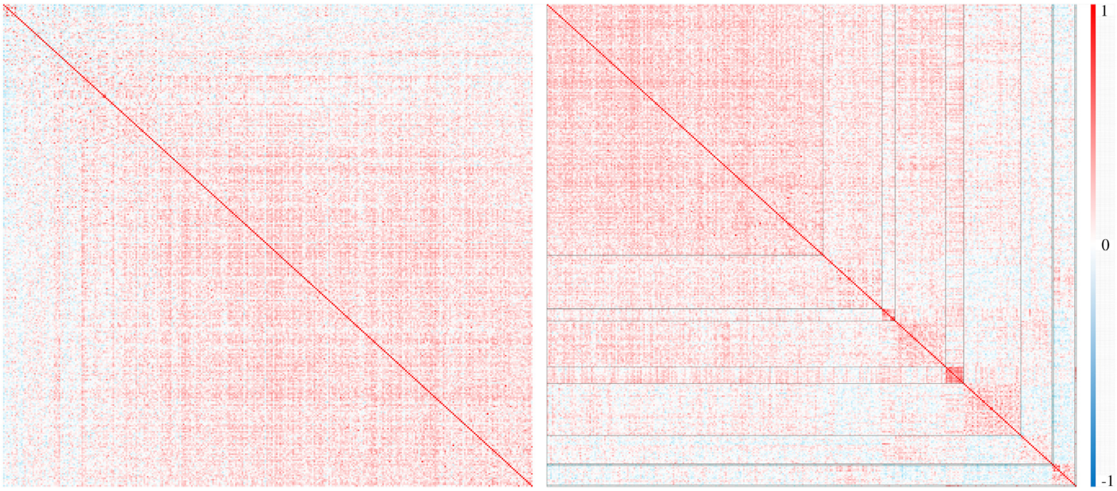

Fig. 4. The partial correlation matrix sorted by market capitalization (left) and by clusters (right). Similar to Fig. 1, 324 firms are used to create the matrix. When sorted by clusters, one can easily identify a few red square blocks. The red square blocks are usually found to be the firms in the same cluster. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

sector, determined by its business, are likely to be grouped together into the same cluster. However, when clustering is done in terms of the firm's partial correlation, all clusters are a mix of firms from different sectors and not one cluster is

Table 5
Two-sample t-tests of the mean value of correlation coefficients in the same cluster vs. the rest of the matrix. Among the ten clusters, three of them (Cluster 3, 8, and 10) were too small to have 30 entries in the lower triangle part so they are omitted from the analysis. For six out of seven cases, the mean values of correlation coefficients in the same cluster were greater than that of the other part of the matrix. The result of the hypothesis test (H_0), p-value, and the 95% confidence interval for the difference of means are reported.

	Cluster 1	Cluster 2	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 9
H_0	Reject	Reject	Reject	Reject	Reject	Fail to reject	Reject
P-value	0	6.8E−06	2.6E−43	4.8E−108	1.1E−54	0.95	4.3E−09
Confidence interval	0.0953, 0.1001	0.0123, 0.0314	0.0668, 0.0889	0.3285, 0.3923	0.0680, 0.0875	−0.0187, 0.0177	0.0538, 0.1076

clearly dominated by a single sector. The implication is significant for practitioners. Many portfolio managers rely on sector diversification to diversify their portfolio and maximize return to risk ratio.

This study shows that, especially for those taking position in their local market only, diversification in regard to sector may not be sufficient because the firms' stock returns may still be correlated with each other. When the firms are sorted by clusters, the correlation of returns shows marked difference between clusters. The firms in the same cluster are mostly positively correlated, but different clusters are more likely to have zero or negative correlation.

Although this study answers some of the questions mentioned in the introduction, much room for further research remains. In this study, only the obvious common factor, the market return, is used to measure the difference between Pearson correlation and Partial correlation, but many other factors still remain which could potentially have a significant influence on the stocks such as foreign exchange rate. Also, in this study, the simplest form of hierarchical clustering approach, an agglomerative clustering method, is used. Since there are many advanced methods of clustering available, it may be possible to explore some other options to better fit the data. The number of clusters is also arbitrarily chosen so that it can be compared with a conventional wisdom, but it is highly likely that using a different number of cluster may result in better clustering results.

References

[1] Z. Ding, C.W.J. Granger, R.F. Engle, A long memory property of stock market returns and a new model, J. Empir. Finance 1 (1993) 83–106.
[2] R. Cont, Empirical properties of asset returns: stylized facts and statistical issues, Quant. Finance 1 (2001) 223–236.
[3] X. Gabaix, P. Gopikrishnan, V. Plerou, H.E. Stanley, A theory of power-law distributions in financial market fluctuations, Nature 423 (2003) 267–270.
[4] B.B. Mandelbrot, Fractals and Scaling in Finance: Discontinuity, Concentration, Risk, Springer, 1997.
[5] J.M. Maheu, T.H. McCurdy, Identifying bull and bear markets in stock returns, J. Bus. Econom. Statist. 18 (2000) 100–112.
[6] W. Gang-Jin, X. Chi, Cross-correlations between WTI crude oil market and US stock market: A perspective from econophysics, Acta Phys. Pol. B 43 (2012) 2021.
[7] Y.J. Zhang, J. Wang, Exploring the WTI crude oil price bubble process using the Markov regime switching model, Physica A 421 (2015) 377–387.
[8] G.M. Caporale, A. Cipollini, N. Spagnolo, Testing for contagion: a conditional correlation analysis, J. Empir. Finance 12 (2005) 476–489.
[9] J.J. Tseng, S.P. Li, Asset returns and volatility clustering in financial time series, Physica A 390 (2011) 1300–1314.
[10] F. Longin, Is the correlation in international equity returns constant: 1960–1990? J. Int. Money Finance 14 (1995) 3–26.
[11] C.S. Eun, S. Shim, International transmission of stock-market movements, J. Finan. Quant. Anal. 24 (1989) 241–256.
[12] T.C. Chiang, B.N. Jeon, H.M. Li, Dynamic correlation analysis of financial contagion: Evidence from asian markets, J. Int. Money Finance 26 (2007) 1206–1228.
[13] K.J. Forbes, R. Rigobon, No contagion, only interdependence: Measuring stock market comovements, J. Finance 57 (2002) 2223–2261.
[14] M. Tumminello, F. Lillo, R.N. Mantegna, Correlation, hierarchies, and networks in financial markets, J. Econ. Behav. Organ. 75 (2010) 40–58.
[15] A. Ang, J. Chen, Asymmetric correlations of equity portfolios, J. Financ. Econ. 63 (2002) 443–494.
[16] H. Markowitz, Portfolio selection, J. Finance 7 (1952) 77–91.
[17] D.Y. Kenett, X.Q. Huang, I. Vodenska, S. Havlin, H.E. Stanley, Partial correlation analysis: applications for financial markets, Quant. Finance 15 (2015) 569–578.
[18] D.Y. Kenett, M. Tumminello, A. Madi, G. Gur-Gershgoren, R.N. Mantegna, E. Ben-Jacob, Dominating clasp of the financial sector revealed by partial correlation analysis of the stock market, PLoS One 5 (2010) e15032.
[19] D.Y. Kenett, Y. Shapira, A. Madi, S. Bransburg-Zabary, G. Gur-Gershgoren, E. Ben-Jacob, Index cohesive force analysis reveals that the US market became prone to systemic collapses since 2002, PLoS One 6 (2011) e19378.
[20] D.Y. Kenett, T. Preis, G. Gur-Gershgoren, E. Ben-Jacob, Dependency network and node influence: application to the study of financial markets, Int. J. Bifurcation Chaos 22 (2012) 14.
[21] L. Uechi, T. Akutsu, H.E. Stanley, A.J. Marcus, D.Y. Kenett, Sector dominance ratio analysis of financial markets, Physica A 421 (2015) 488–509.
[22] J. Lintner, Security prices, risk, and maximal gains from diversification, J. Finance 20 (1965) 587–616.
[23] J. Lintner, The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets, Rev. Econ. Stat. 47 (1965) 13–37.
[24] W.F. Sharpe, Capital-asset prices - a theory of market equilibrium under conditions of risk, J. Finance 19 (1964) 425–442.
[25] R. Roll, A critique of the asset pricing theory's tests part 1: On past and potential testability of the theory, J. Financ. Econ. 4 (1977) 129–176.