Multivariate Time Series Similarity-Based Complex Network in Stocks Market Analysis: Case of NYSE During Global Crisis 2008

Abstract

Maman Abdurachman Djauhari
Institute for Mathematical Research (INSPEM), Universiti Putra Malaysia,
maman_abd@upm.edu.my

Gan Siew Lee
Department of Mathematical Sciences, Universiti Teknologi Malaysia,
slgan3@live.utm.my

Long before we started with the 21st millennium, Stephen Hawking saw the current millennium as the millennium of complex systems. Until present, he was right due to the fast growing technology in computer. Nowadays, in the era of digital world where big data is our daily menu, we cannot escape from complex systems. As big data is characterized by "4V" (Variety, Velocity, Veracity and Volume), statistics such as practiced in traditional way is not enough and sometime is not apt to be used to understand the most important information contained in big data. What people call now data analytics needs to be used as the only complementary. It is mathematically dominated by multivariate data analysis (MVDA) in the French way. Traditional statistics, which is based on mathematical statistics, is to do confirmatory analysis while data analytics is to do exploratory analysis. The former is to do hypothesis testing (micro analysis) and the latter is for hypothesis generation (macro analysis). Macro analysis is more appropriate to deal with big data. The principal mathematical tool to do macro analysis is MVDA in the French way where big data is considered as a complex system. In this regards, the main problem is to define the similarity among objects of the study such as stocks, economic sectors, currencies, and other commodities in financial industry, which are statistically a multivariate time series. Furthermore, the principal tools to filter the important information contained in a complex system are complex network and social network analysis. To demonstrate the advantages of complex network approach in stocks market analysis, in this paper the behaviour of economic sectors played in NYSE during global crisis in 2008 will be presented and discussed. By nature, all stocks are a multivariate time series. Therefore, in that example, we show that the use of Pearson correlation coefficient is useless to define the similarity among them. We use Escoufier's vector correlation coefficient instead.

Keywords: minimal spanning tree, network centrality, stocks network, stock's prices, vector correlation

1. Introduction

The inter-relationships among stocks in a given portfolio represent a complex system of correlation structure. It is numerically represented in the form of a correlation matrix. In practice, it is then a complex network where its level of

complexity is of order $O\left(n^2\right)$. Here n is the number of stocks. As a consequence, when stocks market becomes a big dataset in terms of n, the correlation structure becomes harder and harder to understand. How can we transform such complex network into important economic information? To answer this question, the tools developed in the field of econophysics, see Mantegna and Stanley (2000), are very powerful.

In the current practice, stock is usually represented by its closing price. It is then a univariate (UV) time series which is customarily assumed to be governed by geometric Brownian motion law. This means that the price returns are independent and identically log-normally distributed. In other words, the log returns are independent and identically normally distributed (i.i.n.d). This fundamental assumption is the theoretical basis in stocks network analysis. Under this assumption, the similarity among stocks can be measured in terms of Pearson correlation coefficient (PCC) of log returns. Based on PCC, the stocks network is constructed as dissimilarities network and the important economic information is filtered by using minimal spanning tree (MST).

However, in daily market activities, stock is represented by its opening, highest, lowest and closing (OHLC) prices. This means that stock is a multivariate (MV) time series of those prices. In this case, since PCC is not apt anymore to measure the similarity among stocks, we show that the use of Escoufier vector correlation (EVC) is more advantageous. It generalizes PCC from bivariate into multivariate case. EVC, originally introduced by Escoufier (1973), quantifies the linear relationship of two random vectors while PCC is about two random variables. Nowadays, its application can be found in many areas of statistics. However, to the best of our knowledge, its use in stocks network analysis is still at the beginning as can be seen in Kazemilari and Djauhari (2015) and Djauhari and Gan (2015).

Those authors have showed that MV could describe well the real market situation. For example, (i) in terms of the number of worst performance stocks (leaves), and (ii) the phenomenon of social embeddedness that cannot be detected by using closing price only. This phenomenon, see Halinen and Tornroos (1998), is very important in the study of stocks behaviour under similar management.

Since highest and lowest prices can take place at any time during the trading day, to synchronize the effect of OHLC prices, it is customary to use weekly data as can be seen, for example, in Lo and MacKinlay (1990). Furthermore, like in UV case, MST is used to filter the topological structure of stocks network based on OHLC prices. To construct MST from dissimilarities network, since Kruskal's algorithm or Prim's is computationally slow for large n, we use the algorithm proposed in Djauhari and Gan (2013).

To show the advantages of MV approach compared to the standard UV approach, a case study on NYSE data during global crisis in 2008 has been conducted. We compare the MSTs issued from those approaches in terms of degree centrality measure since it represents market index as remarked, for example, in Kenett et al. (2013).

The rest of the paper is organized as follows. We start by introducing in Section 2 the notion of similarity among stocks each of which is defined by its OHLC prices and stocks network is constructed. Section 3 presents the MSTs of NYSE one year before, during, and one year after global crisis. Some evidences from MST, Jaccard Index, as well as degree centrality will be reported to show the advantages of MV approach. Concluding remarks will be highlighted in the last section.

2. Stocks network in multivariate setting

Let $p_i(t,1)$, $p_i(t,2)$, $p_i(t,3)$ and $p_i(t,4)$ denote the opening, highest, lowest and closing prices of stock i; $i = 1, 2, \ldots, n$. We write,

$$r_i(t,m) = \ln p_i(t,m) - \ln p_i(t-1,m) \quad (1)$$

the log return of the m-th price of stock i at time t; $m = 1, 2, 3, 4$. Under the assumption that each price is a GBM process for all stocks, $r_i(t,m)$ are i.i.n.d for all m. Therefore, $r_i(t,m)$ in (1) can be viewed as the m-th component of a random vector. This viewpoint leads us to consider stock as a multivariate entity.

2.1. Similarity among stocks

Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two random vectors representing two different stocks each of dimension $p = 4$ and $q = 4$. We denote $r_{\boldsymbol{X}}(t,m)$ and $r_{\boldsymbol{Y}}(t,m)$ the m-th component of $\boldsymbol{X}$ and $\boldsymbol{Y}$; $m = 1, 2, 3, 4$, and T the length of time support for the four prices. Let $S_{\boldsymbol{XX}}$, $S_{\boldsymbol{YY}}$ and $S_{\boldsymbol{XY}}$ be the sample covariance matrix of $\boldsymbol{X}$, $\boldsymbol{Y}$, and between $\boldsymbol{X}$ and $\boldsymbol{Y}$ are, In this circumstance, by using EVC, the correlation between the random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ is,

$$RV_{\boldsymbol{XY}} = \frac{Tr\left(S_{\boldsymbol{XY}} S_{\boldsymbol{YX}}\right)}{\sqrt{Tr\left(S_{\boldsymbol{XX}}^2\right)}\sqrt{Tr\left(S_{\boldsymbol{YY}}^2\right)}}. \quad (2)$$

Like PCC, this coefficient is the cosine of an angle spanned by two stocks. It satisfies,

(i) $0 \leq RV_{\boldsymbol{XY}} \leq 1$. It is 1 if $\boldsymbol{X}$ and $\boldsymbol{Y}$ have the same correlation structure and it is 0 if each price of one stock is uncorrelated with all prices of the other stock.

(ii) In bivariate case, $RV_{\boldsymbol{XY}}$ is the squared of PCC.

In terms of these properties, therefore, EVC defines the similarity among stocks in MV setting.

2.2. Stocks network

Let $\boldsymbol{S}_{ij}$ be the covariance matrix between stocks *i* and *j*. According to (2), the *RV* coefficient of these stocks is RV_{ij} obtained by substituting $\boldsymbol{X} = i$ and $\boldsymbol{Y} = j$. If we consider a matrix of size $(n \times n)$ with RV_{ij} as *i*-th row and *j*-th column element, this matrix is symmetric with all diagonal elements equal to 1 and the off-diagonal elements are between 0 and 1. It then represents similarities network among stocks in MV setting of OHLC prices. This generalizes then the notion of correlations network such as presented in Mantegna and Stanley (2000), Bonanno et al. (2003), and Galazka (2011) into the notion of vector-correlations network. By using the idea in Mantegna and Stanley (2000), to analyse that network, we define the distance between two stocks *i* and *j*,

$$\delta_{ij} = \sqrt{2\left(1 - RV_{ij}\right)}. \tag{3}$$

If we denote **D** the matrix of size $(n \times n)$ with δ_{ij} in (3) as the element of its *i*-th row and *j*-th column, then **D** represents the dissimilarities network among stocks that we required.

To filter the information contained in **D**, like in UV case, the method of MST is used. For that purpose, due to computational complexity, we use the algorithm developed in Djauhari and Gan (2013) instead of Kruskal's algorithm or Prim's. Furthermore, the topological properties of MST is analysed in terms of market index.

3. Results on NYSE

NYSE 100 most capitalized stocks are analysed one year before, during, and one year after global crisis in 2008. Data were downloaded from this link: http://www.nyse.com/about/listed/nyid_components.shtml. Four stocks were removed from the analysis because of the incompleteness of data. Therefore, only 96 stocks were analysed.

3.1. Evidence from MST

Based on weekly data of OHLC prices, NYSE 96 most capitalized stocks network is constructed. In Figure 1, the dynamics of half yearly MST is presented from the period of (a) January-June 2007, (b) July-December 2007, (c) January-June 2008, (d) July-December 2008, (e) January-June 2009, and finally (f) July-December 2009. All artworks in this figure are drawn using *Pajek* software, downloaded from http://www.mrvar.fdv.uni-lj.si/pajek/.

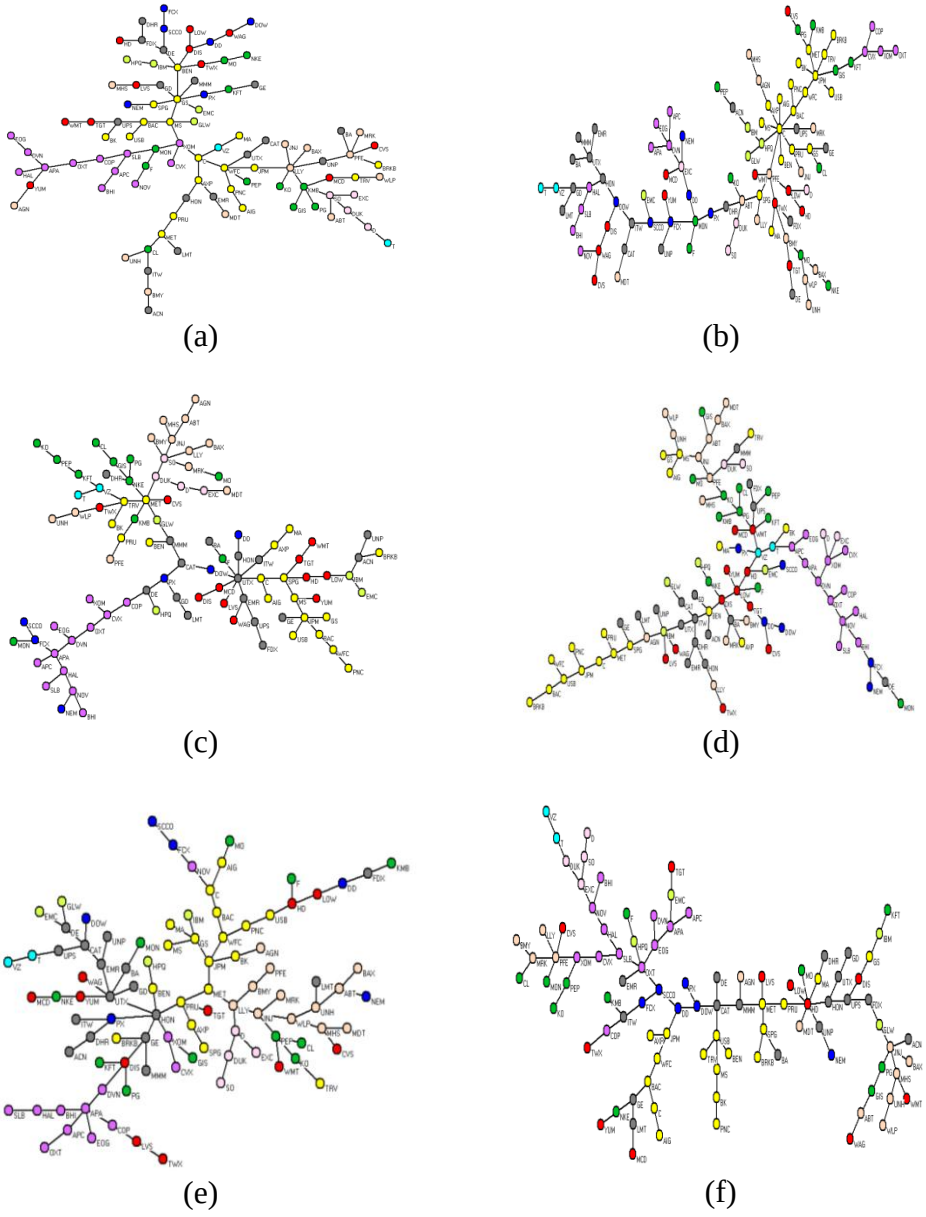

Figure 1. Dynamics of MSTs in Jan-Jun 2007 (a), Jul-Dec 2007 (b), Jan-Jun 2008 (c), Jul-Dec 2008 (d), Jan-Jun 2009 (e), and Jul-Dec 2009 (f) with the following legend:

Colour	Economic sector	Total
Yellow	Financials	18
Light-Orange	Health Care	12
Pine-Green	Consumer Goods	11
Grey	Industrials	16
Thistle	Oil & Gas	12
Red	Consumer Services	11
Pink	Utilities	4
Blue	Basic Materials	6
Green-Yellow	Technology	4
Cyan	Telecommunications	2

The colour of the node (stock) refers to the economic sector where it belongs. At a glance, we see at Figure 1 that one year before the crisis the centre of the network is dominated by Financials. The performance of this sector was dominant in NYSE. However, it was not so anymore during the crisis. In this period all colours are distributed more randomly. The situation becomes severe in the second half of 2008 where Financials moved to the periphery. When the crisis was over, this sector strengthened again. But, in the second half of 2009 it worsened again. This is perhaps due to the Greek debt crisis.

This result is totally different from that given by Djauhari and Gan (2014) who use closing price only. To show other advantages of MV approach, in the next sub-sections the two results will be compared in terms of Jaccard index and degree centrality.

3.2. Evidence from Jaccard index

Jaccard index is used to measure the similarity among MSTs. Here, for each period, we compare MST issued from MV approach and the one given by UV approach. This index reflects the discrepancy in correlation structure. For period $i = 1, 2, \ldots, 6$, see Djauhari and Gan (2014), it is defined by,

$$I_i = \frac{\left|MST_{UV,i} \cap MST_{MV,i}\right|}{\left|MST_{UV,i} \cup MST_{MV,i}\right|} \tag{4}$$

Here, $MST_{UV,i}$ and $MST_{MV,i}$ are the MSTs of the period i issued from UV and MV approaches, respectively, and $|*|$ is the number of elements in the set *. The value of Jaccard index in (4) is between 0 and 1. It is 0 if there is totally no accordance between the two MSTs and it is 1 if they are the same. From NYSE data, the value of this index is very small for each period. The smallest is 0.1377 (Jan-Jun 2007) and the largest is only 0.2338 (Jan-Jun 2009). This indicates that the MST issued from UV approach is really different from that given by MV approach in all periods. This is the advantages of MV approach based on OHLC prices in stocks network analysis.

3.3. Evidence from degree centrality

In this sub-section, the results of MV approach and UV will be compared in terms of number of leaves, diameter of MST, and degree centrality.

3.3.1. Number of leaves and diameter of MST

A leaf is a stock of degree one. It represents a worst performance stock in the market in terms of the number of other stocks being directly linked with. On the other hand, diameter of MST represents the longest path used to propagate the influence of a particular worst stock (called pole) to another pole. Data NYSE show that the number of worst stocks in MV setting is always less than in UV and the diameter of MST under MV is generally greater than under UV. This is a natural consequence from the fact that, unlike in UV setting which focuses on the closing price, each stock in MV brings all information about OHLC prices.

3.3.2. Diameter between and within economic sectors in the MST

Other advantages reveal in the study on,

(i) Diameter between sectors which measures the closeness of a particular sector A to all other sectors in MST in terms of maximum linkage. In other words, it is the sum of diameter of the union $A \cup B$ for all other sector B in MST.

(ii) The cohesiveness among stocks in a given sector in terms of diameter within sector.

The results show that Financials and Industrials are the two most influential sectors before, during, and after the crisis. However, during the second half of 2008, the leadership of Finance was replaced by Industry. Furthermore, severe turbulence occurs among stocks in Financials which is not the case in Industrials sector.

By using the same method, UV approach cannot describe the above market situation. Under this approach, leader in all periods is Industrials and not Financials, and turbulence among stocks also occurs but not severe like under MV. This does certainly not reflect the real market situation.

3.3.3. Degree centrality of economic sector

Degree centrality of an economic sector is defined as the average of all stocks' degree in that sector. By using the same method described in the previous paragraph, we find that MV approach based on OHLC prices presents more dynamics degree centrality of economic sector compared to the case where only closing price is considered.

4. Concluding remarks

This paper is to show that stocks network analysis based on OHLC prices, where similarity among stocks is defined using EVC, is more advantageous than based on closing price only. In the case of NYSE data, we show,

(i) The evidence from Jaccard index in each period. The results issued from both MV and UV approaches are totally different.

(ii) The evidence from the number of worst stocks and the diameter of MST. MV setting reflects the real market situation.

(iii) The dynamics of each sector in terms of degree centrality during the six periods of study can be seen more clearly under OHLC prices-based rather than closing price-based.

(iv) In real situation, Financials sector is at the centre (the leader) in NYSE. This can be described nicely by using MV approach but not UV.

Acknowledgement

The first author is very grateful to the Institute for Mathematical Research (INSPEM), Universiti Putra Malaysia, for providing research facilities during his service as Research Fellow. The second author thanks the Universiti Teknologi Malaysia for the facility to support her PhD program.

References

1. Bonanno, G., Caldarelli, G., Lillo, F. and Mantegna, R.N. (2003). "Topology of correlation-based minimal spanning trees in real and model markets", *Physical Review E*, 68(4), p. 046130-046133.
2. Djauhari, M.A. and Gan, S.L. (2013). "Minimal spanning tree problem in stock network analysis: An efficient algorithm", *Physica A*, 392(9), p. 2226-2234.
3. Djauhari, M.A. and Gan, S.L. (2014). "Dynamics of correlation structure in stock market", *Entropy*, 16(1), p. 455-470.
4. Djauhari, M.A. and Gan, S.L. (2015). "Bursa Malaysia stocks market analysis: A review", *Journal of Science – Academy of Science Malaysia*, 8(2), p. 2226-2234.
5. Escoufier, Y. (1973). "Le traitement des variables vectorielles," *Biometrics*, 29(4), p. 751-760.
6. Galazka, M. (2011). "Characteristics of Polish stock market correlations," *International Review of Financial Analysis*, 20, p. 1-5.
7. Halinen, A. and Tornroos, J. (1998). "The role of embeddedness in the evolution of business network", *Scandinavian Journal of Management*, 14(3), p. 187–205.
8. Kazemilari, M. and Djauhari, M.A. (2015). "Correlation network analysis for multi-dimensional data in stocks market", *Physica A*, 429, p. 62-75.
9. Kenett, D. Y., Ben-Jacob, E., Stanley, H. E. and Gur-Gershgoren, G. (2013). "How high-frequency trading affects a market index", *Nature Scientific Reports*, 3, 2110(1)–(8).
10. Lo, A. W. and MacKinlay, A. C. (1990). "An econometric analysis of nonsynchronous trading", *Journal of Econometrics*, 45, 181-212.
11. Mantegna, R.N. and Stanley, H.E. (2000). *An introduction to econophysics: Correlations and complexity in finance*, Cambridge University Press, Cambridge, UK.