Physica A 429 (2015) 62–75

Contents lists available at ScienceDirect

Physica A

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

Correlation network analysis for multi-dimensional data in stocks market

Mansooreh Kazemilari, Maman Abdurachman Djauhari *

Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor Bahru, Malaysia
Institute for Mathematical Research (INSPEM), Universiti Putra Malaysia, 43400 UPM Serdang, Selangor Darul Ehsan, Malaysia

HIGHLIGHTS

- Stocks network based on opening, highest, lowest, and closing prices is introduced.
- Escoufier's *RV* coefficient is used to measure the similarity among stocks.
- Information in multi-dimensional network is filtered using minimal spanning tree.
- The topological properties of stocks are analyzed by using centrality measures.
- 30 Dow-Jones stocks are used to illustrate the advantage of *RV* coefficients network.

ARTICLE INFO

Article history:
Received 5 February 2014
Received in revised form 7 November 2014
Available online 21 February 2015

Keywords:
Multivariate association
Escoffier's operator
Stocks network analysis
Vector correlation

ABSTRACT

This paper shows how the concept of vector correlation can appropriately measure the similarity among multivariate time series in stocks network. The motivation of this paper is (i) to apply the *RV* coefficient to define the network among stocks where each of them is represented by a multivariate time series; (ii) to analyze that network in terms of topological structure of the stocks of all minimum spanning trees, and (iii) to compare the network topology between univariate correlation based on *r* and multivariate correlation network based on *RV* coefficient.

1. Introduction

The system of stocks market is extremely complex and it is continuously evolving through various heterogeneous interactions between them. Thus, to capture the pricing mechanism of stocks market, it is important to study and figure out the interactions. In this regard, the correlation matrix has long been used to quantify the interactions, and the information generated can be very helpful if enough data has been provided beforehand. However, the extraction of information from the correlation matrix is not as straightforward as it seems [1].

In econophysics, stocks network was introduced by Mantegna [2] for investigating the interaction between stocks using the minimal spanning tree (MST) method. The stocks network visually constructs the relationship between stocks, which is extracted by the MST based on the correlations between stocks' price returns [1]. See also Mantegna and Stanley [3] for a more pedagogical approach.

* Corresponding author at: Institute for Mathematical Research (INSPEM), Universiti Putra Malaysia, 43400 UPM Serdang, Selangor Darul Ehsan, Malaysia.
E-mail addresses: kmansooreh2@live.utm.my (M. Kazemilari), maman_abd@upm.edu.my (M.A. Djauhari).

http://dx.doi.org/10.1016/j.physa.2015.02.052
0378-4371/

Since the work of Mantegna [2], MST has become an indispensable tool in econophysics to filter important information contained in the complex structure of a correlation matrix among stocks in a given portfolio [4]. Some examples of this method were used in the context of financial markets by Bonanno et al. [5], Jung et al. [6], Garas et al. [7], Gilmore et al. [8], Pozzi et al. [9], Tabak et al. [10], and Sieczka et al. [11]. MST has also been used to analyze the properties of FX markets based on correlation in different methods by many authors such as McDonald et al. [12], Brida et al. [13], Ortega and Matesanz [14], Mizuno et al. [15], Naylor et al. [16], Kwapień et al. [17], Keskin et al. [18], Jang et al. [19], and Wang et al. [20,21]. In all of those studies, stock or currency is represented by univariate time series its price. All of the models of stocks market have focused on the behavior of price returns and losing the possibility of embodying information from opening price, highest price and lowest price. The importance of opening, highest, lowest and closing (OHLC in short) prices information is well-known and has been used by a number of authors, including Rogers and Satchell [22], Rogers et al. [23], Rogers [24], Rogers and Zhou [25] and Horst et al. [26].

Nonetheless, in finance, many researchers use the methods of multivariate statistical analysis such as factor analysis, principal component analysis, and cluster analysis. However, the use of vector correlation to describe the similarity among multivariate time series representing stocks is new. Vector correlation, for the first time, was introduced by Hotelling in 1936 [27]. After that, correlation coefficient between two sets of complex vectors was considered by Masuyama in Refs. [28,29] and Rozeboom in 1965 [30]. Then, the measurement of correlation among vector variables was studied by Kshirsagar [31], Escoufier [32], Coxhead [33], Cramer [34], Shaffer and Gillo [35], Cramer and Nicewander [36], Stephens [37], Ramsay et al. [38], Robert et al. [39], and Roy and Cléroux [40]. The majority of the proposed multivariate correlations are actually a function of the canonical correlations. Escoufier [32] and Stephens [37] introduced a vector correlation coefficient by assuming that two sets of vector were perfectly correlated if there exists an orthogonal transformation that could facilitate their collision. Robert and Escoufier [41] also showed that the Escoufier vector correlation (EVC) or, equivalently, *RV* coefficient can be used as a unifying tool in multivariate methods.

Escoufier [32] and Robert and Escoufier [41] demonstrated the use of *RV* coefficient for the presentation of different methods in multivariate analysis. Mathematically, it is a solution of an optimization problem under various constraints. Its application can be seen in several statistical techniques such as STATIS and DISTATIS [42]. More recently, Smilde et al. [43] applied *RV* coefficient in metabolomics where they introduce the idea of matrix correlations analysis to the bioinformatics community.

The present research is the first one that considers the similarity measure among multi-dimensional stocks of OHLC prices by using *RV* coefficient as a theoretical basis to analyze stocks market behavior in multivariate setting. We analyze the correlations network of multidimensional prices and then compare this network with the networks of all individual stock's price. For that purpose, MST is used to filter important information contained in those complex networks. As a real example, 30 Dow-Jones stocks from May 7, 2007 to April 30, 2012 will be used.

2. Methodology

In order to measure similarities and differences in the synchronous time evolution of a pair of stocks, we study the correlation between the weekly logarithmic changes in OHLC prices of two multi-dimensional stocks i and j. Logarithmic change refers to the successive differences of the natural logarithm of price, which is defined by,

$$V_i(t) = \ln\ Z_i(t+1) - \ln\ Z_i(t) \tag{1}$$

where $Z_i(t)$ is the rate of stock i at time t; $i = 1, 2, \ldots, N$, and N is the number of stocks under study. Since each stock is characterized by its OHLC prices, we write

$$V_i(t,k) = \ln\ Z_i(t+1,k) - \ln\ Z_i(t,k); \quad k = 1, 2, 3, 4,$$

where $V_i(t,1), V_i(t,2), V_i(t,3)$, and $V_i(t,4)$ denote the log change of stock i at time t for opening, highest, lowest, and closing prices, respectively. To measure the similarity of stocks i and j based on those four dimensions O, H, L, and C, let n be the length of time support of all time series.

In what follows we introduce the notion of similarity between stocks i and j by using Escoufier's *RV* coefficient. This measure generalizes the similarity among stocks represented by their closing prices only. Let $s_{ij}(k,l) = \langle V_i(k)V_j(l)\rangle - \langle V_i(k)\rangle \langle V_j(l)\rangle$ where $\langle V_i(k)\rangle$ is the average of $V_i(t,k)$ for all t. Statisticians call it the covariance between the kth log price returns of stock i and lth log price returns of stock j with $k, l = 1, 2, 3$, and 4 refer to O, H, L, and C prices. In special case, $s_{ii}(k,k)$ is called the variance of the kth log price returns of stock i. Therefore, $c_{ii}(k,l) = \frac{s_{ii}(k,l)}{\sqrt{s_{ii}(k,k)s_{ii}(l,l)}}$ is Pearson correlation coefficient (PCC) of the kth and lth log prices returns in stock i. The matrix $\boldsymbol{S}_{ii}$ of size (4×4), with $s_{ii}(k,l)$ as its element at the kth row and lth column, is called covariance matrix involving all OHLC prices of stock i. Furthermore, we write $\boldsymbol{S}_{ij}$ the matrix of size (4×4), with $s_{ij}(k,l)$ as its element at the kth row and lth column. It is called covariance matrix of stocks i and j.

By using those notations, now, EVC or, equivalently, *RV* coefficient can be used to define the "correlation" of stocks i and j. By using the definition of *RV* coefficient [32,41], then the "correlation" of stocks i and j is,