CORRELATION BASED HIERARCHICAL CLUSTERING IN FINANCIAL TIME SERIES

S. MICCICHE', F. LILLO AND R. N. MANTEGNA
Dipartimento di Fisica e Tecnologie Relative
Università di Palermo
and
Istituto Nazionale per la Fisica della Materia, Unità di Palermo
Viale delle Scienze - Edificio 18,
I-90128, Palermo, Italy

We review a correlation based clustering procedure applied to a portfolio of assets synchronously traded in a financial market. The portfolio considered consists of the set of 500 highly capitalized stocks traded at the New York Stock Exchange during the time period 1987-1998. We show that meaningful economic information can be extracted from correlation matrices.

1. Introduction

The presence of a high degree of cross-correlation of the synchronous time evolution of a set of equity returns is a well known empirical fact observed in financial markets \(^1,2,3\). For a time horizon of one trading day correlation coefficient as high as 0.7 can be observed for some pair of equity returns belonging to the same economic sector.

The study of cross-correlation of a set of financial equities has also practical importance since it can improve the ability to model composed financial entities such as, for example, stock portfolios. There are different approaches to the study of asset cross-correlation. The most common one is the principal component analysis of the correlation matrix of the data \(^4\). An investigation of the properties of the correlation matrix has been performed by physicists by using ideas and theoretical results of the random matrix theory \(^5,6\). Another approach is the correlation based clustering analysis which allows to obtain clusters of stocks starting from the time series of price returns. Different algorithms exist to perform cluster analysis in finance \(^7,8,9,10,11,12,13\).

In previous work, some of us have shown that a specific correlation based clustering method gives a meaningful taxonomy for stock return time series \(^8,14,15\), for market index returns of worldwide stock exchanges \(^16\) and for volatility increments of stock return time series \(^17\). Here we review the method proposed by applying it to the set of 500 highly capitalized stocks traded in the New York Stock Exchange.
2. A correlation-based filtering procedure

In Ref. 8, it has been proposed a correlation based method able to detect economic information present in a correlation coefficient matrix. This method is a filtering procedure based on the estimation of the subdominant ultrametric 18 associated with a metric distance obtained form the correlation coefficient matrix of set of $n$ stocks. This procedure, already used in other fields, allows to extract from it a minimum spanning tree (MST) and a hierarchical tree from a correlation coefficient matrix by means of a well defined algorithm known as nearest neighbor single linkage clustering algorithm 19. This allows to reveal topological (throughout the MST) and taxonomic (throughout the hierarchical tree) aspects of the correlation present among stocks.

The MST is obtained by filtering a relevant part of the information which is present in the correlation coefficient matrix of the original time series of stock returns. This is done (i) by determining the synchronous correlation coefficient of the difference of logarithm of stock price computed at a selected time horizon, (ii) by calculating a metric distance between all the pair of stocks and (iii) by selecting the subdominant ultrametric distance associated to the considered metric distance. The subdominant ultrametric is the ultrametric structure closest to the original metric structure 18.

The correlation coefficient is defined as

$$\rho_{ij}(\Delta t) = \frac{\langle r_ir_j \rangle - \langle r_i \rangle \langle r_j \rangle}{\sqrt{\langle r_i^2 \rangle - \langle r_i \rangle^2} \langle r_j^2 \rangle - \langle r_j \rangle^2}}$$

(1)

where $i$ and $j$ are numerical labels of the stocks, $r_i = \ln P_i(t) - \ln P_i(t - \Delta t)$, $P_i(t)$ is the value of the stock price $i$ at the trading time $t$ and $\Delta t$ is the time horizon which is, in the present work, one trading day. The correlation coefficient for logarithm price differences (which almost coincides with stock returns) is computed between all the possible pairs of stocks present in the considered portfolio. The empirical statistical average, indicated in this paper with the symbol $\langle \rangle$, is here a temporal average always performed over the investigated time period.

By definition, $\rho_{ij}(\Delta t)$ can vary from -1 (completely anti-correlated pair of stocks) to 1 (completely correlated pair of stocks). When $\rho_{ij}(\Delta t) = 0$ the two stocks are uncorrelated. The matrix of correlation coefficient is a symmetric matrix with $\rho_{ii}(\Delta t) = 1$ in the main diagonal. Hence for each value of $\Delta t$, $n(n-1)/2$ correlation coefficients characterize each correlation coefficient matrix completely.

A metric distance between pair of stocks can be rigorously determined 20 by defining

$$d_{i,j}(\Delta t) = \sqrt{2(1 - \rho_{ij}(\Delta t))}.$$  

(2)

With this choice $d_{i,j}(\Delta t)$ fulfills the three axioms of a metric – (i) $d_{i,j}(\Delta t) = 0$ if and only if $i = j$; (ii) $d_{i,j}(\Delta t) = d_{j,i}(\Delta t)$ and (iii) $d_{i,j}(\Delta t) \leq d_{i,k}(\Delta t) + d_{k,j}(\Delta t)$. 


Figure 1. Minimum spanning tree of 500 highly capitalized stocks traded in the NYSE. The filtering procedure is obtained by considering the correlation coefficient of stock returns time series computed at a 1 trading day time horizon. Each circle represents a stock. Clustered areas are observed. With the letters A, B, C and D we indicate clusters of stocks belonging to the sectors “Finance, insurance, and real estate”, “Transportation, communications, electric and sanitary services”, “Manufacturing” and “Mining and construction” respectively. Some of these stocks act as a “hub” of a local cluster. Examples are GE, TY, C, ITW, KO, and DUK. General Electric is the most connected stock observable at the center of the star like area at the center of the graph.

The distance matrix $D(\Delta t)$ is then used to determine the MST connecting the $n$ stocks.

The MST, a theoretical concept of graph theory 21, is the spanning tree of shortest length. A spanning tree is a graph without loops connecting all the $n$ nodes with $n - 1$ links. We have seen that the original fully connected graph is metric with distance $d_{i,j}$. Therefore the MST selects the $n - 1$ stronger (i.e. shorter) links which span all the nodes. The MST allows to obtain, in a direct and essentially unique way, the subdominant ultrametric distance matrix $D^<(\Delta t)$ and the hierarchical organization of the elements (stocks in our case) of the investigated data set.

The subdominant ultrametric distance between objects $i$ and $j$, i.e. the element $d^<_i j$ of the $D^<(\Delta t)$ matrix, is the maximum value of the metric distance $d_{k,l}$ detected by moving in single steps from $i$ to $j$ through the path connecting $i$ and $j$ in the MST. The method of constructing a MST linking a set of $n$ objects is direct and it is known in multivariate analysis as the nearest neighbor single linkage cluster.
A pedagogical exposition of the determination of the MST in the context of financial time series is provided in ref. Subdominant ultrametric space has been fruitfully used in the description of frustrated complex systems. The archetype of this kind of systems is a spin glass.

As an example of the results obtained with this method here we briefly discuss the results obtained by investigating a set of 500 highly capitalized stocks traded in the New York Stock Exchange (NYSE) during the period January 1987 - December 1998.

Figure 2. Minimum spanning tree of 500 highly capitalized stocks traded in the NYSE. In this figure all the stocks belonging to the “Finance, insurance and real estate” sector are indicated with a black circle. Their identity is indicated by their tick symbol. Two major homogenous areas related to stock belonging to this sector are observed in the graph. A few isolated stocks are also present. The homogeneous area observed in the top area of the graph is composed by insurance companies (AOC, AET, CI, HSB, PGR, PL, SPC, and TMK). The other homogeneous area of financial stocks located at the bottom mostly comprises depository institutions and security and commodity brokers. Examples are the stocks BAC, BK, C, CMB, JPM, MER, CMB, PNC, and RNB

In Fig. 1 we show the minimal spanning tree obtained in this investigation with a time horizon equal to one trading day. Stocks are identified with circles. From the figure it can be seen that clusters of stocks exist. Some of these clusters are composed by stocks which are rather homogeneous with respect to the economic sector of activity. The classification of the stocks used here is the one of the Standard
Industrial Classification system. 24.
To support our statement we show in Fig. 2 the MST by evidencing all the stocks belonging to the “Finance, insurance and real estate” sector. Stocks are identified by black circles and labeled with their tick symbol. Information about the identity of most of the stocks can be found in financial web sites. A few stocks are no more active after 1998 due to merging with other companies or to other reasons.

Figure 3. Minimum spanning tree of 500 highly capitalized stocks traded in the NYSE. All the stocks belonging to the “Transportation, Communications, Electric, Gas, and Sanitary Services” sector are indicated with a black circle. Their identity is indicated by their tick symbol. Four major homogenous areas related to stock belonging to this sector are observed in the graph. The biggest one (bottom right part of the graph) comprises stocks belonging to the electric gas and sanitary services subsector. The second cluster, observed in the left bottom part of the graph, is composed by AIT, BEL, BLS, GTE, SBC, and USW. All these stocks are telephone communication companies. The third and fourth cluster are observed in the left top part of the graph. Both clusters are related to transportation. Specifically, one comprises stocks dealing with transportation by air (ALK, AMR, DAL, LUV, and U) whereas the second one comprises stocks of railroad transportation (CSX, NSC and UNP).

A wide cluster is observed at the bottom of the graph. Another cluster can be detected in the upper part of the graph. In addition to this two prominent clusters and a few isolated stocks are also present. By analyzing in detail the main activity of the companies investigated a rational explanation of the observed behavior is
obtained. In fact the cluster observed in the top area of the graph is composed by insurance companies and the other homogeneous area of financial stocks located at the bottom mostly comprises depository institutions and security and commodity brokers.

A second example is provided in Fig. 3 where we show the stocks belonging to the SIC sector “Transportation, Communications, Electric, Gas, and Sanitary Services”. The title of this sector manifests itself that this classification is unavoidably comprising stocks characterized by a rather different economic activity. This aspect is reflected in the clusters detected in the MST. Four well defined clusters are directly observable. The biggest one is seen at the bottom right of the graph. Most of the stocks that are located here belong to the SIC subsector of “Electric gas and sanitary services”. The second cluster is observed in the left bottom part of the graph. It is composed by AIT, BEL, BLS, GTE, SBC, and USW. All these tick symbols identify telephone communication companies. The third and fourth cluster are observed in the left top part of the graph immediately above the star of stocks surrounding General Electric. Both clusters are related to transportation. One comprises ALK, AMR, DAL, LUV, and U, which are companies working in the subsector of transportation by air whereas the second one comprises CSX, NSC and UNP, which are companies of railroad transportation.

The clustering algorithm allows to obtain a hierarchical tree also. An example is provided in Fig. 4 where we reconsider the clusters observed for the group of stocks classified as “Transportation, Communications, Electric, Gas, and Sanitary Services” by the SIC code. For the sake of clarity we show in the figure only the first 250 stocks which are first present in the construction of the MST. The structure of the cluster is now more refined and precisely described. In the figure each vertical line indicates a stock. Stocks are separated by the ultrametric distance (y-axis) at which lines are observed to merge. All the stocks belonging to the “Transportation, Communications, Electric, Gas, and Sanitary Services” sector are indicated with a thicker solid line. All the other stocks are drawn with a thinner solid line. Specific clusters are observed. From left to right, the first one is labeled numerically by number 0 to 4 and comprises AMR, DAL, U, LUV and ALK. The second one is observed from 61 to 63 (NSC, CSX, and UNP). the third one is rather big extending itself from 89 to 102 (DUK, ED, AEP, BGE, D, NSP, FPL, PEG, CPL, SO, CSR, AYE, PCG, and SCG). Then the telephone communication cluster is observed from 108 to 113 (BEL, AIT, BLS, SBC, USW, and GTE). Other electric services clusters are observed considering stocks WEC, CIN (133-134), SRE, TXU and NES (from 138 to 140), PE, FPL and EIX (from 154 to 156), and finally OGE, FPC and IPL (from 165 to 167).

3. Conclusions

Correlation based networks can be obtained in financial markets by investigating financial time series. Here we have reviewed the basic method by analyzing the
Figure 4. Hierarchical tree of the first 250 stocks entering in the construction of the MST. Each vertical line indicates a stock. Stocks are separated by the ultrametric distance (y-axis) at which lines are observed to merge. All the stocks belonging to the “Transportation, Communications, Electric, Gas, and Sanitary Services” sector are indicated with a thicker solid line all the other stocks are drawn with a thinner solid line. Specific clusters are observed. From left to right, the first one is labeled numerically by number 0 to 4 and comprises AMR, DAL, U, LUV and ALK. The second one is observed from 61 to 63 (NSC, CSX, and UNP). The third one is rather big extending itself from 89 to 102 (DUE, ED, AEP, NSP, FPL, SO, CSR, AYE, PCG, and SCG). Then the telephone communication cluster is observed from 108 to 113 (BEL, AIT, BLS, SBC, USW, and GTE). Other electric services clusters are observed considering stocks WEC, CIN (133-134), SRE, TXU and NES (from 138 to 140), PE, FPL and EIX (from 154 to 156), and finally OGE, FPC and IPL (from 165 to 167).

portfolio of stocks composed by the 500 most capitalized stocks traded at the New York Stock Exchange. The correlation-based networks are obtained with a well-defined filtering procedure, which mainly focuses on the most relevant correlations among stocks. Different filtering procedures have been proposed by different authors and provide different aspects of the information stored in the investigated
sets. The robustness over time of the MST characteristics has been investigated in a series of studies\textsuperscript{13,17,25,26,27,28}. The filtering approach based on the MST can also be used to consider aspects of portfolio optimization\textsuperscript{29} and to perform a correlation based classification of relevant economic entities such as banks\textsuperscript{30} and hedge funds\textsuperscript{31}.

The topology of the correlation based networks depends on the investigated set and on the details of investigation. The observed topology ranges from a star-like one to the complex multi-cluster structure of Fig. 1.

In summary, the study of correlation based financial networks is a fruitful method able to filter out economic information from the correlation coefficient matrix of a set of financial time series. The topology of the detected network can be used to validate or falsify simple, although widespread, market models\textsuperscript{15}.

Acknowledgments

This work has been partially supported by research projects MIUR 449/97 “Dinamica di altissima frequenza nei mercati finanziari” and MIUR-FIRB RBNE01CW3M.

References

10. M. Bernaschi, L. Grilli, L. Marangio, S. Succi and D. Vergni cond-mat/0003025
24. The Standard Industrial Classification system can be found at http://www.osha.gov/oshstats/naics-manual.html