FILTERED CLUSTERING FOR EXCHANGE TRADED FUND

Gloria Polinesi, Maria Cristina Recchioni

1. Introduction

Time series are one of many instruments used to represent data present in a variety of fields, from brain activity to finance. Researchers apply clustering techniques to time series data for many reasons. Zhang et al. (2011) mention three main objectives when detecting similarities between time series: time, shape, and change. Similarity in time means that time series are grouped together when they move similarly in time; similarity in shape occurs when time series share common trends or sub patterns. Finally, similarity in change means that time series show similarity in fitted parameters referring to underlying models, which may be different.

Due to the nature of this type of data, cluster analysis of time series requires particular techniques. Mantegna (1999) and other authors use a raw data approach for stock return time series. A single observation (day, week or month) of the time series represents a characteristic of the element and stocks are grouped together when they are correlated: the Pearson correlation coefficient quantifies the degree of interdependence between pairs of financial assets.

Clustering algorithms allow leading information about structural organization aspects to be extracted from a correlation matrix of return time series whereas correlation matrices can be represented as complete graphs lacking a notion of hierarchy (De Prado, 2016). Clustering tools, spectral methods (random matrix theory), and correlation-based graphs are all algorithms used to extract information from complex systems of correlation matrices. Indeed, correlation matrices are subject to non-stationary market conditions and 'measurement noise' due to the finite length of the time series, which makes the analysis difficult without applying these filtering tools.

We contribute to the existing literature applying the work of Miceli and Susinno (2004) outlined above to obtain a cluster of Exchange Traded Fund (ETF) returns that reflects the classification per investment class provided by the Italian Stock Exchange (commonly known as the Borsa Italiana). In the following sections, we describe hierarchical algorithms and alternative methods to filter correlation matrices and we conclude with some results.

2. Hierarchical clustering algorithm

This section shows the use of a hierarchical clustering algorithm to filter correlation matrices of stock return time series to reduce the number of parameters. In fact, filtering procedure permits to extract the structure hidden in large correlation matrices keeping significant links and removing the noisy ones. The analysis considers both a static financial market and a complex system that evolves over time.

In his first work in 1999, Mantegna investigated the correlation matrix to detect the hierarchical organization of stocks in a financial market. In an ultrametric space, the minimal spanning tree (MST) between stocks reveals the topological layout of the financial market that holds important meaning from the economic point of view. A MST, the minimum structure in terms of sum of distances between nodes, groups stocks with respect to the economic sector of the underlying companies. As a consequence, the MST is the tree associated with the single linkage clustering algorithm so playing the same role as a dendrogram.

Tumminello *et al.* (2010) also confirm that elements (or nodes) share information according to the communities they belong to and that communities are organized in a nested structure. Hierarchical clustering algorithms enable this complex structure to be detected. Furthermore, Spelta and Araújo (2012) describe the minimal spanning tree as the corresponding representation of a fully-connected system (network) where sparseness replaces completeness in a suitable way.

The steps necessary to draw an MST can be summarized as follows.

We start with the correlation matrix of the time series of N stock returns, computed as the difference of the logarithm of stock prices in the time horizon T[1]

$$r_i(t) = logP_i(t) - logP_i(t-1) \tag{1}$$

The elements of the correlation matrix for each pair of stocks,

$$c_{ij} = \frac{E(r_i r_j) - E(r_i)E(r_j)}{\sigma_i \sigma_j} \tag{2}$$

are converted into distance elements:

[1] The prices and returns of stocks, the financial assets in general, can be daily, weekly, monthly, or yearly.

$$d_{ij} = \sqrt{2 - 2c_{ij}} \tag{3}$$

MSTs are based on the distance matrix computed thus. This tree graph allows the number of links connecting stocks to be reduced from (N(N-1))/2 (total number of parameters in the distance matrix) to N-1. In general, minimal spanning trees allow hierarchical organization to be detected in sectors and subsectors of stocks, but the literature shows that the result changes if the frequency of data changes. For more details about algorithms to derive an MST, see Moret and Shapiro (1991).

MSTs are associated with the dendrogram of the single linkage clustering algorithm (SLCA); however, MSTs retain some information that the single linkage dendrogram disregards (Raffinot, 2017).

This author tests some Single Linkage (SL) variants: complete linkage (CL), average linkage (AL) and Ward's method (WM), associated with different dendrograms or hierarchical trees. We recall that:

- at each step, SL combines two clusters that contain the closest (minimum distance) pair of elements

- CL works opposite to SL: At each step, it combines two clusters that hold the farthest (maximum distance) pair of elements

- AL considers the distance between two clusters as the average distance between pairs of elements belonging to those clusters

- at each stage, WM merges two clusters if they provide the smallest increase in squared error.

Other authors concentrate on MSTs as characteristic tree graphs to describe the correlation matrices. For example, Onnela *et al.* (2003b) emphasize the aspects already presented by previous authors, but also criticize the fact that the minimal spanning tree (or simply "asset tree") only represents the static average of an evolving complex system. These authors explore the dynamics of the asset tree by computing the correlation matrix for each rolling window of width T and draw the MST for each period to see how the structure of the minimal spanning tree changes over time. They demonstrate that the basic structure of MSTs is very robust with respect to time, but it shrinks during market crises due to the strong global correlation, which makes the behavior of the assets very homogeneous.

Spelta and Araújo (2012) also propose a measure called "residuality coefficient" that compares the relative strengths of the connections above and below a threshold distance in the tree in order to assess structural changes in the MST over time.

Matesanz and Ortega (2015) draw an MST for each time window to evaluate temporal changes in the time series of countries' debt-to-GDP ratio. They calculate the agglomerative coefficient (Kaufman and Rousseeuw, 2009) of each temporal tree and cophenetic correlation (Sokal and Rohlf, 1962) between hierarchical trees for different times. An agglomerative coefficient close to 1 implies a highly nested tree structure and the cophenetic correlation instead gives an idea of how similar the grouping structure is between two different hierarchical trees. For example, during the market crises beginning in 2008, the value of the agglomerative coefficient was much less than 1 and hierarchical trees for overlapping windows were not correlated.

Although the work of Matesanz and Ortega (2015) refers to different time periods and type of data considered, the results confirm that the structure of hierarchical trees tends to be flat and different from others during market crises.

A critical aspect in considering dynamic MSTs is represented by the fact that the choice of time windows (number and length) is arbitrary, as asserted by Marti *et al.*, 2017. In fact, a trade off exists between data that is too noisy and too smoothed for small and large window widths, respectively (see Onnela *et al.*, 2003a for details).

2.1 Extensions of the MST: different algorithms

Following the work of Marti *et al.* (2017), this section lists the different algorithms used to replace the minimal spanning tree and its corresponding clusters with the goal of improving upon the seminal work of Mantegna (1999). These algorithms are both hierarchical, with correlated graphs, and non-hierarchical. The latter consider a spectral method based on the study of eigenvalues of correlation matrices.

With respect to hierarchical algorithms, Tumminello *et al.* (2005) introduce a graph to filter correlation matrices that preserves the hierarchical organization of the minimal spanning tree but includes more information. This graph is known as the planar maximally filtered graph (PMFG); it represents an extension of the MST. The basic difference between the two is the number of links considered: the MST contains $N-1$ links, compared to $3(N-2)$ for the PMFG, where N is the number of nodes in the graph[2].

Therefore, PMFG holds the hierarchical skeleton of the minimal spanning tree but is enriched with loops and cliques. As explained in Tumminello *et al.* (2010), a clique of k elements is a complete subgraph that links all k elements. Due to

[2] For planar filtered graphs, the genus is equal to 0. According to the definition in Tumminello et al. (2005), the genus is a topologically invariant property of a surface defined as the largest number of non-isotopic simple closed curves that can be drawn on the surface without separating it, i.e., the number of handles in the surface.

Kuratowski's theorem, PMFGs can only have cliques of 3 or 4 elements. The number of 3-cliques and 4-cliques that can be built is $3(N-8)$ and $N-3$ respectively.

Tumminello *et al.* (2007b) introduce the average linkage minimum spanning tree (ALMST), i.e., the spanning tree associated with the average linkage clustering algorithms (ALCA). These authors show that ALMST is able to detect groups defined in terms of economic sectors and sub-sectors of stock return slightly better than the MST.

Musmeci *et al.* (2015) recently introduced the directed bubble hierarchical tree (DBHT), a novel clustering algorithm based on the topological structure of the PMFG. In contrast to other hierarchical techniques, the DBHT first identifies clusters and then sets the intra- and inter-group hierarchy.

From the non-hierarchical side, random matrix theory (RMT) is the main approach used to investigate the structure of return correlation matrices of financial assets.

Random matrix theory has a long history (Mehta, 2004). The first results in the financial sector can be found in Galluccio *et al.* (1998), Laloux *et al. (*1999) and Plerou *et al.* (1999), Plerou *et al.* (2002).

The basic idea of RMT is to compare ordered eigenvalues, $\lambda_k < \lambda_{k+1}$, of the correlation matrix of returns (Eq. 1) to eigenvalues of a random Wishart matrix $R = \frac{1}{T}AA^T$ of the same size. This is done to understand how different the matrix in question is from a random matrix. Here, A is an $N \times T$ matrix containing N time series of length T whose elements are independent, identically distributed random variables with zero mean and variance $\sigma^2 = 1$.

The random correlation matrix of this set of variables as $T \to \infty$ is the identity matrix; when T is finite, the correlation matrix is generally different from the identity matrix.

RMT proves that in the limit $N \to \infty$ and $T \to \infty$ with a fixed ratio $Q = \frac{T}{N} \geq 1$ and $\sigma^2 = 1$, the eigenvalue spectral density is given by:

$$f(\lambda) = \frac{T}{2\pi}\frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda} \tag{4}$$

where $\lambda_\pm = 1 + \frac{1}{Q} \pm 2\sqrt{\frac{1}{Q}}$ represent the minimum and the maximum eigenvalue of the Wishart matrix.

It can be shown that the eigenvalues deviating from those of a random matrix convey meaningful information stored in the correlation matrix. Indeed, information can be extracted from eigenvalues that are higher than λ_+ (deviating eigenvalues), which involves correlations between stocks that belong to the same industry or geographical area. The "bulk" of eigenvalues instead agree with RMT, revealing the random correlations.

Onnela *et al.* (2004) remark that random matrix theory offers an interesting comparative perspective with respect to hierarchical clustering techniques.

Other authors use RMT to filter correlation matrices and construct MSTs on this filtered matrix because, in order to extract the structure hidden in large correlation matrices, trees are easier to interpret than inspecting large matrices. With this procedure, Miceli and Susinno (2004) obtain a clusterization per strategies of hedge fund returns. Hedge fund strategies represent the investment styles stated by fund managers (Lhabitant and Learned, 2002). Conlon *et al.* (2007) have also confirmed this result.

3. Data and results

We consider a data set composed of 85 ETF return time series traded over the period December 2016-November 2017 (for NT total observations).

According to the classification per investment class provided by the Borsa Italiana, Table 1 shows the ETFs classified into 11 asset classes and the number of ETFs belonging to the class. Summary statistics of returns for the asset classes - mean, variance, kurtosis and skewness- are described in the same Table. The mean value is around 0 for each asset class. The standard deviation instead depends on the asset class considered: emerging equity ETFs are slightly more volatile with respect to other classes considered. The distribution of most ETF returns tends to be non-Gaussian as confirmed by high values of kurtosis and negative values of skewness.

Having filtered the correlation matrix using the RMT approach, we then reconstruct the distance matrix from the filtered correlation matrix. Figure 1 shows the MST extracted from the distance matrix, where the size of the vertex represents the node degree (i.e., the number of edges connected to each node) and the color represents the class it belongs to.

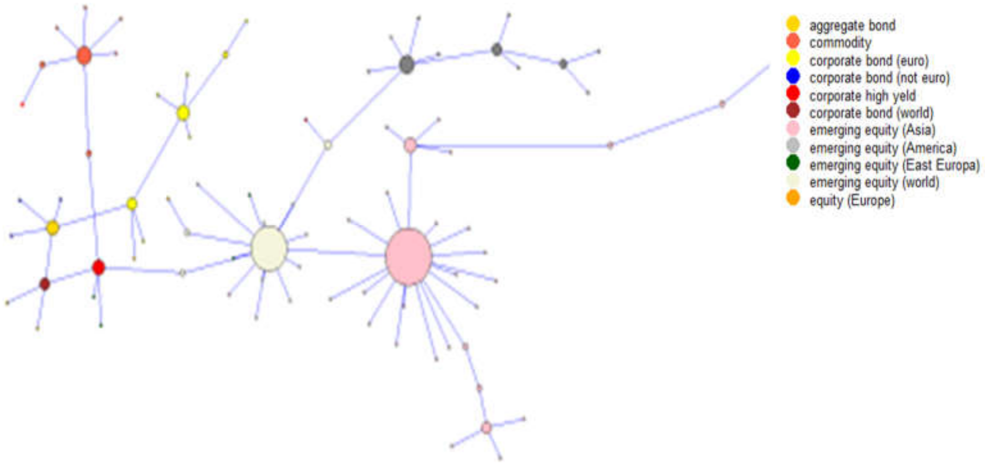
Note that the topological structure of the ETFs in Figure 1 reflects the classification per investment class described in Table 1, where the three main groups are represented by the Equity (emerging and Europe), the Corporate (aggregate, bond and high yield) and Commodity classes, respectively. Indeed, clusters obtained in the MST represent a specific class of ETFs according to the classification per investment classes of Borsa Italiana. Within these groups, specific ETFs act like hubs with higher values of node degree: the MST reveals the importance of the Asian and World Emerging Market classes, which have the highest centralities.

Table 1 – *Summary statistics of ETF returns divided into specific classes.*

ETF Class	# of ETFs per Class	Mean	St. Dev.	Kurtosis (excess)	Skewness
Aggregate Bond	4	0.0002	0.0025	5.5703	-0.4994
Commodity	8	0.0001	0.0061	0.9043	-0.0618
Corporate-euro	8	0.0001	0.0012	0.9405	-0.4330
Corporate- not euro	3	0.0003	0.0030	1.4869	0.0491
Corporate-high yield	2	0.0004	0.0014	6.6871	-0.4438
Corporate-world	1	0.0003	0.0024	1.3115	-0.2290
Emerging Equity-Asia	31	0.0010	0.0070	3.0669	-0.1522
Emerging Equity-America	10	0.0009	0.0129	33.3649	-2.9396
Emerging Equity-East Europe	4	0.0006	0.0133	4.2257	0.2659
merging Equity-world	13	0.0011	0.0060	1.7502	-0.1888
Equity-Europe	1	0.0005	0.0052	0.7576	0.1830

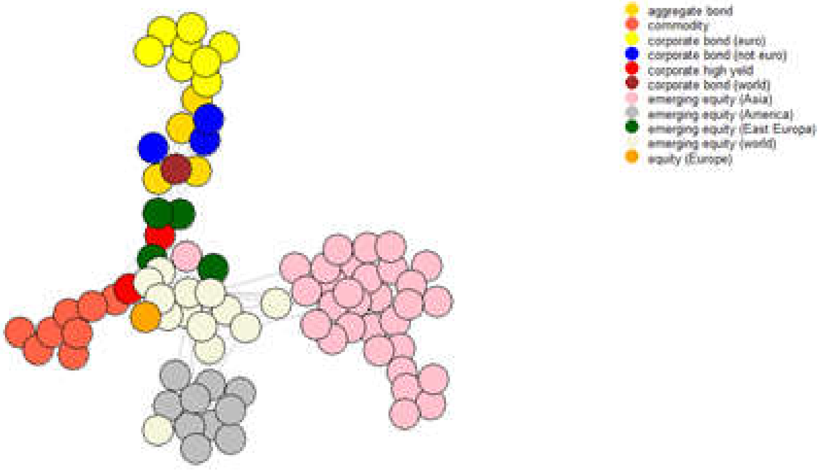
As a robustness exercise of our result, we use the same data to compare the minimum spanning tree with the planar maximally filtered graph (PMFG). Figure 2 shows that correlations of ETFs in the MST are also present in the PMFG, where a classification into investment classes is more evident from the structure of the network.

Figure 1 – *Minimum spanning tree drawn from the filtered distance matrix.*



Source: Own elaboration on ETF data.

Figure 2 – *Planar maximally filtered graph drawn from the filtered distance matrix.*



Source: Own elaboration on ETF data.

4. Conclusion

We have presented hierarchical clustering and spectral methods in order to highlight stronger correlations between time series of financial asset returns. These methods allow information in complex datasets to be filtered by building sparse networks or trees but retaining the relevant edges.

In fact, applying the random matrix approach to the correlation matrix of ETF returns and then drawing a minimal spanning tree as in the work of Miceli and Susinno (2004), allows to obtain clusters of ETFs representing the classification into investment class provided by the Italian Stock Exchange.

We have demonstrated that using RMT to filter a correlation matrix allows trees to be constructed that are easier to interpret with respect to the inspection of large matrices.

References

CONLON T., RUSKIN H. J., CRANE M. 2007. Random Matrix Theory and Fund of Funds Portfolio Optimisation, *Physica A: Statistical Mechanics and Its Applications,* Vol.382, No.2, pp.565-76.

DE PRADO M. L. 2016. Building Diversified Portfolios that Outperform Out of Sample, *Journal of Portfolio Management*, Vol. 42, No.4, pp. 59-69.

GALLUCCIO S., BOUCHAUD J. P., POTTERS M. 1998. Rational Decisions, Random Matrices and Spin Glasses, *Physica A: Statistical Mechanics and Its Applications*, Vol. 259, No. 3, pp. 449-56.

KAUFMAN L., ROUSSEEUW P. J. 2009. *Finding groups in data: an introduction to cluster analysis*. Hoboken: John Wiley & Sons.

LALOUX L., CIZEAU P., BOUCHAUD J.P., POTTERS M. 1999. Noise Dressing of Financial Correlation Matrices, *Physical Review Letters*, Vol. 83, No.7, pp. 1467.

LHABITANT F. S., LEARNED M. 2002. Hedge Fund Diversification: How Much Is Enough?, *The Journal of Alternative Investments*, Vol. 5, No. 3, pp. 23-49.

MANTEGNA R. N. 1999. Hierarchical Structure in Financial Markets, *The European Physical Journal B-Condensed Matter and Complex Systems*, Vol. 11, No.1, pp. 193-97.

MARTI G., NIELSEN F., BINKOWSKI M., DONNAT P. 2017. A Review of Two Decades of Correlations, Hierarchies, Networks and Clustering in Financial Markets., *arXiv Preprint arXiv:1703.00485*.

MATESANZ D., ORTEGA G. J. 2015. Sovereign Public Debt Crisis in Europe. A Network Analysis, *Physica A: Statistical Mechanics and Its Applications*, Vol. 436, pp. 756-66.

MEHTA M. L. 2004. *Random matrices*. Amsterdam: Elsevier.

MICELI M. A., SUSINNO G. 2004. Ultrametricity in Fund of Funds Diversification, *Physica A: Statistical Mechanics and Its Applications*, Vol. 344, No.1-2, pp. 95-99.

MUSMECI N., ASTE T., DI MATTEO T. 2015. Relation Between Financial Market Structure and the Real Economy: Comparison Between Clustering Methods, *PloS One*, Vol. 10, No.3, pp. e0116201.

MORET B. M., SHAPIRO H. D. 1991. An empirical analysis of algorithms for constructing a minimum spanning tree. In *Workshop on Algorithms and Data Structures*, Vol. 519, Berlin: Springer, p. 400-411.

ONNELA J.P., CHAKRABORTI A., KASKI K., KERTESZ J. 2003a. Dynamic Asset Trees and Black Monday, *Physica A: Statistical Mechanics and Its Applications*, Vol. 324, No.1-2, pp. 247-52.

ONNELA J.P., CHAKRABORTI A., KASKI K., KERTESZ J., KANTO A. 2003b. Dynamics of Market Correlations: Taxonomy and Portfolio Analysis, *Physical Review E, Vol.* 68, No.5, pp. 056110.

ONNELA J.P, KASKI K., KERTÉSZ J. 2004. Clustering and Information in Correlation Based Financial Networks, *The European Physical Journal B*, Vol. 38, No.2, pp. 353-62.

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*, Vol. 65, No.6, pp. 066126.

PLEROU, V., GOPIKRISHNAN P., ROSENOW B., AMARAL L.A.N., GUHR T., STANLEY H.E. 1999. Universal and Nonuniversal Properties of Cross Correlations in Financial Time Series, *Physical Review Letters,* Vol. 83, No.7, pp. 1471.

RAFFINOT T. 2017. Hierarchical Clustering-Based Asset Allocation, *The Journal of Portfolio Management*, Vol. 44, No.2, pp. 89-99.

SOKAL R. R., ROHLF F. J. 1962. The Comparison of Dendrograms by Objective Methods, *Taxon*, Vol. 11, No.2, pp. 33-40.

SPELTA A., ARAÚJO T. 2012. The Topology of Cross-Border Exposures: Beyond the Minimal Spanning Tree Approach, *Physica A: Statistical Mechanics and Its Applications,* Vol. 391, No. 22, pp. 5572-5583.

TUMMINELLO M., ASTE T., DI MATTEO T., MANTEGNA R. N. 2005. A Tool for Filtering Information in Complex Systems. In *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 102, No.30, pp. 10421-10426.

TUMMINELLO M., CORONELLO C., LILLO F., MICCICHè S., MANTEGNA R.N. 2007b. Spanning trees and bootstrap reliability estimation in correlation-based networks., *International Journal of Bifurcation and Chaos*, Vol. 17, No. 7, pp. 2319-2329.

TUMMINELLO M., LILLO F., MANTEGNA R. N. 2010. Correlation, Hierarchies, and Networks in Financial Markets, *Journal of Economic Behavior & Organization*, Vol. 75, No.1, pp. 40-5

ZHANG X., LIU J., DU Y., LV T. 2011. A Novel Clustering Method on Time Series Data, *Expert Systems with Applications*, Vol. 38, No.9, pp. 11891-11900.

SUMMARY

Filtered Clustering for Exchange Traded Fund

In this work, we show how time series of Exchange Traded Funds (i.e., ETF) returns can be clustered by reflecting the classification per investment class provided by the Borsa Italiana. We use the random matrix theory (RMT) filter to "clean" noise from a correlation matrix and we then use the reconstructed filtered correlation matrix to draw the hierarchical tree associated with the single linkage clustering algorithm (minimum spanning tree). The main goal of the paper is to show that RMT as a filter for correlation matrices enables the construction of trees that are easier to interpret with respect to large matrices, even for ETF returns.

Gloria POLINESI, Università Politecnica delle Marche, g.polinesi@univpm.it
Maria Cristina RECCHIONI, Università Politecnica delle Marche, m.c.recchioni@univpm.it