Portfolio selection based on a configuration model and hierarchical clustering for asset graphs

Imre Gera
University of Szeged Institute of Informatics
P.O. Box 652
H-6701 Szeged, Hungary
gerai@inf.u-szeged.hu

András London
University of Szeged Institute of Informatics,
Poznan University of Economics, Department of Operations Research
london@inf.u-szeged.hu

ABSTRACT

In this paper we present a null model based clustering method for asset graphs constructed of correlation matrices of financial asset time series. Firstly, we utilize a standard configuration model of the correlation matrix that provides the null model for comparison with the original one. Based on this comparison we define a distance matrix – called asset graph – on which we perform hierarchical clustering procedures. We apply this method to find clusters of similar assets in correlation based graphs obtained form various stock market data sets. We evaluate the performance of the procedure through the Markowitz portfolio selection problem by providing a simple asset allocation strategy based on the obtained cluster structure.

Categories and Subject Descriptors

I.6 [**Simulation and Modeling**]: Applications
; G.1.6 [**Optimization**]: Nonlinear programming

Keywords

Correlation matrix, Complex networks, Clustering, Portfolio selection

1. INTRODUCTION

Correlation matrices are of central importance in financial economics, especially in portfolio theory. Correlations among various assets' returns is used to determine the relative amount of capital should be invested in different assets in order to minimize the investor's risk [4]. Graphs can be easily constructed from correlation matrices in different ways. In asset graphs a node represents a company and a weighted edge between two nodes indicates, for instance, the equal-time Pearson correlation coefficient between their corresponding stock prices [3, 9, 11]. Considering correlation matrices as graphs, a wide range of tools in network analysis, like centrality measures, frequent sub-graph search or community detection, becomes available [1]. Nevertheless, the direct conversion to graphs is not evident, since the problem of information content of correlation matrices plays a key role in applications, especially in risk management. The estimation of the correlation matrix is associated with a significant level of a statistical uncertainty (sometimes called noise) due to the finite length of the asset return time series [15]. Recently, several approaches, that appeared especially in the 'Econophysics' and 'Complex Networks Analysis' literature, have been developed to handle this issue, e.g. [2, 6, 13, 14]. The idea is to filter the 'information core' of the correlation matrix that is robust against statistical uncertainty. One approach is based on random matrix theory and the idea is to compare the empirical correlation matrix with a null model matrix. The null model matrix is defined as the correlation matrix of the same number of random time series of the same length as the empirical one. A barely different approach, preferably used in the finance literature, is the principal component analysis [5]. Other filtering methods perform hierarchical clustering procedures such as single-linkage clustering [9] or average-linage clustering [13].

In this work we follow a standard null model approach for correlation matrices, but consider the information filtering problem as a graph based data mining task. We should emphasize, that in [8] the authors showed that treating the original correlation matrix as a weighted graph directly and apply modularity maximization for clustering using a standard null model approach may lead to biased results. This is due to the fact that the configuration null model doesn't necessarily give enough importance to node pairs with stronger correlations, however this is often desired in clustering algorithms. They also provided several versions of the modularity function for correlation matrices. We choose a much simpler way: we filter the original correlation matrix using a null model matrix and transform the filtered matrix to a distance matrix in a proper way. Then a hierarchical clustering procedure on the distance matrix is performed, regarded as a heuristic to maximize a modularity-like function.

The paper is organized as follows. In Section 2 we briefly describe some ways to construct asset graphs from correlation matrices, and present a heuristic for community detection (i.e. clustering) for these graphs. In Section 3 we present our experiments in various stock market data sets through the Markowitz portfolio selection problem by providing an asset allocation strategy based on the obtained cluster structure. Finally, we summarize in Section 4.

2. METHODS

Let $X_i \equiv \{x_i(t) : t = 0, 1, \ldots, T\}$ be a time series represents the value of some unit i ($i = 1, 2, \ldots n$) at time t. Particularly, in financial markets i is an asset and $x_i(t)$ is the logarithmic return of it, i.e. $x_i(t) = \log P_i(t)/P_i(t-1)$, where $P_i(t)$ is the price of asset i at time t. The system of n assets is often investigated via the correlation matrix $\mathbf{C}$ that statistically measures the pairwise dependencies, where C_{ij} is the Pearson correlation coefficient of assets i and j. It is calculated as

$$C_{ij} = \frac{\mathrm{Cov}(X_i, X_j)}{\sqrt{\mathrm{Var}(X_i) \cdot \mathrm{Var}(X_j)}},$$

where

$$\mathrm{Cov}(X_i, X_j) = \overline{X_i \cdot X_j} - \overline{X_i} \cdot \overline{X_j}$$

is the covariance of X_i and X_j, $\mathrm{Var}(X_i) = \mathrm{Cov}(X_i, X_i) = \sigma_i^2$ is the auto-covariance of X_i and $\overline{X_i}$ denotes the temporal average of the observations of X_i, i.e.

$$\overline{X_i} = \frac{1}{T}\sum_{t=0}^{T} x_i(t),$$

$$\overline{X_i X_j} = \frac{1}{T}\sum_{t=0}^{T} x_i(t) x_j(t).$$

We assume that X_i is standardized as $(X_i - \overline{X_i})/\sigma_i$.

2.1 Asset graphs

Since the correlation matrix $\mathbf{C}$ is a symmetric $n \times n$ matrix, it can be viewed as the adjacency matrix of a weighted graph. In this graph, nodes represent the assets and edges represent correlation coefficient of asset pairs. In the literature, $\mathbf{C}$ is often transformed into a distance matrix $\mathbf{D}$ with entries $D_{ij} = \sqrt{2(1 - C_{ij})}$. This is motivated by the hypothesis that ultrametric spaces[1] are meaningful in economic perspective [10].

A simple filtering technique is to threshold the values of $\mathbf{C}$ (or $\mathbf{D}$), leaving only those edges that are greater than an arbitrarily chosen value. Although the method effectively discards the weakest correlations, that are likely to caused by random fluctuations in the time series, using an inappropriate threshold value may hide important structural features of the asset graph.

A different technique, that does not require to choose a global threshold value is the minimal spanning tree approach. It reduces the number of edges of the graph from $n \cdot (n-1)/2$ to $n - 1$. The procedure is closely related to agglomerative hierarchical clustering performed with the single-linkage distance definition [9]. The approach assumes that the original correlations are approximated well by the filtered ones, and similarly to the threshold based filtering it discards all the weaker correlations. To discard less information, one can use the planar maximally filtered graph approach [14]. The method retains both the correlations used to create the minimal spanning tree and additional information as well, provided that the result is a planar graph.

[1]Ultrametric spaces are defined by an ultrametric distance that satisfy the axioms (i) $D_{ij} = 0 \Leftrightarrow i = j$, (ii) $D_{ij} = D_{ji}$ and (iii) $D_{ij} \leq \max\{D_{ik}, D_{kj}\}, \forall(i, j, k)$.

An important technique, based on fundamental results of random matrix theory, decomposes the correlation matrix $\mathbf{C}$ into a 'structured' and a 'random' part [7]. This is done by comparing eigenvalues of the empirical correlation matrix with the correlation matrix of the same number of random time series of the same length. The latter is known to be given by the Marchenko-Pastur distribution [12]. We use a similar technique in this work, but we choose a so-called configuration model to construct the null model matrix that will be compared with the original one.

2.2 Configuration model and community detection in graphs

A null model correlation matrix $\mathbf{C}^0$ is an $n \times n$ matrix, where C^0_{ij} is the mean value of the correlation between assets i and j under some null model benchmark. For example, under the assumption that every asset is uncorrelated then $\mathbf{C}^0$ would be the $n \times n$ identity matrix. Here, we use a configuration model as null model to generate C^0_{ij} by replacing edges (of the correlation graph) independently at random. The assumption is that the generated $\mathbf{C}^0$ correlation matrix preserves the *strength* of each asset i, i.e. $C_i = \sum_j C_{ij}$ is fixed as much as possible, while randomizing the 'correlation structure'.

We consider $\mathbf{C}' = |\mathbf{C} - \mathbf{C}^0|$ as the filtered (i.e. 'cleaned') correlation matrix. Then we define the re-scaled $\mathbf{D}_c = -\mathbf{C}' + |\min \mathbf{C}'| + |\max \mathbf{C}'|$ distance matrix, that may be interpreted as a weighted graph related to the correlation matrix. Here smaller distance between two nodes refers larger correlation between the corresponding assets. We then apply hierarchical clustering to $\mathbf{D}_c$. This method can be regarded as a heuristic to maximize a modularity-type function, used for clustering, given as $\sum_{i,j}[C_{ij} - C^0_{ij}]\delta_{ij}$, where $\delta_{ij} = 1$ if i and j assigned to the same cluster, and $\delta_{ij} = 0$ otherwise. Hierarchical clustering results in a *dendrogram* that can we cut at an arbitrary level h from the root to get h clusters of stocks.

3. EXPERIMENTS

Correlation (covariance) matrices often used in portfolio optimization (a widely-used model will be described). The performance of the different noise filtering procedures is generally measured via various performance metrics of composed portfolios using filtered correlation matrices.

3.1 Data sets

For our experiments we have relied on the daily closure price time series of three different stock data sets available at *Yahoo! Finance*. The selection of the stocks was based on global indices in two cases (FTSE100 and DOW30), and we also chose the 30 stocks that were active for the longest period among the available time series data. For the sake of simplicity, we refer to these data sets as "FTSE" ($n = 32$ stocks, 1183 records from 16-05-2011 to 27-01-2016), "DOW" ($n = 29$ stocks, 2849 records from 19-03-2008 to 12-07-2019) and "Active30" ($n = 30$ stocks, 5398 records from 19-01-1995 to 27-06-2016).

3.2 Markowitz portfolio selection

The Markowitz portfolio selection problem is an optimization problem where the investor would like to create an opti-

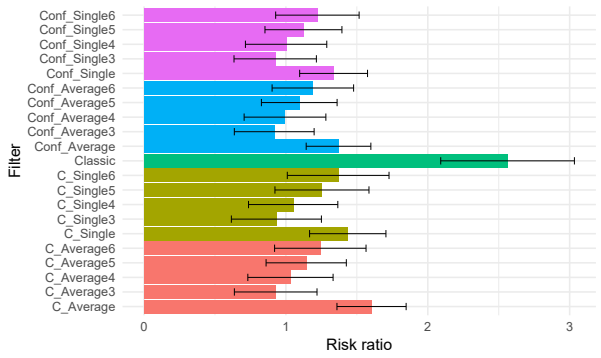

Figure 1: Risk ratios on the 'FTSE' dataset. The lower, the better.

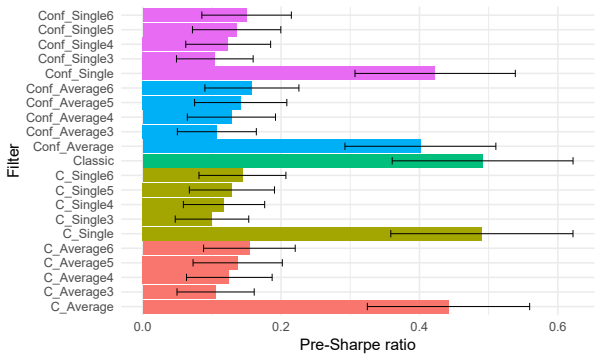

Figure 2: Sharpe ratios on the 'FTSE' dataset. The greater, the better.

mal portfolio of assets with minimum risk, given an expected return in advance. The portfolio is represented as a vector $\mathbf{p}$ that consists of the fraction of wealth to be invested in each asset. We also assume that $\sum_i p_i = 1$, i.e. 100% of wealth is invested. For example $\mathbf{p} = (0.2, 0.8)$ means investing 20% of our wealth in stock #1 and 80% in stock #2. To reach the optimum, the portfolio has to satisfy two conditions. Firstly, it has to achieve an expected return $\overline{r}_p = \sum_i p_i \overline{X}_i$, where $\overline{X}_i$ is the mean log-return of stock i, greater than a specified value R (this is an arbitrary choice). Secondly, it has to provide minimal risk, measured as $\sigma_p^2 = \mathbf{p}\mathbf{\Sigma}\mathbf{p}^T$, where $\mathbf{\Sigma}$ is the covariance (i.e. not normalized correlation) matrix of the assets considered. Negative p_i weights, also referred to as *short-selling*, are allowed.

3.3 Methodology

We used the following rolling window approach to calculate the correlation (and covariance) matrices from the time series data and perform the optimizations described previously. In each dataset we calculated the correlation matrix on the time range $[t_0, t_0 + \Delta T]$, performed a filtering procedure, in a similar way as in [13], and the optimization which gave us a portfolio $\mathbf{p}$. This meant four main optimizations per each t_0 starting day: optimization without filtering ("Classic"), filtering using hierarchical clustering on (i) asset graph $\mathbf{D}$ ("C_Single", "C_Average") and (ii) on configuration model based asset graph $\mathbf{D}_c$ ("Conf_Single", "Conf_Average"). In case of clustering procedures, our portfolio selection strategy was choosing only one asset from each cluster at random and performed portfolio optimization considering only the pre-selected assets. We then evaluated the performance of the portfolios on the interval $[t_0 + \Delta T, t_0 + 2 \cdot \Delta T]$, where $t_0 \in \{0, 30, 60, \dots\}$ and $\Delta T = 100$.

For each portfolio $\mathbf{p} = (p_1, p_2, \dots)$ we calculated the realized return as

$$\sum_{i=1}^{n} p_i \frac{P_i(t_0 + 2\Delta T) - P_i(t_0 + \Delta T)}{P_i(t_0 + \Delta T)},$$

the Pre-Sharpe ratio ($\overline{r}_p / \hat{\sigma}_p^2$), and the risk ratio ($\sigma_p^2 / \hat{\sigma}_p^2$), that is the fraction of the 'realized' and estimated risk. We calculated the mean of each metric but trimmed the data by 20% (10% on the lower and 10% on the upper end) to remove possible outlier values.

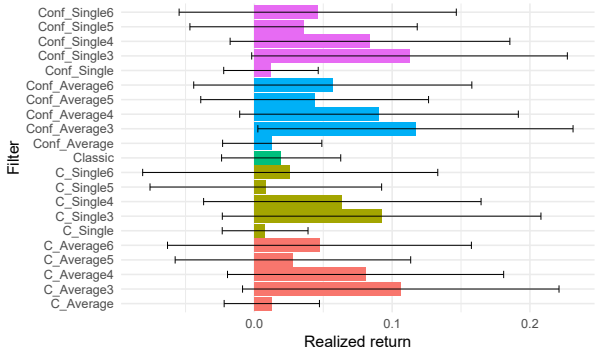

Figure 3: Realized returns on the 'FTSE' dataset. The greater, the better.

3.4 Results

Our experiments show that the resulting portfolios in general had significant improvements in all metrics when filtering methods were applied to the correlation (and hence covariance) matrix. The configuration model based approach provided lower realized risk and lower difference between estimated and realized risks than the other filtering methods (Fig. 1). The risk estimation was even better when we only used one stock per cluster (using 3 or 4 clusters provided the best risk ratios), but the estimated risk increased (the increase of the estimated risk brought it closer to the realized one). In these cases we chose a random element of the cluster, hence it was not guaranteed that we chose the assets with the lowest risk overall. Regarding Sharpe ratios (Fig. 2), it can be noted that the single-linkage clustering was the closest one to the original Markowitz-model, although when using only one stock per cluster, the value significantly decreased (due to the fact that the estimated return did not grow, but the risk increased). The configuration model performed similarly, albeit a bit worse than the other methods.

Regarding realized returns, as Fig. 3 shows, the cluster-based asset selection improved performance. When looking at the results of all the clustering-based approaches, the configuration model provided the highest realized returns with 3 clusters (and thus 3 stocks). The worst performer was the single-linkage clustering. Filtering procedures show a similar shape over time and outperform the classic method

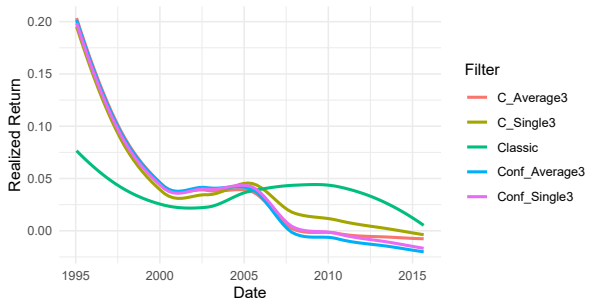

Figure 4: Realized returns of three filters on the 'Active30' dataset from 19-01-1995 to 31-08-2015.

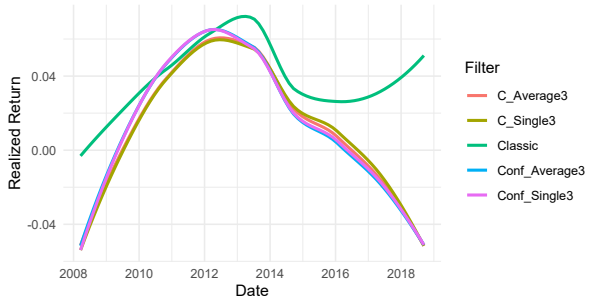

Figure 5: Realized returns of three filters on the 'DOW' dataset from 19-03-2008 to 12-09-2018.

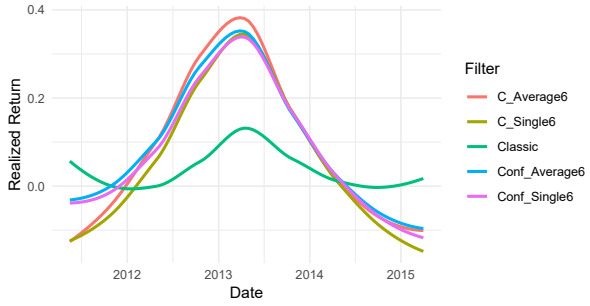

Figure 6: Realized returns of three filters on the 'FTSE' dataset from 16-05-2011 to 01-04-2015.

in certain intervals (Figs. 4-6). However, understanding the shape of the curves and the underlying causes are worth further investigation.

4. SUMMARY

In this work, by combining techniques used to investigate correlation matrices and used in graph based data mining, we performed clustering procedures for asset graphs constructed of filtered correlation matrices of financial asset time series. We provided an asset allocation strategy based on the obtained cluster structure and using Markowitz' portfolio optimization. The above discussion of our findings shows that the utilized methodology is able to provide reliable portfolios in terms of risk estimation and is competitive with classical methods in terms of return realization as well. Defining asset graphs based on different filtering procedures and cluster based asset selection strategies leave open many questions for further investigations.

Acknowledgments

The project has been supported by the European Union, co-financed by the European Social Fund (EFOP-3.6.3-VEKOP-16-2017-00002). IG was supported by the UNKP-19-2 New National Excellence Program of The Ministry of Human Capacities. AL was partially supported by the National Research, Development and Innovation Office - NKFIH, SNN-117879.

5. REFERENCES

[1] A.-L. Barabási et al. *Network science*. Cambridge University Press, 2016.

[2] J. Bun, J.-P. Bouchaud, and M. Potters. Cleaning large correlation matrices: tools from random matrix theory. *Physics Reports*, 666:1–109, 2017.

[3] K. T. Chi, J. Liu, and F. C. Lau. A network perspective of the stock market. *Journal of Empirical Finance*, 17(4):659–667, 2010.

[4] E. J. Elton, M. J. Gruber, S. J. Brown, and W. N. Goetzmann. *Modern portfolio theory and investment analysis*. John Wiley & Sons, 2009.

[5] R. F. Engle, V. K. Ng, and M. Rothschild. Asset pricing with a factor-arch covariance structure: Empirical estimates for treasury bills. *Journal of Econometrics*, 45(1-2):213–237, 1990.

[6] L. Laloux, P. Cizeau, J.-P. Bouchaud, and M. Potters. Noise dressing of financial correlation matrices. *Physical Review Letters*, 83(7):1467, 1999.

[7] L. Laloux, P. Cizeau, M. Potters, and J.-P. Bouchaud. Random matrix theory and financial correlations. *International Journal of Theoretical and Applied Finance*, 3(03):391–397, 2000.

[8] M. MacMahon and D. Garlaschelli. Community detection for correlation matrices. *Physical Review E*, 5:021006, 2013.

[9] R. N. Mantegna. Hierarchical structure in financial markets. *The European Physical Journal B*, 11(1):193–197, 1999.

[10] R. N. Mantegna and H. E. Stanley. *Introduction to econophysics: correlations and complexity in finance*. Cambridge University Press, 1999.

[11] J.-P. Onnela, K. Kaski, and J. Kertész. Clustering and information in correlation based financial networks. *The European Physical Journal B*, 38(2):353–362, 2004.

[12] A. M. Sengupta and P. P. Mitra. Distributions of singular values for some random matrices. *Physical Review E*, 60(3):3389, 1999.

[13] V. Tola, F. Lillo, M. Gallegati, and R. N. Mantegna. Cluster analysis for portfolio optimization. *Journal of Economic Dynamics and Control*, 32(1):235–258, 2008.

[14] M. Tumminello, T. Aste, T. Di Matteo, and R. N. Mantegna. A tool for filtering information in complex systems. *PNAS*, 102(30):10421–10426, 2005.

[15] M. Verleysen and D. François. The curse of dimensionality in data mining and time series prediction. In *International Work-Conference on Artificial Neural Networks*, pages 758–770. Springer, 2005.