STOCK MARKET INDEX CLUSTERS

László Nagy[66]
Mihály Ormos[67]
DOI: https://doi.org/10.31410/EMAN.2018.181

Abstract: *We introduce a spectral clustering-based method to show that stock prices contain not only firm, but also network level information. We cluster different stock indices and reconstruct the equity index graph from historical daily closing prices. We find that tail events have a minor effect on the equity index structure. Gaussian clusters can explain a substantial part of the total variance. Thus, mean-variance analysis with Gaussian clusters gives significant regression estimations. In addition, cluster-wise regressions also provide significant and stationer results.*

Key words: *cluster analysis, equity index networks, machine learning*

1 Introduction

The average surface temperature of the Earth is 15 °C. Everybody feels the temperature; however, it does not say too much about the current local conditions. Seasonal and geographical adjustments are required. Similarly, the global stock market structure has to be well understood to analyze local economic trends. Institutional economic surveys mostly provide qualitatively identified network structures e.g. emerging markets, developed markets.

The main goal of this study is to provide quantitative techniques to discover the equity index network structure.

The baseline concept follows the CAPM, in which similarity measures are calculated from correlations between logarithmic returns (Yalamova 2009). The anomalies of CAPM indicate a two dimensional mean-beta framework that gives only a simplified picture of the real market structure. The proposed non-linear similarity kernels are able to deal with higher order terms, hence clusters would be more accurate.

We show that normalized Laplacian based spectral clustering techniques can be used for recognizing well separated clusters in the global financial markets. Analyzing the correlation structure of stock indices turns out clusters are homogenously connected with each other, hence the normalized Newman-Girvan modularity matrix brings better clustering results (Bolla 2013).

2 Methodology and Data

The current study presents detailed analysis of 59 stock indices. We apply USD denominated stock splits and dividends adjusted daily closing prices between 26/9/1990 and 21/9/2015. Data is provided by Thomson Reuters. In order to underline the highly different characteristics of individual stock indices we present some monthly descriptive statistics.

[66] Department of Finance, Budapest University of Technology and Economics, Magyar tudósok körútja 2, Budapest H-1117, Hungary
[67] Department of Finance and Accounting, Eötvös Loránd University, Szép utca 2. Budapest H-1053, Hungary and Department of Economics, J. Selye University, Bratislavská cesta 3322, SK-94501 Komárno, Slovakia

Table 1
Descriptive statistics of monthly returns

Index	Mean	Variance	Skewness
.CSI300	0.018	0.056	-0.336
.XU100	0	0.026	-0.809
.DJI	0.012	0.009	-0.819
.UAX	-0.034	0.037	-0.721
.WORLD	0.004	0.002	-1.889

Notes: Table 1 shows the descriptive statistics of the monthly returns, where CSI300, XU100, DJI, UAX, WORLD represent the Shanghai Composite 300, Brose Istanbul 100, Dow Jones, Ukraine UX index and MSCI World index respectively.

Our selection criterion for covered stock indices is based on their classification in IMF Economic Outlook 2015 and the MSCI WORLD Index composition in 2015. In our analysis we allocate approximately the same weight to each region. Although, the number of countries is not equal in each region and the market capitalization differs as well, we rebalance the sample by choosing approximately ten indices from each IMF group. We are also interested in the role of well diversified indices e.g. MSCI WORLD and EURO STOXX600, hence we put them into the analysis.

2.1 Spectral clustering

In the 20th century, the appearance of large, complex data sets brought new challenges to develop methods which can be used to understand complicated structures. Spectral clustering techniques provide optimal, lower dimensional representation of multidimensional data sets. The idea is to represent the data structure as a weighted graph, and cut the graph along the different clusters. This approach leads to penalized cut optimization problems. Linear algebra and cluster analysis give powerful methods to find the optimal representations and minimized cuts.

2.2 Similarity matrix

If we would like to cluster different items, first the measurement of similarity has to be decided. In this study similarity of two stock indices (i, j) will be denoted by $W_{i,j}$. The goal is to penal differences and reward similarities. Logarithmic returns are easy to handle and keep all the information about the price processes.

$$r_i(t) = \ln\left(\frac{S_i(t)}{S_i(t-1)}\right) \quad (1)$$

where $S_i(t)$ represents the price of index i. The current study analyses multiple similarity approaches.

First, the Markowitz based squared correlation is considered as a similarity metric.

$$W_{i,j} = \mathrm{Corr}^2(r_i, r_j) \quad (2)$$

We argue this approach because logarithmic returns are not normally distributed, hence non-linear effects also could be important. However; correlation is linear, hence squared correlation similarities take into account only linear dependences.

The problem of higher-order moments can easily be solved by using symmetric and positive-definite kernel functions. The idea comes from the functional analysis. Data can be transformed into a reproducing kernel Hilbert space (RKHS) where applying the usual statistics provide the same outcomes which can be reached by using non-linear statistics in the original Hilbert space. In practice, the Gaussian-kernel is widely used (Leibon et al. 2008).

$$W_{i,j} = \exp(-\| r_i - r_j \|^2) \tag{3}$$

We notice that, if the sets of the relevant information and sensitivities are similar, then the relative entropy of the distribution of return processes is small. Otherwise, we can say stock indices are sensitive to different sets of information in a different manner (Ormos and Zibriczky 2014). This means similarity function has to be monotonically decreasing in symmetric Kullback-Leibler distance, thus we can construct a similarity measure such that:

$$W_{i,j} = \frac{1}{1+[KL(p(r_i)\|p(r_j))+KL(p(r_j)\|p(r_i))]/2} \tag{4}$$

where $p(r_i)$ denotes the probability distribution function of logarithmic returns of index i and $KL\big(p(r_i) \parallel p(r_j)\big) \stackrel{\text{def}}{=} \sum_x p(r_i = x) \ln\left(\frac{p(r_i=x)}{p(r_j=x)}\right)$ is the relative entropy of indicies i and j.

Another perspective says that large deviations are riskier, hence similarities should be defined with tail distributions. We calculate the differences of return series and count the number of at least two standard deviation peaks. The logic implies indices are similar if their price processes jump together. Similarity function has to be decreasing in the number of large deviations, hence we propose the following metric;

$$W_{i,j} = \frac{1}{1+\sum_{t=1}^{T} \delta(|\, z_i(t)-z_j(t)|>2)} \tag{5}$$

where z_i represents the normalized return of index i. In the current study we compare each approaches.

2.3 Normalized modularity

The equity index structure is strongly connected. We cannot say that events in Africa do not have any kind of effects on European markets, hence we have to find methods which can be used to cluster dense graphs. Let $G(V_{Nx1}, W_{NxN})$ be a weighted graph, where V denotes the set of vertices and W represents the weights of the edges. A k-partition of graph $G(V, W)$ can be defined as the partition of vertices such that $\bigcup_{a=1}^{k} V_a = V$ and $V_i \bigcap V_j = \delta_{i,j} V_i \;\; \forall i, j \in \{1, \dots, k\}$. The $W_{i,j}$ value represents the strength of the connection between nodes (i, j). If we assume that nodes are independently connected, then the guess of weight $W_{i,j}$ will be the product of the average connection strength of i and j. The average connection strength d_i and d_j are given by $W, d_i = \frac{1}{N}\sum_{u=1}^{N} W_{i,u}$. Thus, $W_{i,j} - d_i d_j$ captures the information of the network structure (Bolla 2011), hence if we would like to maximize the sum of information in each cluster, then we get:

$$\max_{P_k \in \mathcal{P}_k} \sum_{a=1}^{k} \sum_{i,j \in V_a} (W_{i,j} - d_i d_j) \tag{6}$$

where P_k stands for specific k-partition in $\mathcal{P}_k$.which represents the set of all possible k-partitions.

Let $M \coloneqq W - dd^T$ denotes the modularity matrix of $G(V, W)$. If we would like to get clusters with similar volumes then we have to add some penalty to Equation (6) hence we get the normalized Newman-Girvan cut.

$$\max_{P_k \in \mathcal{P}_k} \sum_{a=1}^{k} \frac{1}{Vol(V_a)} \sum_{i,j \in V_a} (W_{i,j} - d_i d_j) \tag{7}$$

where $Vol(V_a) = \sum_{u \in V_a} d_u$.

Let us define the so called normalized modularity matrix;

$$M_D \coloneqq D^{-1/2} M D^{-1/2} \tag{8}$$

If we would like to cluster a weighted graph $G(V, W)$ then eigenvectors of its modularity (M) and normalized modularity matrices (M_D) can be used. Modularity and normalized modularity matrices are symmetric, and 0 is always in the spectrum of M_D.

$$M_D = \sum_{i=1}^{N} \lambda_i u_i = \sum_{i=1}^{N-1} \lambda_i u_i$$

where $1 > \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \geq -1$ denote the eigenvalues of M_D.

If we would like to maximize Equation (7) we can use the k-means clustering algorithm on the optimal k-dimensional representation of vertices,
$\left(D^{-\frac{1}{2}} u_1, \ldots, D^{-\frac{1}{2}} u_k\right)^T$.

where $u_1, \ldots, u_k$ denote the corresponding eigenvalues of $|\lambda_1(M_D)| \geq \cdots \geq |\lambda_k(M_D)|$. Moreover, if the normalized modularity matrix has large positive eigenvalues, then the graph has well separated clusters, otherwise clusters are strongly connected.

Another natural approach is to minimize the normalized cut (Luxburg 2007)

$$\min_{P_k \in \mathcal{P}_k} \sum_{a=1}^{k-1} \sum_{b=a+1}^{k} \left(\frac{1}{Vol(V_a)} + \frac{1}{Vol(V_b)}\right) W_{i,j} \tag{9}$$

The optimization problem is similar to Equation (7). Instead of the normalized-modularity matrix the normalized Laplace matrix gives the solution (Shi and Malik 2000).

$$L_D \coloneqq D^{-\frac{1}{2}} (D - W) D^{-\frac{1}{2}} \tag{10}$$

This technique works when clusters are well separated otherwise normalized modularity gives better figures.

2.4 Algorithm

In empirical analysis, the following steps are the backbone of the calculation (Filippone et al. 2007).

1. Constructing similarity matrix (W),
2. Calculating normalized modularity matrix (M_D),
3. Based on the spectral gap, determine the number of clusters and optimal k-dimensional representation,

4. Apply k-means clustering.

2.5 Assessment of clustering methods

Relevance of different clustering techniques can be tested in multiple ways. The most common metrics follows a regression based logic. In this framework we suppose that variance has two components, the within and the between cluster components. Therefore, the explanatory power of given clusters can be described as

$$\frac{\sum_{j=1}^{k}\sum_{j=1}^{N_i}(X_{i,j}-\overline{X})^2 - \sum_{i=1}^{k}\sum_{j=1}^{N_i}(X_{i,j}-\overline{X}_i)^2}{\sum_{i,j=1}^{N_i,N_j}(X_{i,j}-\overline{X})^2} \tag{11}$$

where k represents the number of clusters, N_i shows the size of clusters and $\overline{X}$, $\overline{X}_i$ stands for the total and cluster wise average (Zhao 2015). The formula penalizes dispersions within clusters, hence dense clusters would give number close to 1. Moreover, calculating the ratios with different number of clusters highlights the optimal number of clusters as well.

3 Empirical results

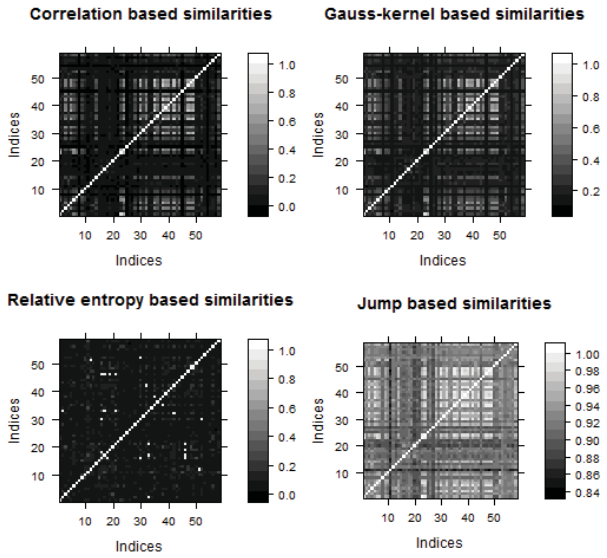
Current study presents a broad analysis of the equity index network structure. Logarithmic returns of 59 stock indices are clustered in different ways. The investigation reveals stock indices are homogenously connected and large price movements have limited effect on the network structure.

3.1 Similarity metrics

Defining similarity is a key aspect in clustering. In general it is hardly possible to find an optimal kernel, but different approaches can be tested and compared on specific data sets. This study analysis correlation, jump, entropy and Gaussian based similarity kernels. Calculating the similarity matrices we expect strongly connected indices have coefficients close to one, whereas loosely connected close to zero.

Figure 1

Level pots of daily similarity matrices



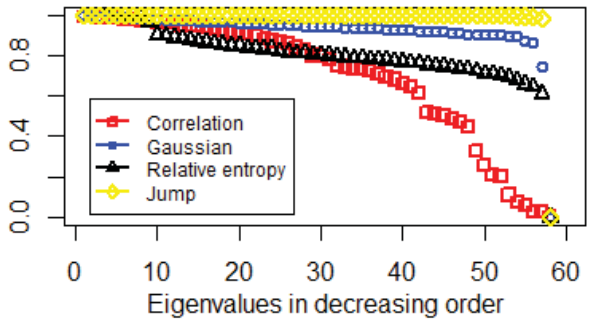
Different similarity measures imply similar patterns which are in line with our a priory intuition. However, the spectrum of normalized Laplace and normalized modularity matrices help us to find the most adequate kernel function, because the wider the spectral gap the better the clustering property. This means, we have to find similarity metrics which implies large gaps in the spectrum of normalized Laplacian and modularity matrix.

Empirical evidences (Figure 2.) show relative entropy and Gaussian-kernel also can be used to cluster the stock index network, while correlation and jump based similarities are not promising.

Correlation based similarity approach implies roughly uniform eigenvalue density on [0,1]. This means, a lot of gaps appear in the spectrum, hence we could not say anything about the optimal number of clusters. Moreover, lower dimensional represenatiations will not contain all the information, because of some of the large eigenvalues are not considered. These hurdles highlight the problems of squared correlation similarity matrices.

Counting at least two standard deviation jumps results small number of eigenvalues with large multiplicity. Therefore, lower dimension representation can not be used to cluster the data points. Accordingly, jumps are random that do not say much about the network structure.

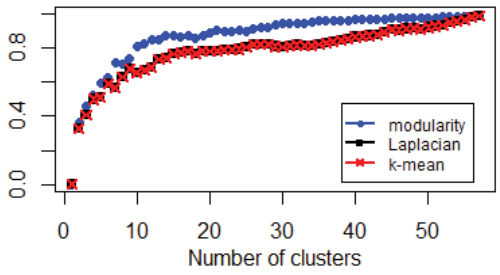
Figure 2
Eigenvalues of normalized Laplacian matrix in decreasing order



Notice that, these results are in line with Figure 1. Because, normalized Laplacian minimize the normalized cut (Equation (9)), which is small if and only if clusters are loosely connected. Whereas, modularity approach maximize the information of clustering, hence it can be used in homogeneous network structure as well. Investigating the spectrums, especially the positions of spectral gaps, gives some guidance on the optimal number of clusters. Considering the previous results the spectrums of Gaussian and relative entropy based normalized modularity matrices are suitable. Figure 2. shows indices could be put into 2, 3 or 5 clusters. We apply the elbow method to identify the optimal number of clusters. This approach is rather computation intensive, because of the percentage of variance explained as a function of clusters has to be estimated (Eq. 11); thus, the whole process has to be repeated many times. However; in our case we have 59 stock indices, hence the elbow method can be used as well. Figure 5 and 6. give evidences for using 2, 3, 4 or 5 clusters.

László Nagy *is a PhD student at the Department of Finance, at the School of Economic and Social Sciences, Budapest University of Technology and Economics. His main area of research is financial risk measures and asset pricing.*

Figure 4
Explained percentage variance of Gaussian-kernel based clusters of representations



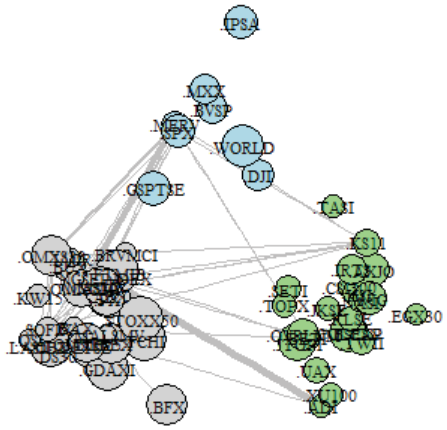
The Figure 2 and 3 show Gaussian-kernel implies the clearest spectrum property. Relative entropy based kernel also gives usable results. Whereas, jump and correlation based approaches are ineffective.

3.2 Equity index network structure

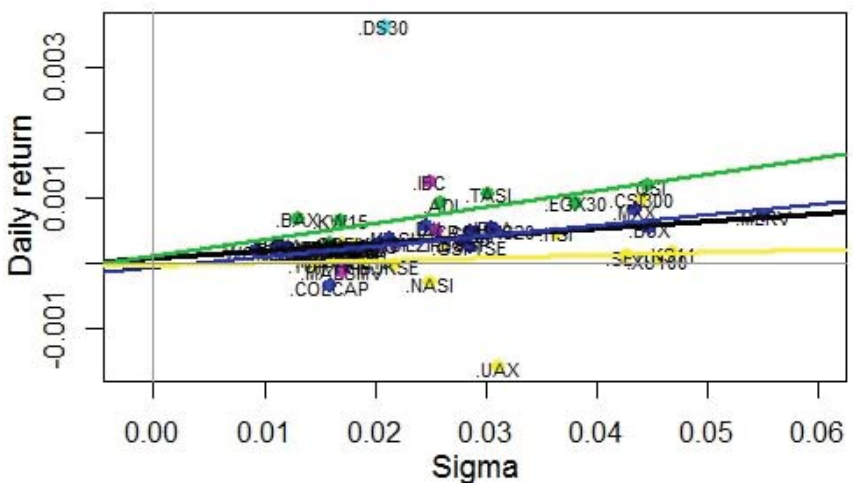
Spectral gap (Figure 2.) and variance analyses (Figure 4.) imply equity indices can be studied by using 2, 3 and 5 clusters. The explanatory power of two clusters is 38%. This means roughly one-third of the total variance comes from the sample heterogeneity. If we increase the number of clusters and investigate the three cluster case we get similar explanatory power. However, spectral gap appears between the third and fourth eigenvalues (Figure 2.), so theoretically three cluster is proposed. The next gap is between the fifth and sixth eigenvalues. Explanation power of five clusters is 52%. This means, half of the total variance of data can be explained by five clusters. Additional clusters have little explanatory power which is in line with spectrum properties. In practice, mean-variance plots can be used to represent risks and rewards. Intuitively, indices with similar risk and return can thought to be similar. This approach applies k-means algorithm to cluster the two dimensional (mean, standard deviation) representation of logarithmic returns. We have seen this naïve method does not give optimal cuts. However; if we calculate Gaussian similarities and normalized modularity matrix based representation, then we get clusters with higher variance explanatory power. We have seen stock indices can be put into 2, 3 or 5 clusters. In Figure 6. we can see clusters which are optimizing the modularity cut are concave in mean-variance framework. If we have a closer look at the indices in Appendix Table A1 we could say qualitative approach also works, because east-west geographical clustering would imply almost similar results. Putting the indices into three different clusters gives a complicated figure, but we could still say first cluster is dominated by European countries, second by American and third is a mixture of indices from the rest of the world.

Mihály Ormos *is a Professor of Finance at Department of Finance and Accounting, Institute of Business, Eötvös Loránd University and at Department of Economics at J. Selye Univerity. His area of research is financial economics especially asset pricing, risk measures, risk perception and behavioral finance. He serves as one of the contributing editors at Eastern European Economics published by Taylor and Francis. His teaching activities concentrate on financial economics, investments and accounting.*

Figures 5
Three Gaussian-kernel based normalized modularity clusters



Figures 6
Five Gaussian-kernel based normalized modularity clusters



Calculating five different clusters help us to have a deeper understanding on the network. The first surprising result is despite of the penalty of different cluster sizes Dhaka stock exchange (.DS30) is separated into cluster three. In addition, cluster four contains only two African and two American indices. Another interesting result is the first cluster. Arabian indices except Morocco are put into this cluster. Cluster two contains developed while cluster five contains emerging markets.

Table 3
Descriptive statistics of daily linear regressions

	p-value of intercept	p-value of s.d.	R^2
Total Sample	0.62	0.12	0.05
First cluster	0.62	0.02	0.68
Second cluster	0.29	0.00	0.59
Fifth cluster	0.93	0.71	0.01

Notes: This table shows the p statistics and R^2 values of daily linear regressions. Returns are regressed on standard deviations. Calculation is done for total, only the first, second and fifth clusters.