Cluster Analysis of Market States

Michael Tan, Ph.D., CFA

8 May 2008

Presented during the
Mathematical Adventures in Data Mining
course given by
Prof. Esteban Tabak
New York University

www.michaeltanphd.com
Apothem Capital Management, LLC
330 East 38th Street 14L
New York, NY 10016
Tel: 212-922-1265
mltan@apothemcapital.com

How to use cluster analysis for prediction?

- Usually the goals of cluster analysis are twofold:

  1. Organize a set of objects into subsets or “clusters” whose members are similar in some way.

  2. Arrange clusters into a natural hierarchy such that clusters within the same “supercluster” at each level are more similar than those in different superclusters.

- Corresponding to these goals are the two major styles of clustering:

  1. Partitioning (“k-clustering”)
  2. Hierarchical Clustering (“tree clustering”)

- Partitioning divides a set of objects into k clusters, where k is known or given a priori, while hierarchical clustering automatically discovers the “natural” classes in data where no “obvious” partitions can at first be seen.

- With simple add-ons, this paradigm for classification of data can be used for prediction as well.

- For example, given an unknown object, one could “predict” the cluster to which it belongs by determining the k nearest neighboring objects whose cluster memberships are known. The cluster to which the majority of these known objects belong is then the predicted cluster for the unknown object.

Using cluster analysis to predict the stock market

- This paper describes a data mining exercise based on cluster analysis to predict the stock market.

- The basic idea is as follows:

    1. Let the "state" of the market at a given time be specified by a cross-sectional set of stock attributes (such as returns, earnings, book value, etc).

    2. Treat the current "state" of the market as an object and associate it to other objects which are past states of the market.

    3. Compute the cluster structure for all past states of the market.

    4. Determine the cluster of past states to which the current state most likely belong using, for example, a k nearest neighbor followed by majority vote procedure.

    5. Infer the future state of the market from the states that immediately follow (in time) the states in the cluster assigned to the current state.

- The rest of the paper is organized into two parts, viz. a brief review of cluster analysis followed by specification and results of the data mining exercise.

Brief review of cluster analysis

- In cluster analysis, the causal relationship between two objects is quantified by a *distance function* giving the degree of association or similarity between the objects.

- For a function $d(x_i , x_j)$ of x_i and x_j drawn from an input set S to qualify as a distance function, it must satisfy the “Euclidean metric”:

$$d(x_i, x_i) = 0 \qquad \text{(i)}$$

$$d(x_i, x_j) = d(x_j, x_i) \qquad \text{(ii)}$$

$$d(x_i, x_j) \le d(x_i, x_k) + d(x_k, x_j) \qquad \text{(iii)}$$

Distance functions

- I have tested many distance functions in the context of analyzing security return relationships but will focus only on the following three in this paper:

- Let x_{it} and x_{jt} denote the returns of stock i and stock j at time t respectively. The return of stock is defined as $\log(P_t / P_{t-1})$ where P_t and P_{t-1} are the prices of the stock at time t and t-1 respectively. Let σ_i and σ_j denote the standard deviation of these returns, and let the vector notation $\mathbf{x}_i$ denote the time series of returns of stock i.

  - *Euclidean metric:* $d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^2}$

    The distance tends to be small for stocks with similar return volatility or beta and therefore groups stocks with similar betas together.

  - *Standardized Euclidean metric:* $d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\left[\frac{(\mathbf{x}_i - \overline{\mathbf{x}_i}}{\sigma_i} - \frac{\mathbf{x}_j - \overline{\mathbf{x}_j}}{\sigma_j}\right]^2} = \sqrt{2(1-\rho_{ij})}$

    This metric can be expressed as a function of the correlation ρ_{ij} between stock i and stock j which vanishes when they are perfectly correlated.

  - City block metric: $d(\mathbf{x}_i, \mathbf{x}_j) = \sum_t \left| x_{it} - x_{jt} \right|$

    Dampens large return differences and thus gives more weight to small differences as compared to the Euclidean metric.

Agglomerative clustering

- One way to perform hierarchical clustering is by *agglomeration*:

  Step 1: Start with a set S of n objects.

  Step 2: Place elements of S into singleton sets S_1 , S_2 … S_n .

  Step 3: Devise a *cost function* which determines the pair of sets $\{S_i , S_j\}$ that is “cheapest” to merge.

  Step 4: Remove S_i , S_j from the list of sets and replace with $S_i \cup S_j$.

  Step 5: Repeat steps 3 and 4 until only one set remains.

- Agglomerative clustering then differ only in the definition of the cost function or *linkage algorithm*.
- A linkage algorithm links objects together into clusters based on the “cost” of each link which in turn depends on the distances between objects.

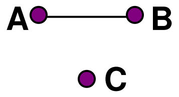
Linkages

- Some common linkage algorithms and their cost functions are given below:

Algorithm	Cost Function
Single Linkage	$\min_{x_i \in S_i, x_j \in S_j} d(x_i, x_j)$
Complete Linkage	$\max_{x_i \in S_i, x_j \in S_j} d(x_i, x_j)$
Average Linkage	$\frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} d(x_i, x_j)$
Ward Linkage	$c(S_i) = \sum_{p=1}^{n(S_i)} \sum_{q=1}^{n(S_i)} \left[d(x_p^{(i)}, x_q^{(i)})^2 \right]$ where $n(S_i) =$ number of elements in S_i

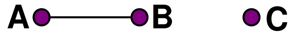
Spherical versus elongated clusters

- Complete, Ward, and average linkage tend to produce “spherical” clusters, i.e. those whose members are classified together because they are close to each other or to a “bellwether” member:



C must be close to both A and B to make {A, B, C} a cluster

- Single linkage tend to produce “elongated” clusters:



C need only be close to B (but not necessarily to A) to make {A, B, C} a cluster

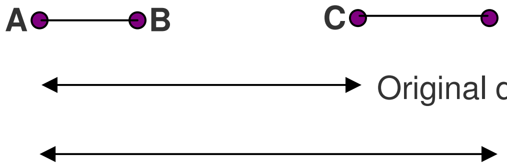
- Complete linkage is “furthest neighbor” linkage while single linkage is “nearest neighbor” linkage.

- To the extent that there are bellwether stocks and lead-lag relationships among stocks, single linkage is probably not suited for the analysis of stock clusters.

Cophenetic coefficient

- The cophenetic coefficient is the correlation between distances in the original set of objects and their corresponding “ultrametric” distances.

- For example, suppose 4 objects A, B, C and D are agglomerated into 2 clusters {A, B} and {C, D} via complete linkage:



Original distance between A and C

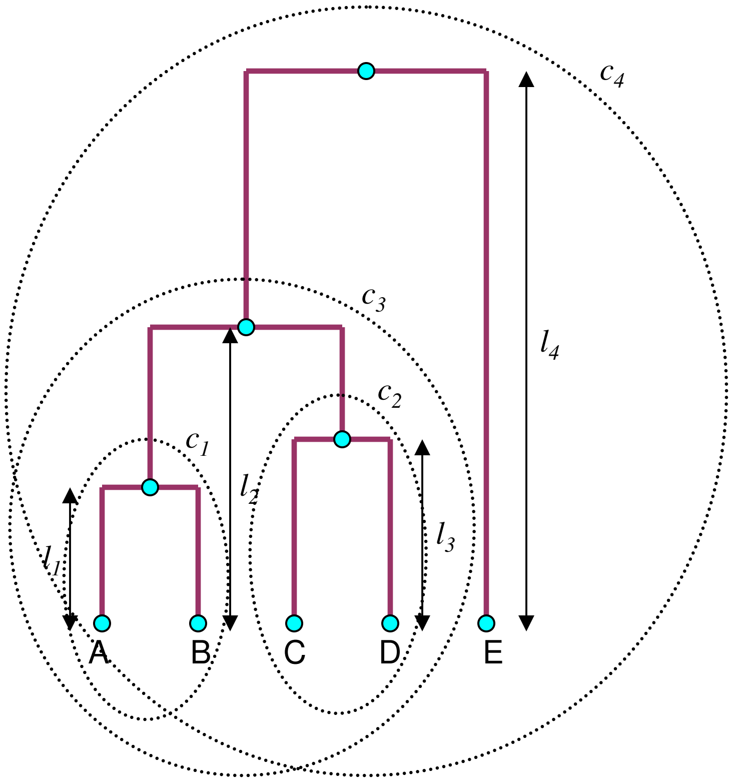
Ultrametric distance between A and C (since the distance between A and D, the two furthest neighbors, is taken as the distance between the two clusters).

- The original object-to-object distance AC is to be correlated with the distance between the clusters to which A and C belong.

- The cophenetic coefficient measures the distortion of the distance information in the original data set caused by linkage into clusters.

Dendrograms

- A *dendrogram* or *hierarchical cluster tree* is a graphical depiction of the distances between clusters.
- The tree ends in leaves at the bottom of the graph corresponding to singleton clusters, i.e. single stocks.
- Branches are drawn such that height of node at which one branch meets another is equal to the distance between clusters at the end of the branches.



● represents a node

The objects A, B, C, D, E are *leaf nodes*.

Number of objects is 5.

Number of clusters c_1, c_2, c_3, c_4 is always one less than the number of objects.

The height of the links l_1, l_2, l_3, l_4 correspond to distances.

Inconsistent coefficient

- An *inconsistent coefficient* can be computed for each link in a cluster tree.

- It compares the length of the link to the average length of links below it to a specified depth (number of levels in the hierarchy).

- For example, the inconsistent coefficient to depth one for link l_2 in the tree on the previous page is

$$\frac{l_2 - (\text{mean of } l_1, l_2, \text{and } l_3)}{(\text{standard deviation of } l_1, l_2, \text{and } l_3)}$$

- For depth m, the mean and standard deviation in the above formula are taken over the lengths of all links stemming from l_2 down m levels in the hierarchy.

- By definition, links connecting leaf nodes have inconsistent coefficients of zero.

- A large inconsistent coefficient implies that the link connects two very distinct clusters.

- A small inconsistent coefficient implies that the link connects two clusters that are not differentiated.

Generating clusters

- Clusters can be generated by grouping together the leaves at the bottom of a link whose inconsistent coefficient is larger than a specified *cut-off value*.
- The node from which the link emanates may in principle be several levels above the leaves if the links below it have inconsistent coefficients smaller than the cut-off.
- In this case, a large cluster may be generated.
- Thus, a large cut-off value (say larger than 1) will usually produce large clusters while a small cut-off value will produce small clusters.
- Inconsistent coefficients typically have a numerical range between 0 and 3.
- Different clustering configurations can be obtained by adjusting the cut-off value.

Previous work on cluster analysis of financial data

- Only one substantive thread in the published literature, viz. the work of R.N. Mantegna et al[1,2,3] from 1999 through 2007, can be found on cluster analysis of financial data.

- In 1999, Mantegna “discovered” a hierarchical cluster structure in stock returns[2].

- Later, when average linkage under a standardized Euclidean metric is used for clustering the returns, Mantegna and co-workers found that the clusters correspond to economic sectors.

- They also used cluster analysis as a noise reduction technique to improve the measurement of the return correlation matrix[3].

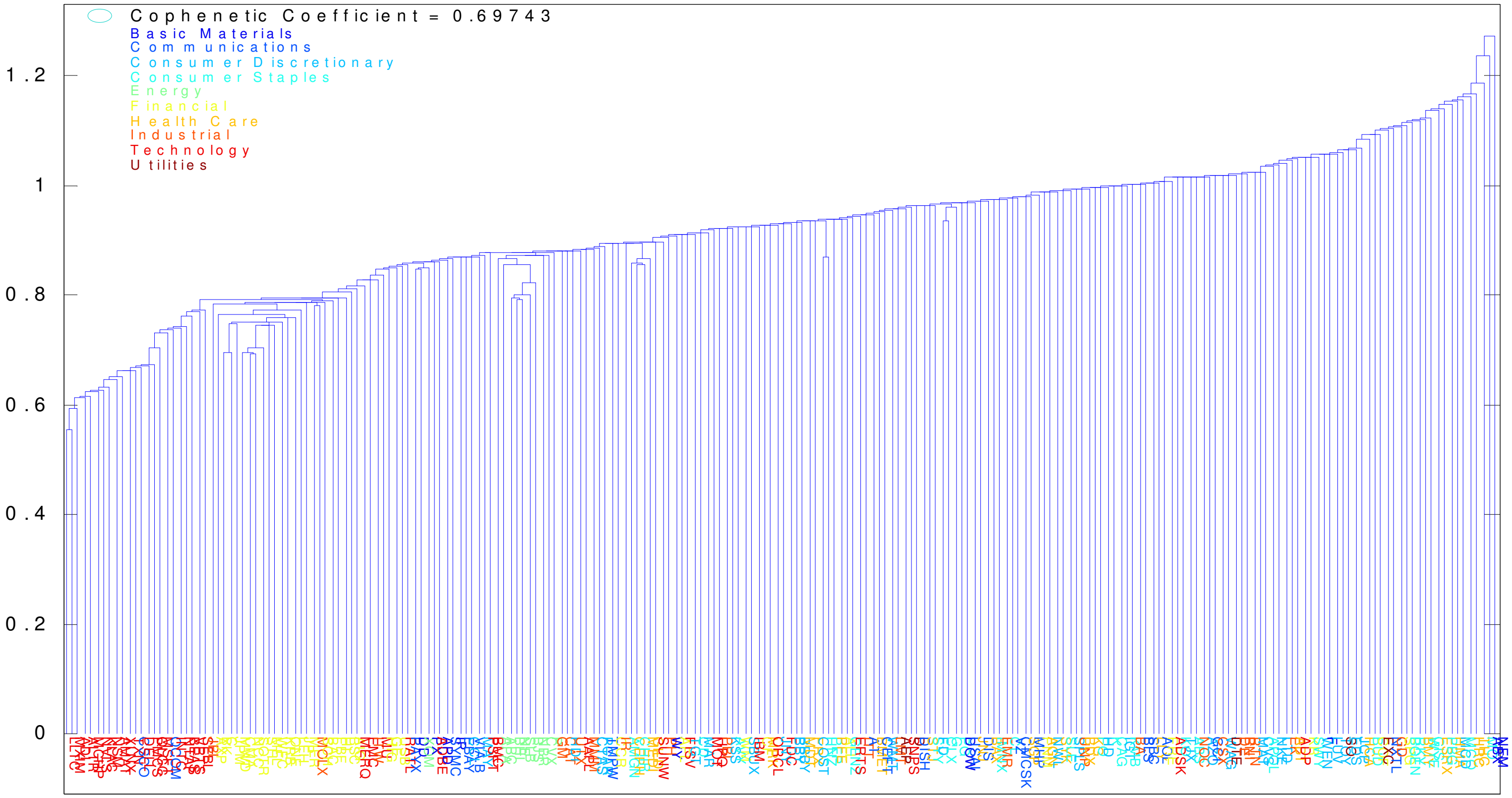
- In this technique, the out-of-sample risk of the Markowitz mean-variance optimal portfolio computed using the improved correlation matrix is closer to its in-sample value as compared to other methods of shrinking the correlation matrix.

Is there a natural hierarchy among stocks?

- Mantegna et al[2] claim that a hierarchical structure exists among stocks which is exemplified by the dendrogram below.

- The dendrogram is produced using single linkage of the standardized Euclidean metric for the returns of the top 200 stocks ranked by market capitalization from the S&P500 Index. The date range for the stock returns is 31 Dec 1998 to 31 Dec 2004.

- The nodes at the bottom of the graph are labeled with their stock symbols and color-coded with their economic sector designations.



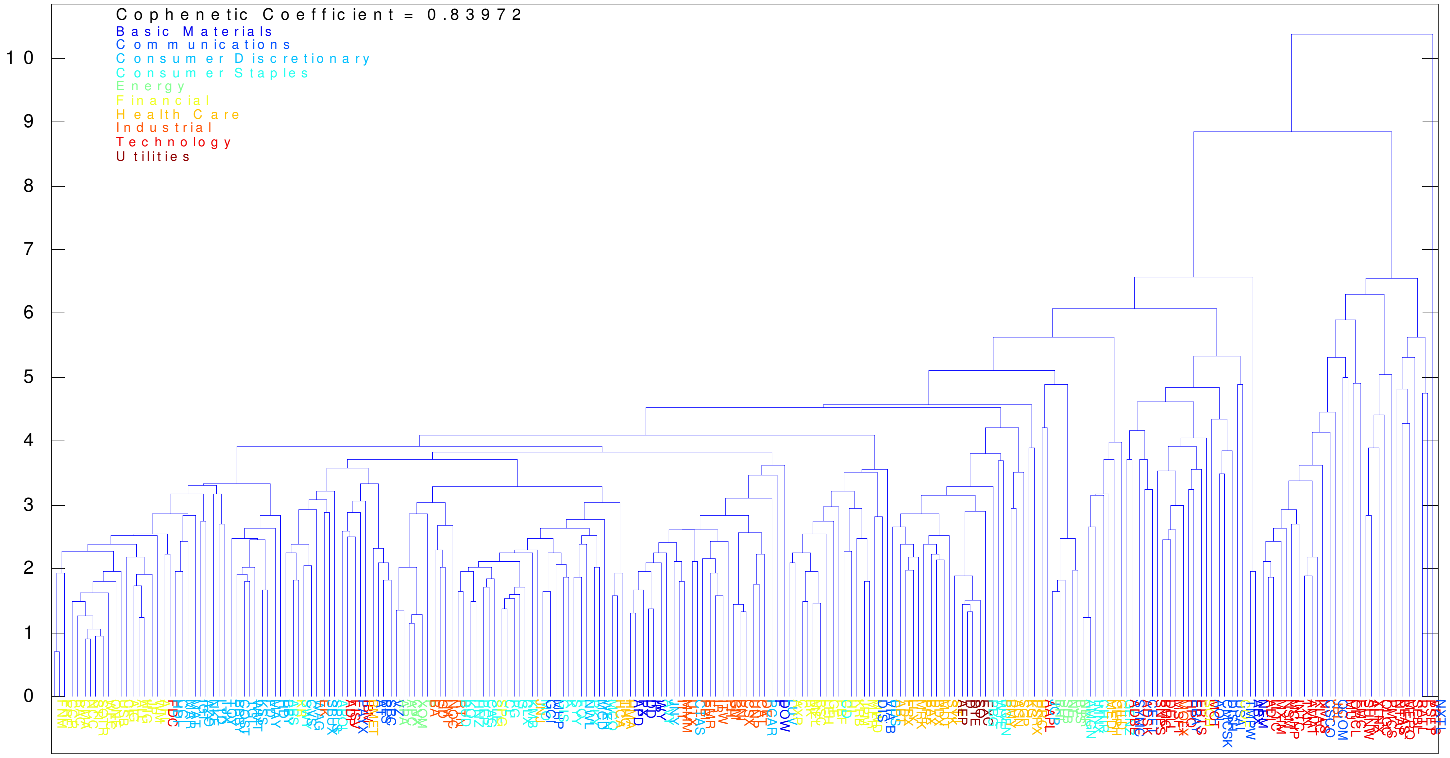
Is there a natural hierarchy among stocks?

- I do not think that there is any exalted economical meaning to the dendrogram shown on the previous page.

- Because single linkage creates “elongated” clusters, only naturally large clusters, such as those containing technology and financial stocks which form the largest two sectors in the S&P 500 Index, would show up in the dendrogram (on the left side). The rest of the clusters merely reflects the accidental ordering (due to noise in the correlation matrix) of stocks based on the magnitude of their correlations to each other.

Complete linkage reproduces economic sector classifications

- The dendrogram below shows the cluster structure of the same 200 stocks produced using complete linkage of the city block metric for returns. The hierarchical structure seen in the single linkage dendrogram disappears when we use complete linkage.
- Clusters now correspond to economic sectors, as can be seen in the bunching together of similarly colored symbols at the bottom of the dendrogram. This is the same observation reported by Mantegna et al[3].
- This is not surprising since the clustering procedure depends solely on the correlation matrix of the returns (distances being inversely related to correlations), and returns of stocks within the same sector are highly correlated to each other.



Can we do better than just classifying stocks?

- Being able to classify stocks into their economic sectors using only price data is a good thing, but can we do better by using cluster analysis to predict returns?

- I found no previously published work that attempts to do this, so I propose the following study:

  Consider a stock market consisting of N stocks. Consider the N x W dimensional space of data points where each data point is the vector

  $$\mathbf{X}_t = \left(x_{1,t}, x_{1,t-1}, x_{1,t-2} \cdots x_{1,t-W+1}, x_{2,t}, x_{2,t-1} \cdots x_{2,t-W+1} \cdots x_{N,t}, x_{N,t-1}, \cdots x_{N,t-W+1}\right)$$

  Here $\mathbf{x}_{i,t}$ is the return of stock i over the time period from $t - 1$ to t, and W is the width of "lookback" window. If the time discretization is one day and we have M days of historical data, then t would vary from W to M. Thus each data vector represents the "state" of the stock market on day t characterized by short histories of the returns of all stocks over the past W days all strung up together.

- Note that we are clustering all the states of the stock market observed at different times in the past.

Predictability hinges on the similarity of the successor states

- The question I will study in the rest of this paper is whether the states immediately following (in time) the states in a given cluster also close together in the given metric. Predictability would hinge on the answer to this question.

- If these “successor states" are also close together in some sense, then we can say "history tends to repeat itself".

- Thus given an out-of-sample state, we simply check to which cluster it belongs and look at the successor states corresponding to the predecessor states in the cluster. By studying these successor states, we may be able to predict what will happen in the future of the out-of-sample state.

- How do we tell if a set of successor states are “close”?

- By “close”, we mean “closer than it should be” given a joint distribution of the returns that imply unpredictability.

The data set for the study

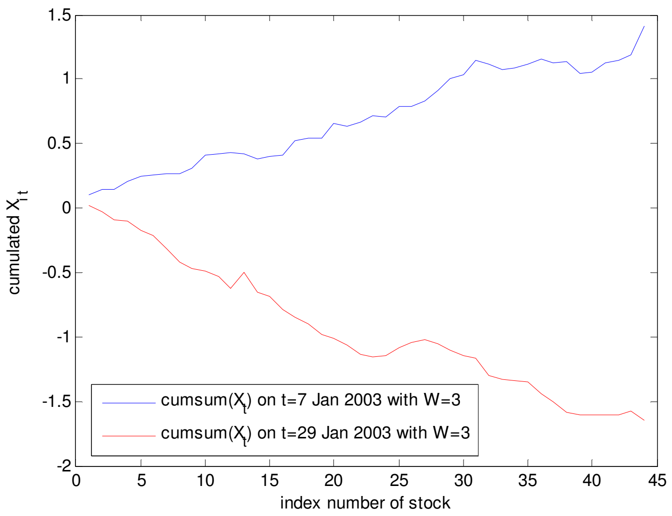
- We use the following data set for the study:

  1. Take the top 44 stocks sorted by market capitalization from the CRSP[5] database on 31 Dec 2002.

  2. Construct the market state vector $\mathbf{X}_t$ as defined on the previous page from the daily total returns of these stocks between this date and 30 Apr 2008.

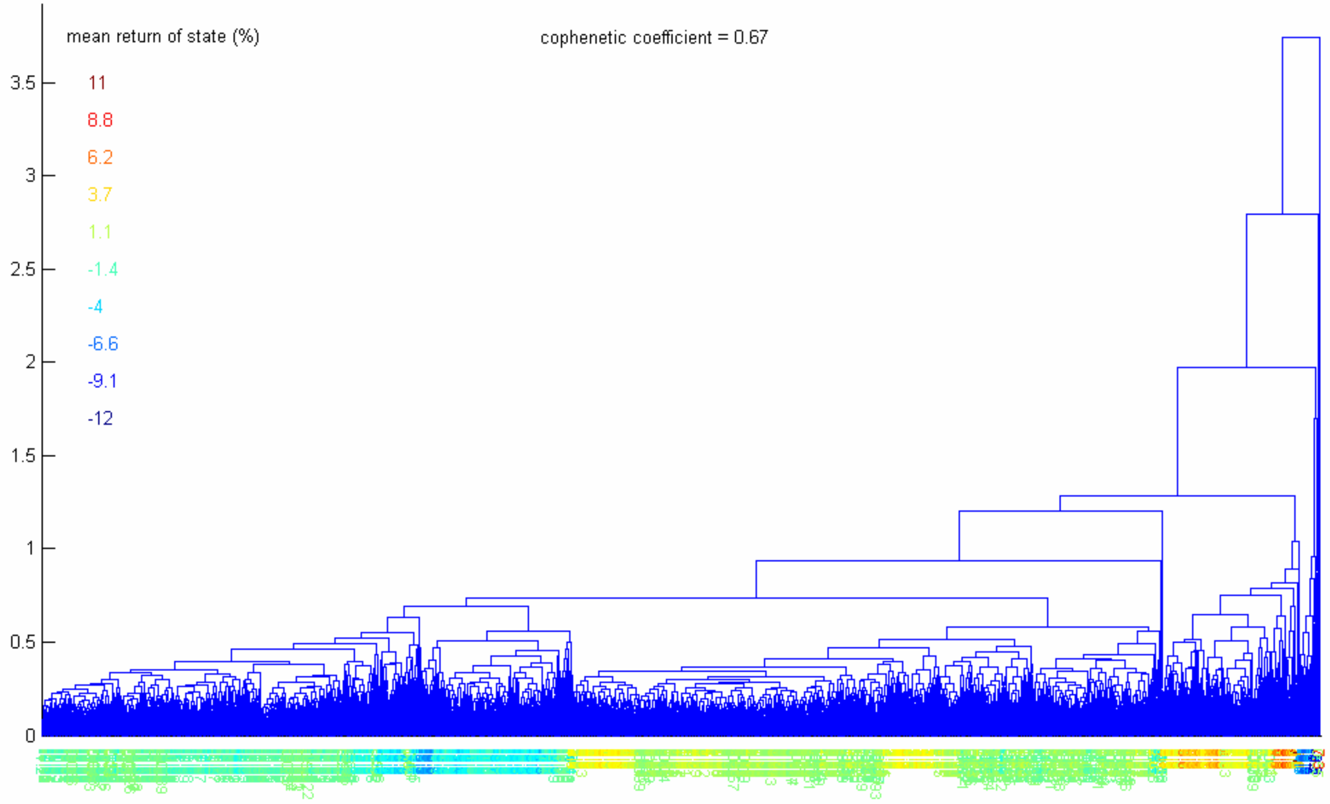
  3. For stocks that drop out due to takeovers or other restructuring events, fill their slots with the average returns of the remaining stocks.

- The plot below depicts a few typical $\boldsymbol{X}_t$ vectors (as a cumulative sum, i.e. $c(i) = \Sigma_i X_{it}$). These plots trend either up or down, reflecting the fact that most stocks are highly correlated and move largely in the same direction contemporaneously.



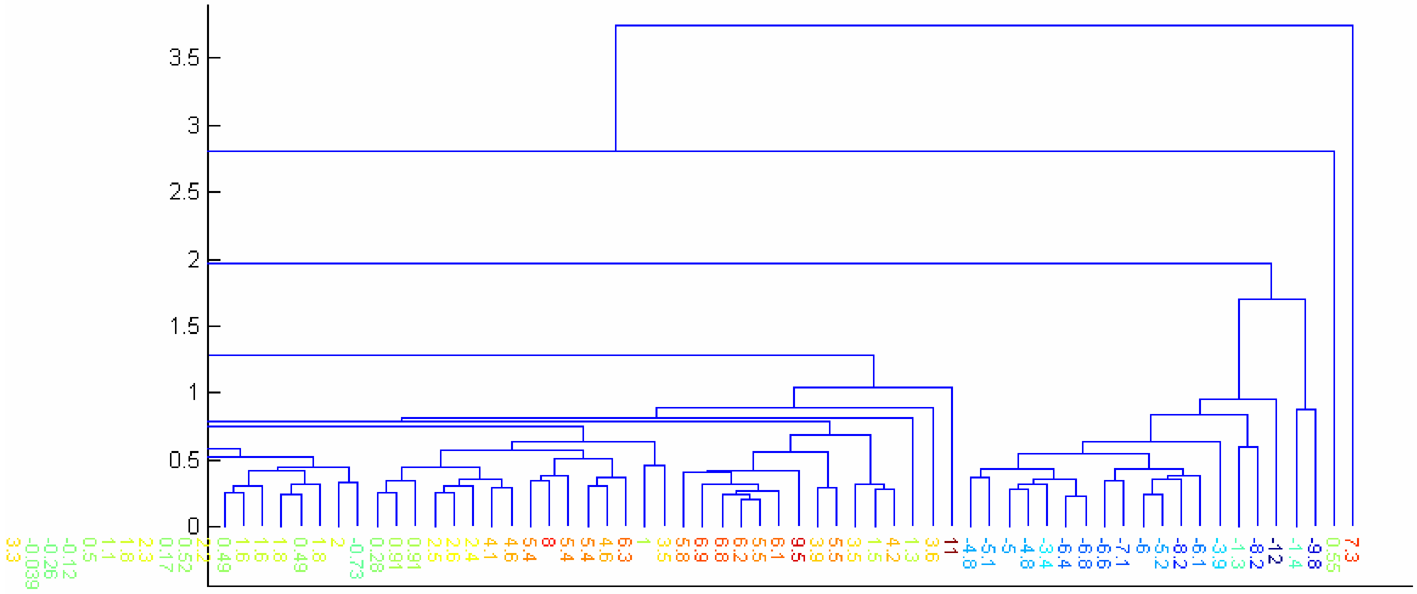
Market states cluster according to mean return of market

- The dendrogram below shows the cluster structure of the market states produced using complete linkage of the Euclidean metric for the returns. The market state vector $\mathbf{X}_t$ is constructed from 3-day returns, i.e. each time step equal to 3 trading days and W=1.
- The nodes at the bottom are labeled with the mean return of the state and color-coded as shown in the legend.
- It is clear that the states cluster according to market return. A few states (to the far right) are very different from the rest and correspond to very negative or very positive returns.



Market states cluster according to mean return of market

- The far right portion of the dendrogram on the previous page is expanded in the plot below.
- Large positive returns cluster together as do large negative returns. The clusters to which both types of returns belong are very different from each other.



Clustering of normalized states

- If we normalized the state vectors, that is, we cluster instead the following objects:

$$\tilde{\mathbf{X}}_t = \frac{\mathbf{X}_t - \overline{\mathbf{X}}_t}{\mathrm{std}(\mathbf{X}_t)}$$

where

$$\overline{\mathbf{X}}_t = \frac{1}{N}\sum_{i=1}^{N} x_{it}$$

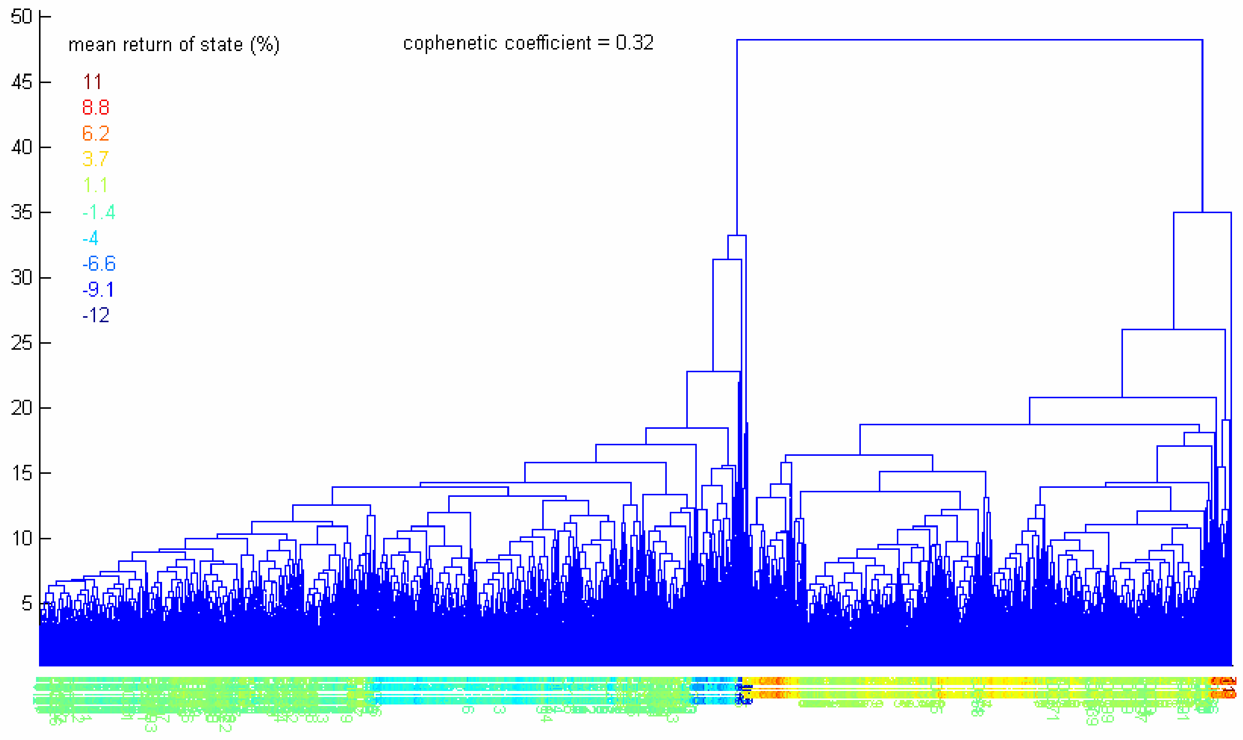
represents a vector with N elements each equal to the same cross-sectional mean stock return and

$$\mathrm{std}(\mathbf{X}_t) = \sqrt{\frac{1}{N-1}\left(\sum_{i=1}^{N} x_{it} - \frac{1}{N}\sum_{i=1}^{N} x_{it}\right)^2}$$

is the cross-sectional standard deviation of the returns, then we cluster based on the time-wise correlation of the return patterns of the states with the magnitude and direction of the mean return of each state normalized out.

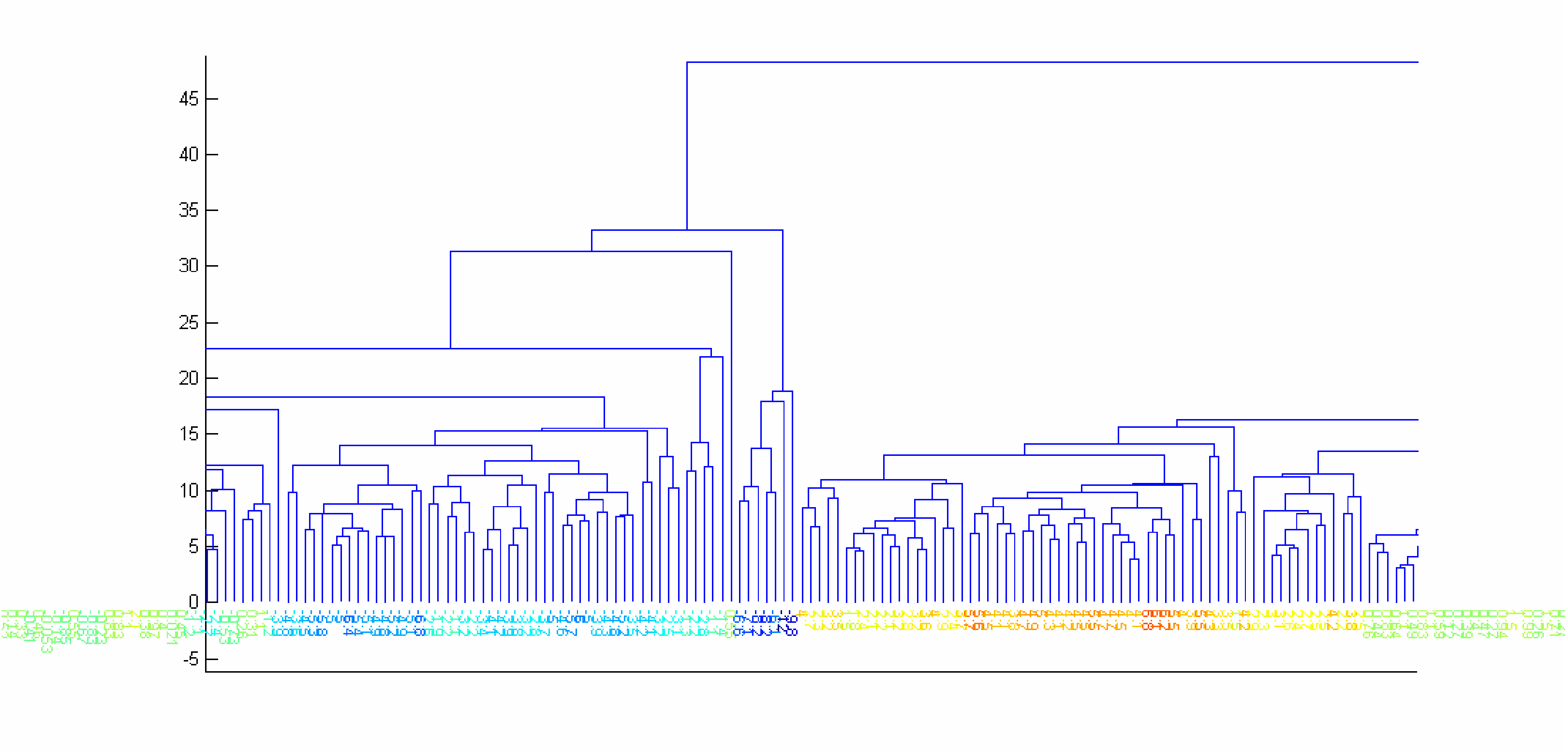
Clustering of normalized states

- The dendrogram of the normalized states produced using complete linkage is shown below.
- As before the nodes are labeled with the mean return of the state (before normalization) and colored coded as shown in the legend.



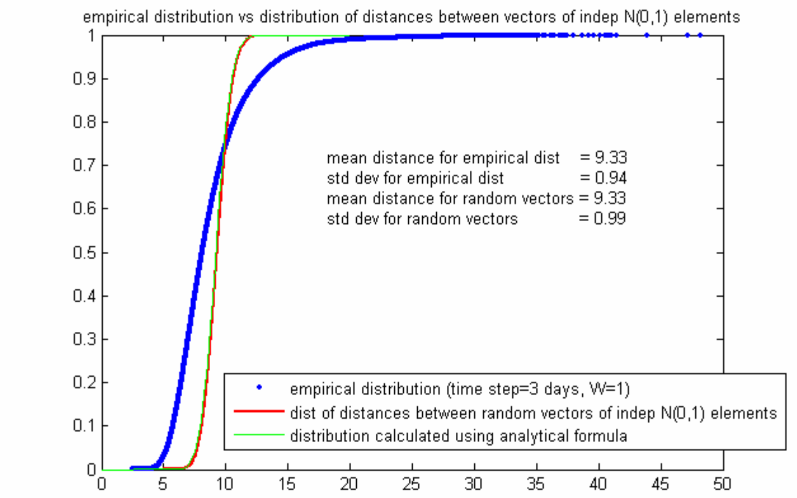
Clustering of normalized states

- The center of the dendrogram on the previous is expanded in the plot below.
- Interestingly, despite normalizing the state vector, states with roughly the same average return still cluster together, indicating that a relationship exists between return and correlation.



Empirical distribution of the distances between states

- The empirical cdf of the distances between normalized market states are shown below together with a reference distribution of distances between vectors whose elements are independent $N(0,1)$ variates.



- The reference distribution is computed two ways: (1) using Monte Carlo to generate random vectors of independent $N(0,1)$ elements (with the same dimension as the market state vector); and (2) using the analytical formula[4]

$$P_n(s)=\frac{s^{n-1}\exp\left(-\frac{s^2}{4\sigma^2}\right)}{2^{n-1}\Gamma\left(\frac{n}{2}\right)\sigma^n}$$

This formula gives the probability of finding a distance s between two points drawn from a spherical (isotropic) gaussian density in n-dimensional space whose pdf is

$$\rho_n(r)=\frac{A}{(2\pi)^{n/2}\sigma^n}\exp\left(-\frac{r^2}{2\sigma^2}\right) \qquad \text{where} \qquad A=\lim_{R\to\infty} n\frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2}+1\right)}\int_0^R \rho_n(r)\,r^{n-1}\,\mathrm{d}r$$

Empirical distribution of the distances between states

- We see from the graph that:

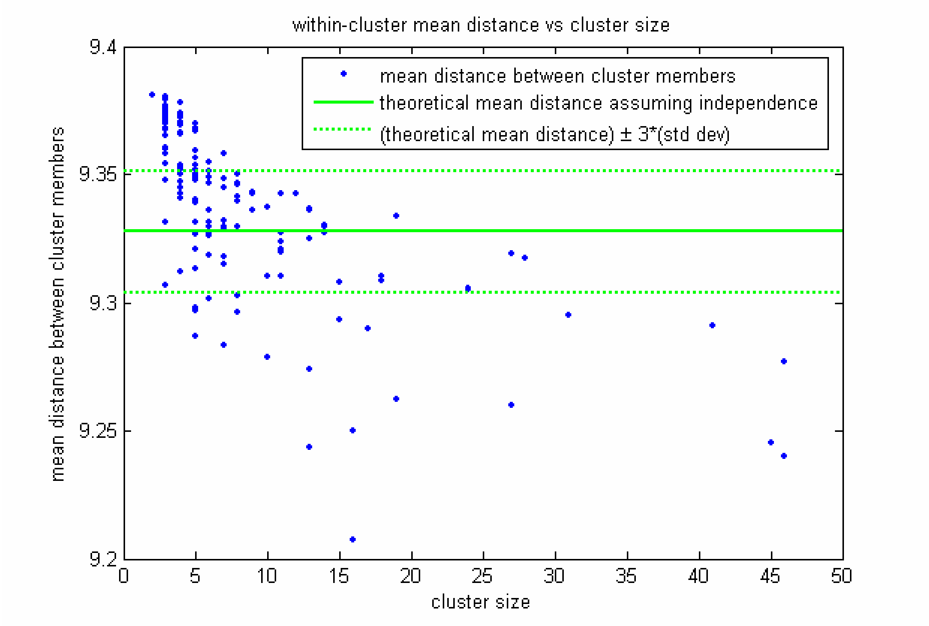
   1. The analytical cdf agrees with the Monte Carlo cdf.

   2. The means of the analytical, Monte Carlo and empirical distributions are all the same.

   3. The empirical distribution has fatter tails and more mass for small distances.

Average distance between market states decreases with cluster size

- Using an inconsistent cut-off value of 1.15, a time step of 3 days for calculating returns and look back window W of one time step, the market states cluster into 211 clusters whose average distance between members of each cluster are shown below versus the size of the cluster.



- Also shown is the theoretical mean distance for independent state vectors of the same dimension and its standard error bounds. The error bounds are computed from 1000 Monte Carlo samples of 20 (which is the typical size of a cluster) random state vectors with $N(0,1)$ variates as elements.

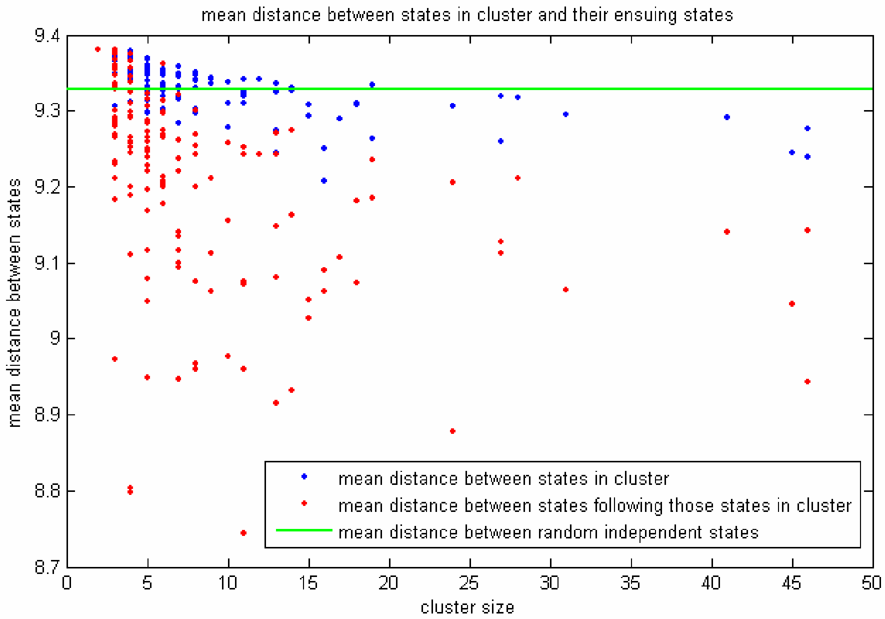
- Large clusters appear to contain states that are statistically significantly closer together then if the states are independent.

Conditional distribution of the successor market states

- One of the key questions of this study is whether the market states that follow in time immediately after the states belonging to a given cluster are also similar in some way.

- If cluster analysis is able to group states together with some common characteristic whose successor states also share some common characteristic, then we may be able to use cluster membership as a “leading indicator”.

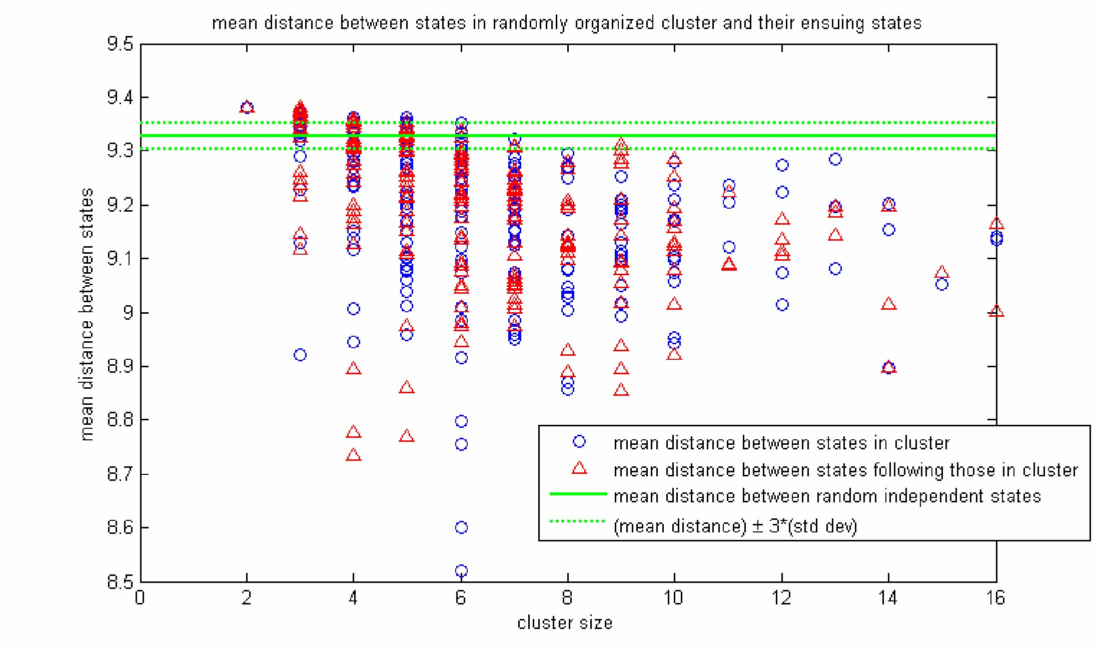
- The mean distance between the successor market states corresponding to predecessor states in each cluster is shown below juxtaposed on the mean distance between the latter.



- Surprisingly, the mean distance between the successor states appears to be generally shorter than that between the predecessor states in the clusters. This effect is strongest for large clusters.

Conditional distribution of the successor market states

- As a control, the mean distance between the successor market states of randomly organized clusters is shown below juxtaposed on the mean distance between the states in the cluster.



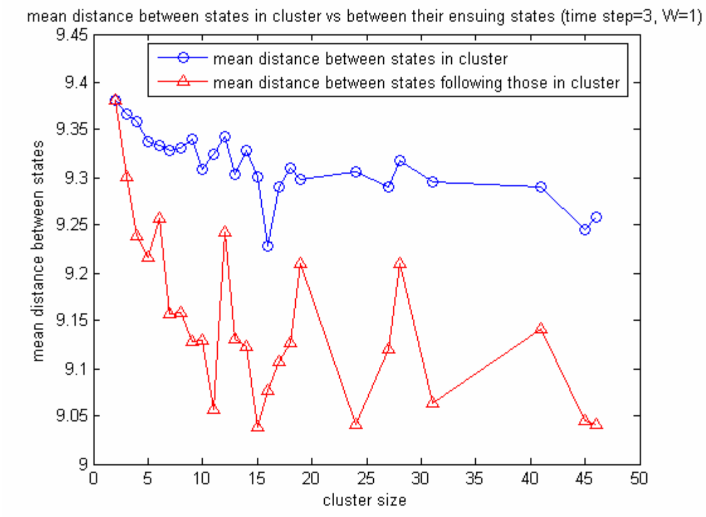
- The clusters were randomly created by generating 1339 (the number of all possible pairs of the market states) random numbers uniformly distributed between 1 and 211 (the total number of clusters based on an inconsistent cut-off of 1.15).

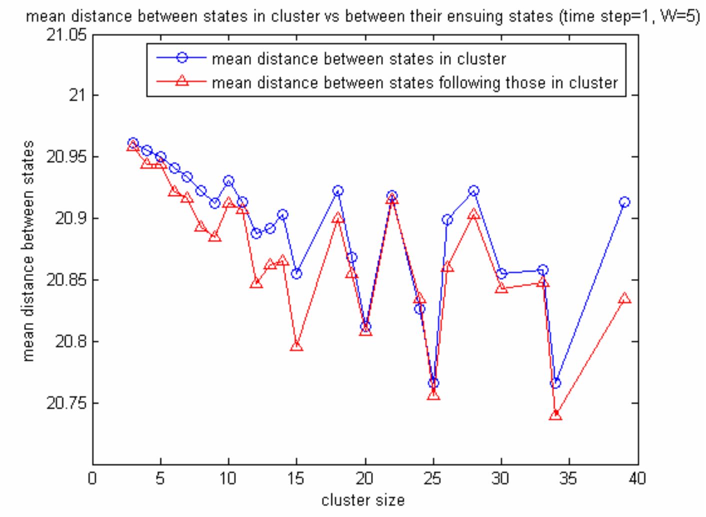
- There is no clear demarcation of the states in the randomly organized cluster versus their successor states, unlike the graph for the states that were organized based on the actual distances.

Conditional distribution of the successor market states

- The fact that the successor states appear to be closer together than their predecessor states is robust to different definitions of the market state.

- For example, we see the same effect if the time step is set to 1 day and W to 5 (cf. graph to the right). In the graphs here all distances between objects in clusters of the same size are averaged.

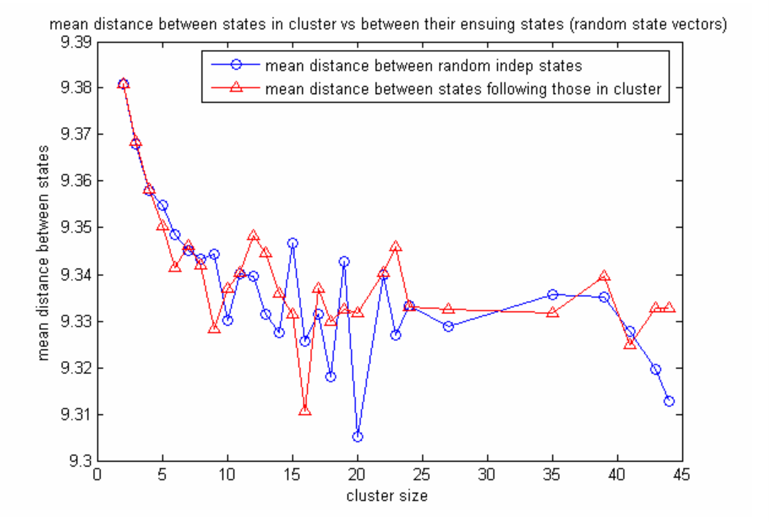


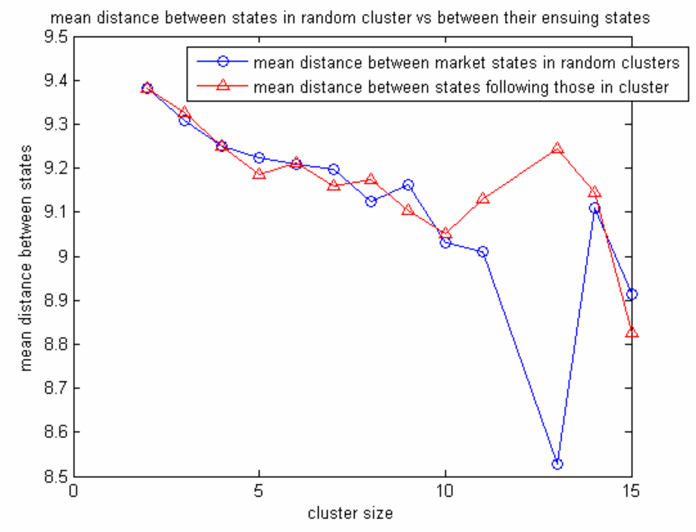


Conditional distribution of the successor market states

- As controls, the mean distance versus cluster size relationship is computed for random state vectors (vectors of random N(0,1) elements) and random clusters.

- In these graphs, there is no distinction between the mean distance between successor states and that between predecessor states.





- The fact that in real-life successor states appear to be closer together is quite baffling. I am quite sure the explanation, if and when I find it, is quite mundane, but I don't have any intuition at this moment.

Summary of Findings

- Cluster analysis is able to discern the economic sector classification of stocks.

- Market states cluster mostly according to the mean return of the state. Large positive or large negative mean return states cluster strongly together, that is to say, the intra-cluster distance is much smaller than the inter-cluster distance.

- The empirical distribution of the distance between market states has the same mean as the theoretical distribution corresponding to random independent state vectors.

- The empirical distribution has fatter tails and more mass at short distances.

- Large clusters have statistically significantly smaller average within-cluster distances then the theoretical average for independent state vectors.

- Successor states have smaller average distances then their predecessor states in clusters. This finding is counter-intuitive and should be further investigated.

Bibliography

[1] M. Tumminello, T. Di Matteo T., T. Aste, and R.N. Montegna. Correlation based networks of equity returns sampled at different time horizons. Eur. Phys. J. B 55, 209–217 (2007)

[2] R.N. Montegna. Information and hierarchical structure in financial markets. Computer Physics Communications 121–122 (1999) 153–156

[3] V. Tola, F. Lilloc, M. Gallegatia, and R.N. Mantegna. Cluster analysis for portfolio optimization. Journal of Economic Dynamics & Control 32 (2008) 235–258.

[4] S-J. Tu and E. Fischbach. Random distance distribution for spherical objects: general theory and applications to physics. J. Phys. A: Math. Gen. 35 (2002) 6557–6570.

[5] CRSP is Center for Research in Security Prices. The CRSP database is arguably the most complete and precise database of historical prices of U.S. stocks.