Introduction

In this report we identify different clusters based on the daily equity return data of market-cap weighted industry indices. We use hierarchical clustering with a **Spearman correlation-based distance** to measure dissimilarity between observations, and **complete linkage** to estimate dissimilarities between clusters. We find an optimal number of **6 clusters** and we conclude that our clustering is robust after employing a stability test based on bootstrapping.

Methodology

- We use hierarchical clustering with Spearman correlation-based distance to measure dissimilarity between observations in our dataset. To identify the best linkage method, we try several options and use the cophenetic correlation as a criterion to choose the method that generates the best result.
- To identify the optimal number of clusters we construct the dendrogram based on the selected hierarchical clustering model. As a robustness check, we employ a more quantitative approach that relies on the Calinski-Harabasz Index, a measurement of cluster variation.
- Based on the optimal hierarchical clustering model and the selected number of clusters, we assign industries to different groups. To check the stability of our clustering model we use an iterative algorithm based on bootstrapping and Jaccard score to quantify the cluster-wise similarities.
- We finally use our domain knowledge to find meaningful interpretations of the clusters and to identify whether certain industries tend to cluster together.

Data cleaning and pre-processing

To illustrate the proposed methodology, we use the 48 industry portfolios from Kenneth French's website. In particular, we use market-cap weighted industry indices. The original data consists of 24,099 observation for each industry index, from January 1926 to October 2017. As a first step in our data cleaning process, we drop the missing observations. This reduces the sample significantly to 12,196 observations for each industry index (from 1969 to 2017). We note that our approach to handling missing values, is reasonable because we do not expect data to be stationary over a long time, so it would be better to consider a restricted time window.

Subsequently we standardise our time series for each industry index to zero mean and unit variance. This is an important step for clustering analysis because clusters are defined based on the distance between points in mathematical space. Therefore, if one of the variables is measured on a much larger scale than the other variables, then whatever measure we use will be overly influenced by that variable.[1]

Clustering analysis

Why Hierarchical Clustering?

Our choice to use hierarchical clustering, is motivated by the seminal work of Mantegna.[2] Returns of different assets display a high dependency, across industries and asset classes. This is generally explained in terms of synchronizations among market participants, due to common flows of information and overlapping investment strategies. In addition to exploring the correlations

[1]University of California, Berkley: "Cluster Analysis", https://www.stat.berkeley.edu/~s133/Cluster1.html
[2] Mantegna R. (1998): "Hierarchical Structure in Financial Markets', https://arxiv.org/pdf/cond-mat/9802256.pdf"

between asset returns, we are also interested in discovering and extracting hierarchical structure within data.

Therefore, in this report, we employ hierarchical clustering with a correlation-based distance to measure dissimilarity between observations. The correlation coefficient itself cannot be used as a distance between pairs of financial instruments because it does not fulfil the three axioms that define an Euclidean metric.[3]

A generalized metric can be defined to use correlation coefficient as an appropriate measurement of distance. The formulation of the correlation-based distance between observations i and j is given by: $d(i,j) = \sqrt{\left(2 \cdot \left(1 - \rho_{i,j}\right)\right)}$, where $\rho_{i,j}$ is the correlation between observations i and j.

Choice of correlation coefficient

To inform our choice of correlation coefficient, we have visually analysed the pairwise relationship between industry indices in our dataset. We observe that:

- Histograms of returns for each industry index suggest that returns are not normally distributed. This is consistent with the result we get form the Shapiro-Wilk test for normality.[4]
- The scatter plots do not support linearity between pairs.

These observations suggest that Pearson correlation coefficient might not be appropriate for our analysis. Instead, we need to use a correlation measure which: i) does not carry any assumptions about the underlying distribution of the data and ii) captures non-linear dependencies. Spearman's correlation is an appropriate choice,[5] and it is measured as the Pearson correlation between the rank values of two variables: $\rho_{(i,j)} = \frac{Cov(Rank_x, Rank_y)}{\sigma_{Rank_x}\sigma_{Rank_y}}$.

How does the Hierarchical Clustering Algorithm work?

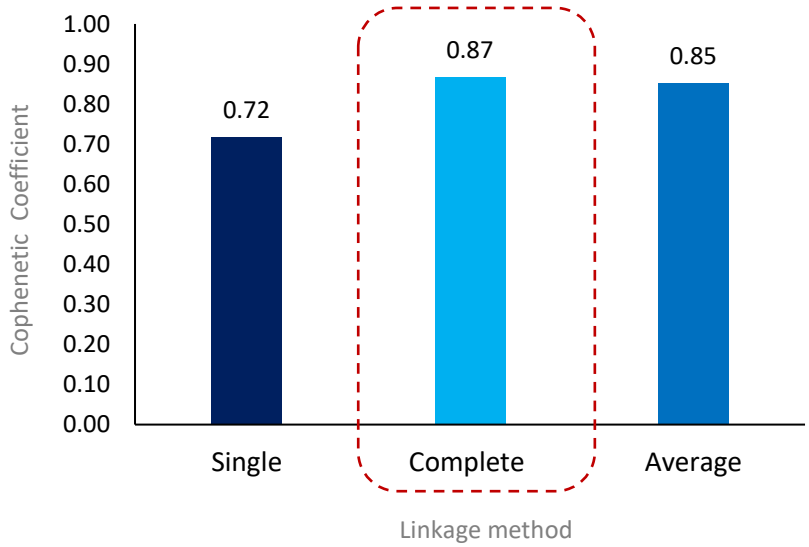
- The idea is to build a hierarchy of clusters by first merging nodes and then clusters. We start with N nodes, and each node is its own cluster.
- We use the **Spearman correlation-based dissimilarity** to compute the distance between pairs of observations.
- We extend the definition of dissimilarity so we can compute the distance **between pairs of clusters**. We will try different linkage methods and select the one that generates **the highest cophenetic coefficient.** [6] We have summarised the definitions of various linkage methods in Appendix A1.1.
- We go through an iterative procedure where at each step we merge the two clusters that are the closest according to the dissimilarity measure.
- The final output is a dendrogram, that is a tree showing the hierarchical structure. A cluster partition can be obtained by choosing the number of clusters and cutting the dendrogram at the appropriate level.

[3] Chehreghani et al. (2020): "Hierarchical Correlation Clustering and Tree Preserving Embedding"
[4] Shapiro, S. S.; Wilk, M. B. (1965). "An analysis of variance test for normality (complete samples)"
[5] Corder, G. W. & Foreman, D. I. (2014). "Nonparametric Statistics: A Step-by-Step Approach"
[6] Saracli et al. (2013): "Comparison of hierarchical cluster analysis methods by cophenetic correlation".

Selecting the optimal linkage method

To identify the optimal linkage method in our analysis we rely on the cophenetic correlation coefficient. Since its introduction by Sokal and Rohlf, the cophenetic correlation coefficient has been widely used as a criterion for evaluating the efficiency of various clustering techniques. More precisely, the cophenetic coefficient is a measure of how faithfully a dendrogram preserves the pairwise distances between the original unmodeled data points. [7] The higher the cophenetic values, the more the ultrametric distances are representative of the true distances. [8] See Appendix 1.2 for further details. Based on the cophenetic correlation coefficient we choose **'complete linkage'** as the optimal method for our hierarchical clustering.

Figure 1: Cophenetic Coefficient for different linkage methods



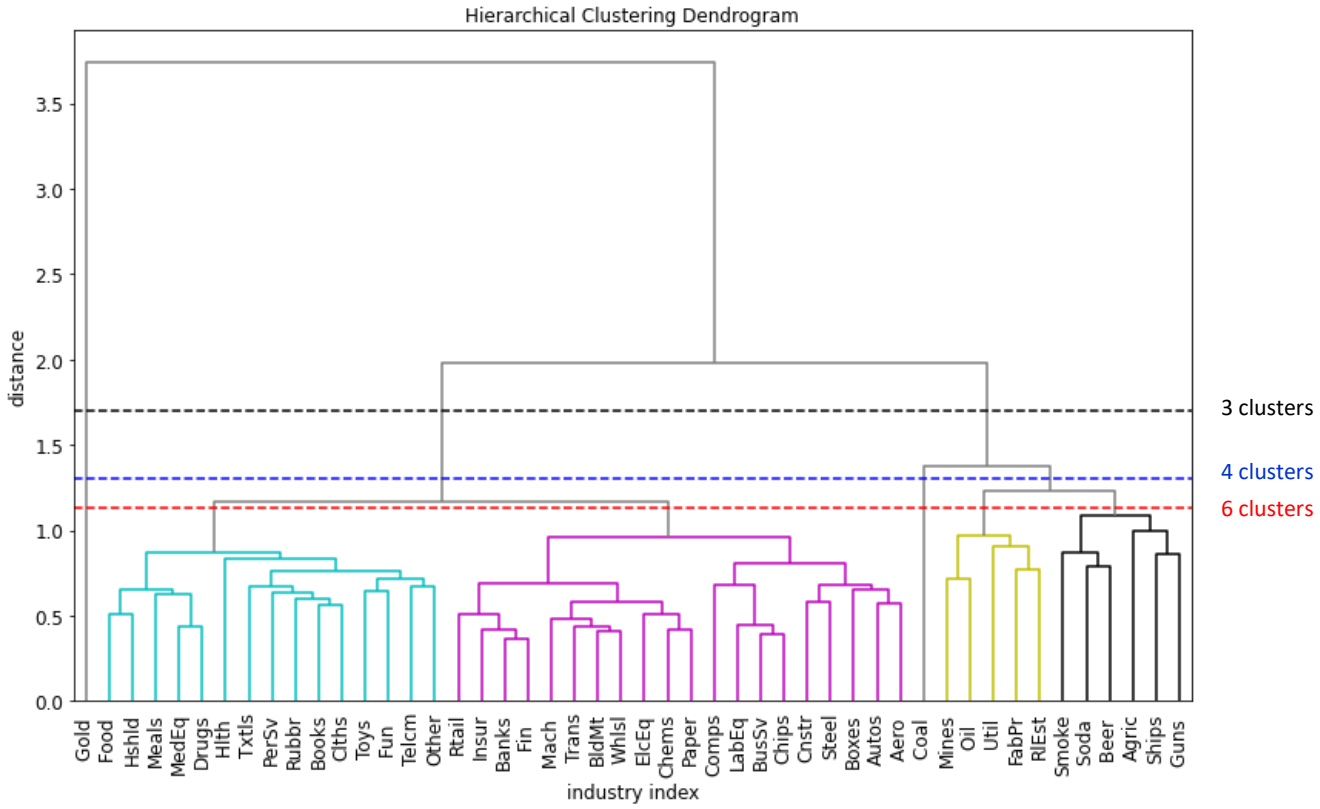
Selecting the optimal number of clusters

To make an informed decision on the optimal number of clusters, we initially inspect the hierarchical structure of the dendrogram. Dendrogram displays the distance between each pair of sequentially merged objects in a feature space. One common approach to decide the number of clusters is to analyse the dendrogram and look for groups that combine at a higher distance.[9] In our dendrogram, we construct three lines representing different distance thresholds to qualitatively identify the number of clusters implied by each line.

We obtain **6 clusters** when we cut the tree at a distance threshold of 1.13. This choice does not support the idea that the identified clusters are combined at a higher dendrogram distance (3 clusters would be more appropriate for that), however it **can lead to a more intuitive interpretation of the cluster components.**

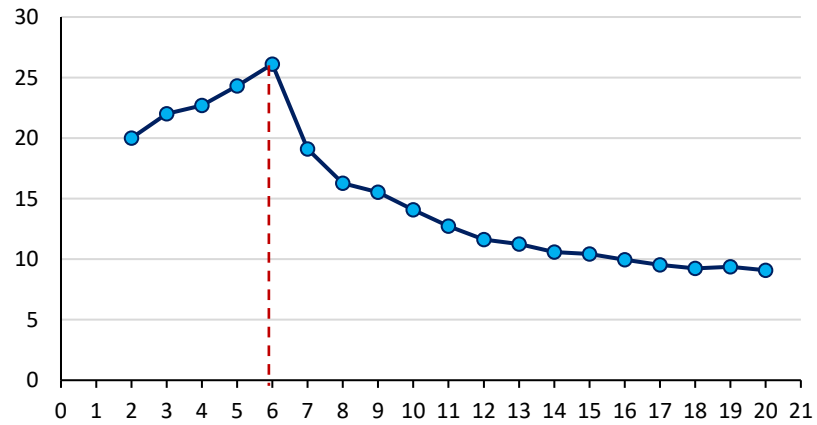
[7] Sokal R, Rohlf F: "The Comparison of Dendrograms by Objective Methods"
[8] Ultrametric distances represent the height of the link at which two elements are first joined together in the dendrogram.
[9] The general idea being that the identified clusters are combined at a much higher dendrogram distance and hence can be treated as individual groups for this analysis.

Figure 2: Dendrogram of our hierarchical clustering algorithm based the correlation dissimilarity measure and complete linkage



Moreover, we can further cross-check our choice using a more robust quantitative approach based on the **Calinski-Harabasz Index**, also known as the variance ratio criterion.[10] CH index is the ratio of the sum of between-clusters dispersion[11] and of inter-cluster dispersion for all clusters.[12] Well separated and compact clusters should maximize this ratio. Further details on how it is calculated can be found in Appendix A1.3. Figure 3 suggests **that the optimal number of clusters is 6,** which supports our choice above.

Figure 3: Estimating Calinski-Harabasz index for different number of clusters



[10] Calinski T , Harabasz J . "A dendrite method for cluster analysis". Communications in Statistics, 1974
[11] Between group sum of squares that measures the dissimilarity between different clusters.
[12] Within Group Sum of Squared that measures dissimilarity within clusters.

Cluster Stability

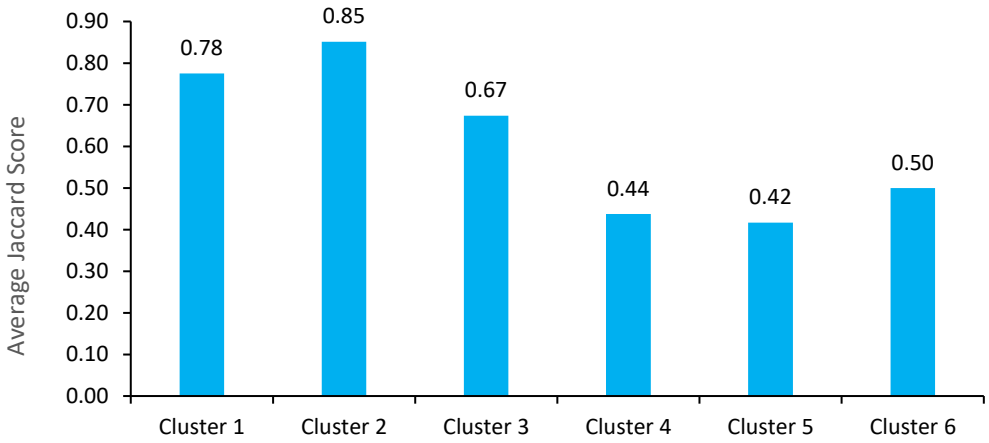
An important question when evaluating clusters is whether a given cluster is "real"- i.e., does the cluster represent actual structure in the data? Validation is very important in cluster analysis, because clustering methods tend to generate groupings even for fairly homogeneous data sets. Therefore, we need to check the stability of our clustering algorithm. Stability means that a meaningful valid cluster should not disappear easily if the data set is changed in a non-essential way. One way to assess that, is to see if the cluster holds up under plausible variations in the dataset.[13]

We use **bootstrapping with resampling to evaluate the cluster-wise stability of our model.** We employ the methodology below: [14]

- Cluster the original data using the hierarchical clustering algorithm defined before.
- Draw a bootstrap sample of the same size by resampling the original data with replacement. Cluster this new data.
- Compute the Jaccard similarities of the original clusters to the most similar clusters in the resampled data.[15] The mean over these similarity scores is used as an index of the stability for each cluster.
- Repeat step 2 and 3 for 1000 bootstrap iteration.

As a rule of thumb, clusters with a stability value less than 0.6 should be considered unstable. Values between 0.6 and 0.75 indicate that the cluster is measuring a pattern in the data, but there is not high certainty about which points should be clustered together. Clusters with stability values above 0.85 can be considered highly stable. **Based on our analysis, Clusters 1,2 and 3 can be considered stable. We conclude that our clustering is generally stable** because clusters 1,2 and 3, correspond to the majority of our observations i.e., 40/48 industries indices in our dataset.

Figure 4: Average Jaccard Score for each cluster (1000 iterations)



[13] Hennig C.: "Cluster-wise assessment of cluster stability", Department of Statistical Science, University College London
[14] Ibid.
[15] The Jaccard coefficient measures similarity between finite sample sets and is defined as the size of the intersection divided by the size of the union of the sample sets.

Cluster Interpretation

Based on the specifications of the hierarchical clustering model and the optimal number of clusters discussed above, we use Python's fcluster function to assign industries to different clusters.[16] The distribution of industry indices across different clusters is summarised in Appendix A1.4.

Economic intuition supports the idea that the systematic component of returns is the main source of variation for firms in similar industries. Indeed, in the dendrogram above, we observed that industries which are highly correlated, i.e., have a small distance to their neighbours, are linked together at an early stage. One example is the financial services industry i.e., Banks, Trading, and Insurance. The same pattern is consistently observed for other industries as well, such as: Medical Equipment and Drugs; Food products and Consumer goods; Construction and Steel etc.

However, industry indices can be categorized into homogeneous groups using criteria other than industry affiliation. Groupings can be based on similarity with respect to firm attributes such as market capitalization, targeted end-customer (i.e., B2B vs. B2C), valuation ratios, commonality in operating performance etc.[17] This can explain the clustering of industry indices that are seemingly unrelated. For instance, we can see that Wholesale, Business Supplies, Construction, Aircraft etc. are categorised in the same cluster. Despite being seemingly unrelated in terms of the industry affiliation, these firms may share other commonalities such as: their B2B target, high market capitalisation, high valuation ratios etc.

Based on the selected optimal number of clusters we can observe that industries are not evenly distributed across different groups. Clusters 1 and 2 hold a particularly large amount of industry indices, while clusters 5 and 6 consist of a single index. The classification of Gold in a separate category is not surprising, because Gold is generally considered as a 'safe haven' asset. As such it has a negative correlation with most of the industries in our sample, which implies a greater distance from other observations based on our correlation dissimilarity metric.

In our post-cluster analysis, we have attempted to find meaningful interpretations of clusters that go beyond direct industry affiliations (*see Appendix A1.4 for further details).* For instance:

- **Cluster 1: Healthcare and consumer** firms with a B2C target and greater diversity in terms of firm size.
- **Cluster 2: Financial services, Technology and Transport** companies, with a highly B2B focus. Average market capitalisation might be significantly higher than for other clusters. (i.e., dominated by big firms). Industry indices in this cluster may be characterised by relatively higher P/E ratios.
- **Cluster 3: Energy** firms.
- **Cluster 4: Import/Export** firms.
- **Cluster 5: Coal** (standalone cluster due to low correlation with other indices)
- **Cluster 6: Gold** (standalone cluster due to 'safe haven' asset).

A potential practical application of our clustering algorithm would be to construct diversified portfolios, consisting of the best performing indices for each cluster.

[16] https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.fcluster.html

[17] Chan et al. (2007): "Industry classifications and the co-movement of stock returns"

Appendix

A1. 1 Linkage methods

In this appendix we summarise the definitions of different linkage methods that we have explored in our hierarchical clustering analysis:

- **Single linkage:** the dissimilarity between two clusters is the ***minimum*** dissimilarity between pairs of observations in the 2 clusters. It is computed as:
  - $D(A, B) = \min_{i \in A, j \in B} d(i, j)$
- **Complete linkage:** the dissimilarity between two clusters is the ***maximum*** dissimilarity between pairs of observations in the 2 clusters. It is computed as:
  - $D(A, B) = \max_{i \in A, j \in B} d(i, j)$
- **Average linkage:** the dissimilarity between two clusters is the ***average*** dissimilarity between pairs of observations in the 2 clusters. It is computed as:
  - $D(A, B) = \text{mean}_{i \in A, j \in B} d(i, j)$

A1. 2 Cophenetic Correlation

The cophenetic correlation coefficient (CC), gives a measure of the correlation between two matrices: D, the matrix of the distances and Δ, the matrix of the ultrametric distances (i.e., levels in a dendrogram at which the pairs of points join).

It is defined as: $CC = \frac{\sum(d_{ij} - \overline{D})(\delta_{ij} - \overline{\Delta})}{\sqrt{\sum(d_{ij} - \overline{D})^2 \sum(\delta_{ij} - \overline{\Delta})^2}}$

where d_{ij} is the distance between elements i and j in D, δ_{ij} is the ultrametric distance between elements i and j in Δ; i.e., the height of the link at which the two elements i and j are first joined together, and $\overline{D}$ and $\overline{\Delta}$ are the average of D and Δ, respectively. The higher the CC values, the more the matrix Δ is representative of matrix D and, consequently, the more the ultrametric distances are representative of the true distances.[18]

A1. 3 Calinski-Harabasz index

Calinski-Harabasz index is a measurement of cluster variation and it is estimated as follows:

$$\frac{SS_B}{SS_W} \times \frac{N - k}{k - 1}$$

Where k is the number of clusters; and N is the total number of observations; SSW is the overall within-cluster variance; SSB is the overall between-cluster variance. SSB measures the variance of all the cluster centroids from the dataset's grand centroid (A big SSB value means that the centroid of each cluster will be spread out and they are not too close to each other). As we increase the number of clusters, the within-cluster variance (SSW) will decrease. Therefore, for the Calinski-Harabasz Index, the ratio of SSB/SSW should be the biggest at the optimal clustering size.[19]

---

[18] Battaglia et al. (2019): "Unsupervised quantitative methods to analyze student reasoning lines: Theoretical aspects and examples"

[19] https://ethen8181.github.io/machine-learning/clustering_old/clustering/clustering.html

A1. 4 Distribution of industry indices in different clusters

Table 1: Distribution of industry indices across different clusters[20]

Clusters	Industry index	Cluster interpretation
1	• Textiles; Rubber and Plastic Products. • Medical Equipment; Healthcare; Pharmaceutical Products; Telecommunication • Clothes; Household; Fun; Books; Food; Personal Services; Meals; Toys • Others	-Healthcare and consumer firms. - Dominated by B2C firms (i.e., industries that target end-customers) - More diversified in terms of market capitalisation (i.e., we expect small firms to comprise a significant portion of this cluster)
2	• Insurance; Trading; Banks; • Retail; Wholesale; Business Supplies • Transportation; Aircraft; Automotive; Shipping Containers; • Construction materials; Steel; Electrical Equipment; Machineries; Measuring and Control Equipment; Chemicals • Electronic Equipment; Computers; Business Services	-Financial services, Technology, Transport and Construction - Dominated by B2B firms (industries that target business customers) -Average market capitalisation might be significantly higher than for other clusters. (i.e., dominated by big firms) -Similar performance in terms of valuation ratios (i.e, may be characterised by generally high P/E ratios)
3	• Mines; Real Estate; Fabricated Products; Utilities; Oil	- Energy companies
4	• Agriculture; Tobacco Products; Candy & Soda; Beer & Liquor • Shipbuilding, Railroad Equipment; Defence	- Import/Export firms - This cluster is dominated by industries that are characterised by a high rate of trading activity (i.e for import/export purposes).
5	• Coal	Coal
6	• Gold	Gold

[20] We have used this website to obtain details on the industry definitions and abbreviations: http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library/det_48_ind_port_old.html