Unsupervised Learning on Daily Equity Return

Rik Harng Loong
Student ID: 18061593

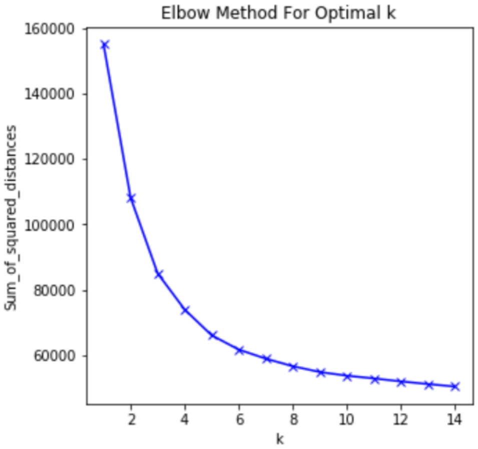
February 12, 2020

1 Introduction

In this report, we first discuss two common clustering algorithms: K-means and Hierarchical Clustering applied on our daily equity returns data for 48 industries. We aim to identify which industries tend to cluster together and see if our results make sense. Following this, we perform Principal Component analysis (PCA) to reduce the dimension of our features (i.e. industries) while preserving the maximum information in the original data. Additionally, we can gain a better understanding of which industries are related to each other from the output of PCA.

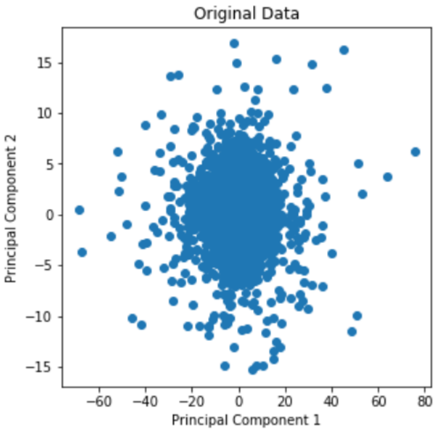
2 K-means Clustering

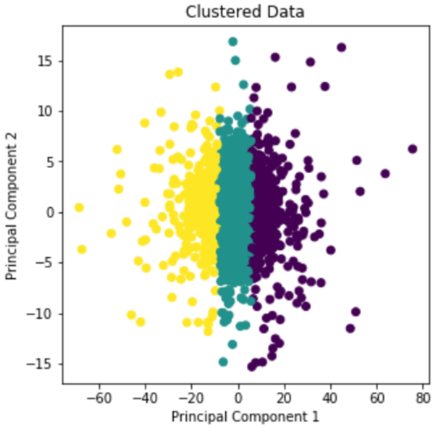
Our data consists of 3231 observations with 48 features (i.e. industries) from year 2005 onward until 2017. We aim to partition these observations into K distinct clusters satisfying two properties: (1) Each observation must belong to at least one cluster. (2) The K clusters are non-overlapping [1]. Note that K is the number that has to be specified beforehand. We first identify the optimal number of K which minimizes the sum of squared Euclidean distances of observations to their nearest cluster centre. This is done via the so-called elbow method. We then validate our result by computing the Silhouette coefficients which assess the performance of the resulting clusters. This coefficient lies within [-1,1] with value near 1 indicates perfect clustering, 0 indicates an overlapping cluster and -1 indicates wrong clustering [2].



Silhouette coefficient for k^{th} cluster
k = 2 : 0.2787786148444873
k = 3 : 0.3001364299136892
k = 4 : 0.2247316117132846
k = 5 : 0.1940334436654053
k = 6 : 0.1505659144228625
k = 7 : 0.1118525182326850
k = 8 : 0.1142517714164020

From the plot above, the optimal k is on the elbow which is 3. Indeed, it is found that the Silhouette coefficient is the highest also for k = 3. Let us reduce the dimension of our features by using PCA for the sake of visualizing the clusters on the 2D graph.

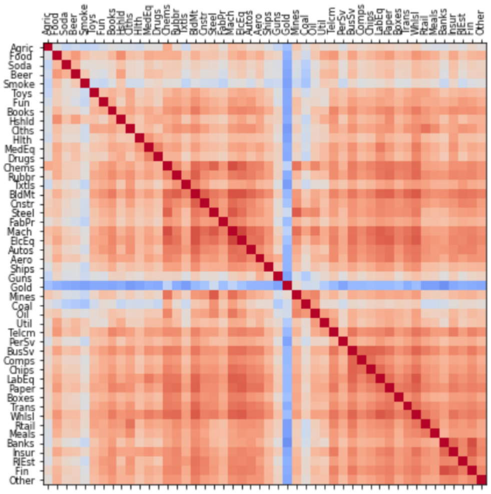




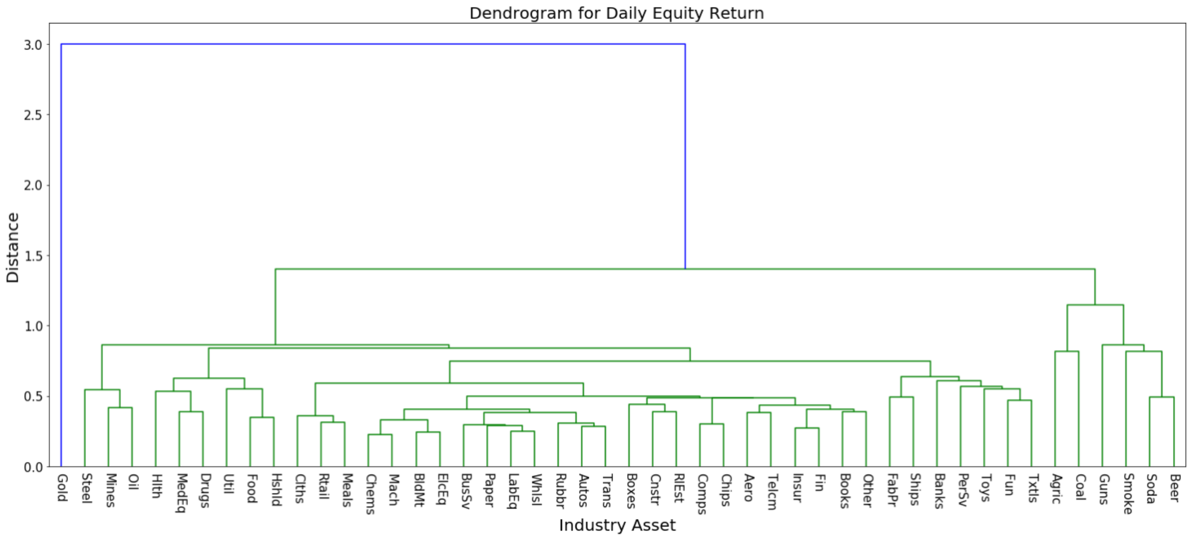
It is not clear that how many clusters exist in the original data. In fact, one of the assumptions of K-means clustering stating that all features must have equal variances is broken in this case. Moreover, the ”messy” clusters are not well separated here so it is reasonable to say that K-means does not work well on our data. We do not gain any insights about clustering of the industries. Instead, let us consider hierarchical clustering which gives a much bigger picture than K-means.

3 Hierarchical Clustering

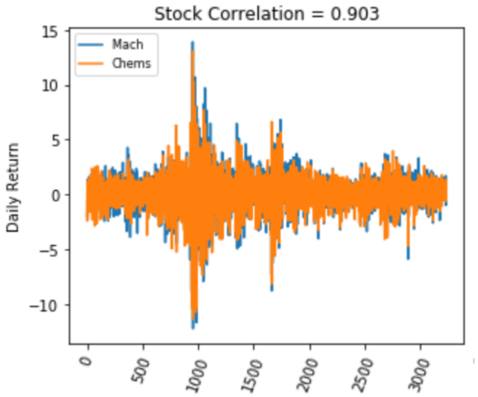
Unlike K-means, we do not have to specify the parameter K. We will focus on bottom-up (or agglomerative) hierarchical clustering. This approach looks for clusters without having to know exactly what these clusters mean. In order to cluster different industries, we need a distance measurement between them. The most well-known one which we will use is the standard correlation coefficient. The correlation matrix is plotted below with the red/blue colours indicate high/low correlation. Notice that a perfect correlation lies on the diagonal as expected since all industries have a perfect correlation with themselves.

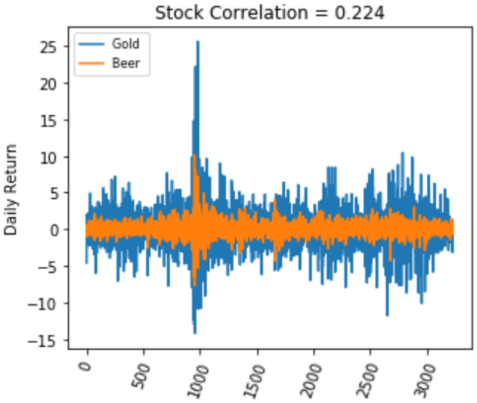


One primary advantage of this approach over K-means is that we do not have to guess how many clusters there might be in our data. The algorithm works by first assigning each industry to its own cluster, and then merges pairs of clusters from the bottom to the top of the hierarchy. At each particular stage, the clusters are joined together using the notion of linkage. We will use average linkage here which measures the average distance between elements of each cluster. The reason is that average linkage method works well in most situations and possesses a certain robustness with respect to slight distortions [3]. We also tend to use standardized features for equally weights in their distance representations. The best way to visualize this clustering is through a dendrogram.



It gives us a deep insight of the convergence of different clusters at each step and we have the flexibility to decide where to cut the dendrogram to figure out the desired number of clusters. The main interpretation of the dendrogram is that the longer the distances (as indicated by the vertical lines) are, the less correlated two clusters are. Based on the dendrogram above, the two most correlated industries are **Chemicals** and **Machinery**. It seems like the machinery equipment often involves heavy use of chemical substances, so it makes sense that they would be strongly correlated. Let us plot the graph below to see how well they correlate, and also pick two industries that are not well correlated, say **Gold** and **Beer** to compare.



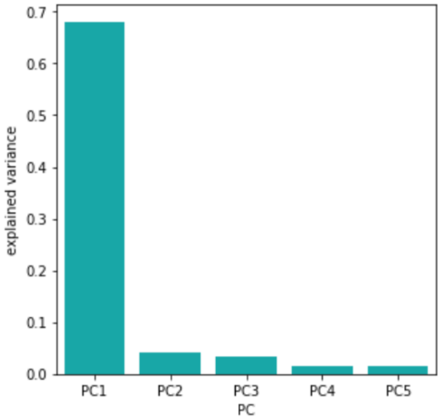


Indeed the plots support our claims above. In general, the best choice of number of clusters is the number of vertical lines that can be cut by by a horizontal line, that can traverse the maximum distance up and down without intersecting other cluster [4]. In our case, best choice for number of clusters will be 3 (the cut-off point is somewhere between distance 1.2 and 1.4). It also matches the optimal number of clusters that we get from the result of K-means. For instance, **Beer**, **Soda** and **Tobacco products** tend to cluster together. It makes sense as they are frequently bought together. The **Steel** up until **Textiles** industries on the dendrogram above tend to cluster together, and finally **Gold** forms its own cluster. **Gold** are precious metals and neither of the industries activities affect their sales, so it makes sense that it does not correlate well with other industries.

However, just by knowing these 3 clusters is not enough to gain an insight into our data. We need to break down the large cluster that contains from **Steel** up until **Textiles** industries into smaller sub-clusters. There is no hard rule getting the optimal number of clusters. Instead, we can choose our desired number of clusters based on our intuition or to see if the combinations of the cluster make sense. For instance, **Steel**, **Mines** and **Oil** are clustered together perhaps they belong to the oil and mining industry. **Healthcare services**, **Medical equipment** and **pharmaceutical products** tend to cluster together to form the basis of the healthcare industry. **Utilities**, **Food products** and **Consumer goods** form the basis of the typical household industry. We can keep inspecting the sub-clusters to better understand the clustering of certain industries.

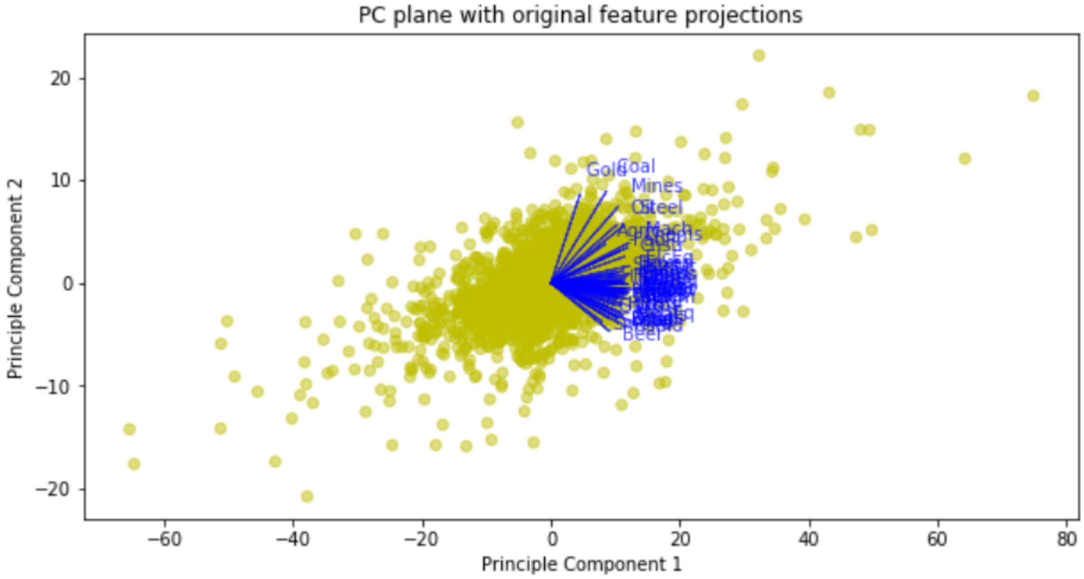
4 Principal Component Analysis

Assuming that our daily returns data are measured on the same scale, we still need to standardize the data by transforming it to the unit scale (zero mean and unit variance). Eigendecomposition is then performed on our covariance matrix generated from the standardized data. Each eigenvector is obtained as a linear combination of the covariance matrix with its corresponding eigenvalue, and the sum of the eigenvalues represents all the variance in the entire data. Next, we wish to select the number of eigenvectors, also known as the principal components (normalized linear combination of the 48 features in the original data) that explain most of the variance in our data. The plot below shows the explained variance by our principal components:



From the plot, we can see that the first principal component explains pretty much most of the variation (information) within our data. In my opinion, it is good enough to select the first 2 principal components that account for around 72% of the variance in the entire data.

Note that our data contains 48 features and it would be hard to visualize how they are related with each other on a graph. PCA has helped us to reduce down to only two features, and this is a major reduction from the initial 48. Let us visualize our two principal components on a biplot to see the relation between the components and the original features.



We can see from the plot that all industries are positive in the first principal component and most of the industries are negative in the second principal component. The length of the line represents the magnitude of their relations. It is clear that **Mines**, **Steel** and **Oil** are aligned towards the same direction (PC 2), and so does the **Machinery** and **Chemicals**. These look intuitive as we have already seen there seems to be a linear relationship between the group of steel, mining and oil industries. The same goes for both the machinery and chemical industries.

5 Conclusion

It is found that hierarchical clustering approach produces a more informative clustering of the industries than K-means. The disadvantage of using K-means in our case is that it is difficult to guess the number of clusters in advance, unless we have been told by some domain experts. The initial seeding/choosing initial cluster centre also strongly affects the final result of K-means [5]. Hence, hierarchical clustering is more favourable for the analysis of our dataset. However, if we were to deal with a much larger dataset (e.g. 100 features), the process may take a while and the dendrogram becomes much harder to interpret visually. Finally, even though PCA is useful in exploring patterns in a large dataset, we need to be aware of some of its limitations. The most crucial one is that PCA assumes linearity of features in each principal component. Therefore, PCA may not give a sensible result if the features are related in a non-linear manner.

References

[1] https://moodle-1819.ucl.ac.uk/pluginfile.php/362752/modresource/content/1/Part3 ULv1.pdf [online]

[2] Plot.ly. (2019). Selecting the number of Clusters with Silhouette Analysis on KMeans Clustering. [online]

[3] https://www.quora.com/What-is-the-best-linkage-criterion-for-hierarchical-cluster-analysis [online]

[4] https://medium.com/data-science-group-iitr/clustering-described-63e62833099e [online]

[5] http://santini.se/teaching/ml/2016/Lect10/10c UnsupervisedMethods.pdf [online]