GROUP WORK PROJECT # ___
GROUP NUMBER: __6217_____

FULL LEGAL NAME	LOCATION (COUNTRY)	EMAIL ADDRESS	MARK X FOR ANY NON-CONTRIBUTING MEMBER
Yang Xue	China	KierenXue@gmail.com	
Carlos García Guillamón	Spain	carlosgguil@outlook.es	
Ebenezer Yeboah	Ghana	ebenezeryeboah46@gmail.com	

Statement of integrity: By typing the names of all group members in the text boxes below, you confirm that the assignment submitted is original work produced by the group (excluding any non-contributing members identified with an "X" above).	
Team member 1	**Yang Xue**
Team member 2	**Carlos García Guillamón**
Team member 3	**Ebenezer Yeboah**

Use the box below to explain any attempts to reach out to a non-contributing member. Type (N/A) if all members contributed. **Note:** You may be required to provide proof of your outreach to non-contributing members upon request.

1. On top left of your screen click on File →Download → Microsoft Word (.docx) to download this template
2. **Upload the template in Google Drive and share it with your group members**
3. **Delete this page with the requirements before submitting your report. Leaving them will result in an increased similarity score on Turnitin.**

Keep in mind the following:

- Make sure you address all the questions in the GWP assignment document published in the Course Overview.
- Follow the "Submission requirements and format' instructions included in each Group Work Project Assignment, including report length.
- **Including in-text citations and related references is mandatory for all submissions.** You will receive a '0' grade for missing in-text citations and references, or penalties for partial completion. Use the **In-Text Citations and References Guide** to learn how to include them.
- Additional writing aids: Anti-Plagiarism Guide, Academic Writing Guide, Online Writing Resources.
- To avoid an increase in the Turnitin similarity score, **DO NOT copy the questions** from the GWP assignment document.
- Submission format tips:
    - Use the same font type and size and same format throughout your report. You can use Calibri 11, Arial 10, or Times 11.
    - Do NOT split charts, graphs, and tables between two separate pages.
    - Always include the axes labels and scales in your graphs as well as an explanation of how the data should be read.
- Use the LIRN Library for your research. It can be accessed via the left navigation pane inside the WQU learning platform.

The PDF file with your report must be uploaded separately from the zipped folder that includes any other types of files. This allows Turnitin to generate a similarity report.

Step 1

The categories and topics chosen have been:

- Category 1:

BASICS

LASSO Regression is short for **least absolute shrinkage and selection operator**. It has a penalty function:

$$f() = \sum_{j=1}^{p} |\beta_j| = ||\beta||_1$$

In this category, penalty regressions are discussed for this case. Penalty regressions are dealt with too many variables or high dimension data for regression analysis. A penalty function is introduced in this case, as it can pull the value of a coefficient close to 0 when the coefficient value is too large. The action of pulling a coefficient value towards 0 or to a certain preset value is called **shrinkage.**

For the Lasso regression, it is also a type of penalty regression. The above penalty function is also called as L1 penalty function. This is because the function is the absolute value of the magnitude of coefficients in the model.

As LASSO regressions are still based on the regression method, hence, it is a supervised learning when it tries to model data and relationships.

Keywords:

- ❖ Regression
- ❖ Penalty function
- ❖ Coefficient shrinkage

- Category 2:Hierarchical Clustering

BASICS

Hierarchical Clustering: Hierarchical clustering is a method of cluster analysis which aims to build a hierarchy of clusters. It is an unsupervised machine learning technique used to group similar data points into clusters based on their similarity. There are two approached:

- ❖ Agglomerative Hierarchical Clustering (Bottom-Up Approach): This approach starts with each data point as its own cluster and merges the closest pairs of clusters iteratively until only one cluster remains or the desired number of clusters is reached.

- Divisive Hierarchical Clustering (Top-Down Approach): This approach starts with all data points in a single cluster and recursively splits clusters into smaller clusters until each data point is its own cluster or the desired number of clusters is achieved.

Measures of Distance

- Euclidean Distance: The straight-line distance between two points in Euclidean space.

  $sqrt((x2 - x1)\text{^}2 + (y2 - y1)\text{^}2)$

- Hamming Distance: Counts the number of differing coordinates between two sets, used for text and non-numerical data.
- Minkowski Distance: A generalization of Euclidean and Manhattan distances.

  $D\_M(x, y) = (\sum\_\{i = 1\}\text{^}\{n\} |x\_i - y\_i|\text{^}p)\text{^}\{1/p\}$

- Maximum Distance: The maximum absolute difference between coordinates of a pair of points.

  $max|ai - bi|$

- Canberra Distance: Used to compare ranked lists

  $\Sigma \frac{|xi - yi|}{|xi + yi|}$

- Manhattan Distance: The sum of the absolute differences of their coordinates, suitable for grid-based distance calculations.

  $\Sigma |a\_i - b\_i|$

Linkage Criteria

- Single Linkage (Nearest Linkage): The minimum distance between a pair of data points in two different clusters.
- Complete Linkage (Farthest Linkage): The maximum distance between a pair of data points in two different clusters.
- Average Linkage: The average distance between all pairs of points in two different clusters.
- Centroid Linkage: The distance between the centroids (geometric mean) of two clusters.
- Ward's Linkage: Minimizes the increase in the sum of squared errors within clusters at each iteration

Keywords

- ❖ Distance Metric
- ❖ Linkage Criteria
- ❖ Dendrogram
- ❖ Cluster Algorithm
- ❖ Hierarchical
- ❖ Cluster Similarity

- Category 3: Principal Components

BASICS

The principal components (PCs) are the orthonormal directions in a dataset which represent the most variance of the dataset. They are sorted so that the first PC direction represents the direction with most of the variance, the second PC represents the direction capturing most of the variance after the first one, and so on. PCs directions correspond to the eigenvectors of the covariance matrix from the data subject to study.

PCs are used in the Principal Component Analysis (PCA) technique for dimensionality reduction, which is part of unsupervised learning. The key idea is that since only a few PCs can capture most of the variance of the data (a variance threshold should be defined, to select the correct amount of PCs that can retrieve this total variance), these PCs can be used to represent the original data with fewer variables (the selected number of PCs) than the original features. Transformation between PCs and the original features is done through a linear relationship.

More details about the working of PCA is given in the next sections.

KEYWORDS

- ❖ Dimensionality reduction
- ❖ Loading factors
- ❖ Covariance
- ❖ Eigenvalues
- ❖ Eigenvectors
- ❖ Eigenproblem

Step 2

Category 1: Lasso Regression

- Advantages: Lasso regression has several advantages in the context of linear regression models. It includes feature selection, prevents overfitting, and handles multicollinearity. It is a great tool to create predictive models that are both accurate and interpretable, particularly when dealing with large numbers of predictors.
- Computation: For example, a multi-variable regression model is analyzed and modeled in this case. The dependent variable is DXY which is the U.S. Dollar Index daily return. Independent variables are the Gold Index daily return, Silver Index daily return, 13-week Treasury Bills daily return, 10-year Treasury Bond Yield daily return, and EURUSD daily return. Two years of data are collected in this case, one year for the training, and the other for the testing.
- Disadvantages: As the Lasso regression is a supervised machine learning, the performance depends on the data selection. In this case, one of its disadvantages is variable selection bias. It tends to select only one variable from a group of highly correlated variables, then, the rest of them are ignored. In addition, it is also sensible to the sample size. When the dataset has a small sample size, the performance of LASSO regressions will also be limited. Moreover, as it only has an L1 penalty function, it has parameter sensitivity and bias in coefficient as its disadvantages.
- Equations: the LASSO regression is actually an extension to the linear regression or the OLS regression (Ordinary Least squares regression). This equation can be referred as follow:

  $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_n X_n$$

  As stated in the basics section, the LASSO regression is a supervised penalty regression that has a penalty function. Its penalty function is an L1 penalty function which can be referred to as follows:

  $$f() = \sum_{j=1}^{p} |\beta_j| = ||\beta||_1$$

  This equation is used for the coefficient adjustment as it adjusts the parameters for different coefficients.

- Features: One of the significant features is the LASSO regression is capable of dealing with multi-variable cases, it can flexibly adjust the coefficient of variables to ensure the performance of regression models. With the presence of a penalty function, it can shrink the coefficients of variables that don't contribute a lot.
- Guide: LASSO regression has simple inputs and outputs. The inputs are independent and dependent variables, and a regression model will be output accordingly.
- Hyperparameters: there is a hyperparameter which is the regularization parameter λ that requires tuning. This parameter determines the amount of shrinkage applied to the coefficients. A larger value of λ increases the penalty, leading to more coefficients being shrunk towards zero, which can result in a simpler model with fewer predictors.
- Illustration: In the computation case, U.S. Dollar Index daily return is tried to be modelled by multiple independent variables including the Gold Index daily return, Silver Index daily return, 13-week Treasury Bills daily return, 10-year Treasury Bond Yield daily return, and

EURUSD daily return. By using the LASSO regression, it can model the curve of the U.S. Dollar Index daily return, however, it seems to have trouble in modeling the local extremes. Based on the R-squared test, the adjusted R-squared value, for this case, is 0.39, which indicates the model can interpret the somehow of data, but not very well.

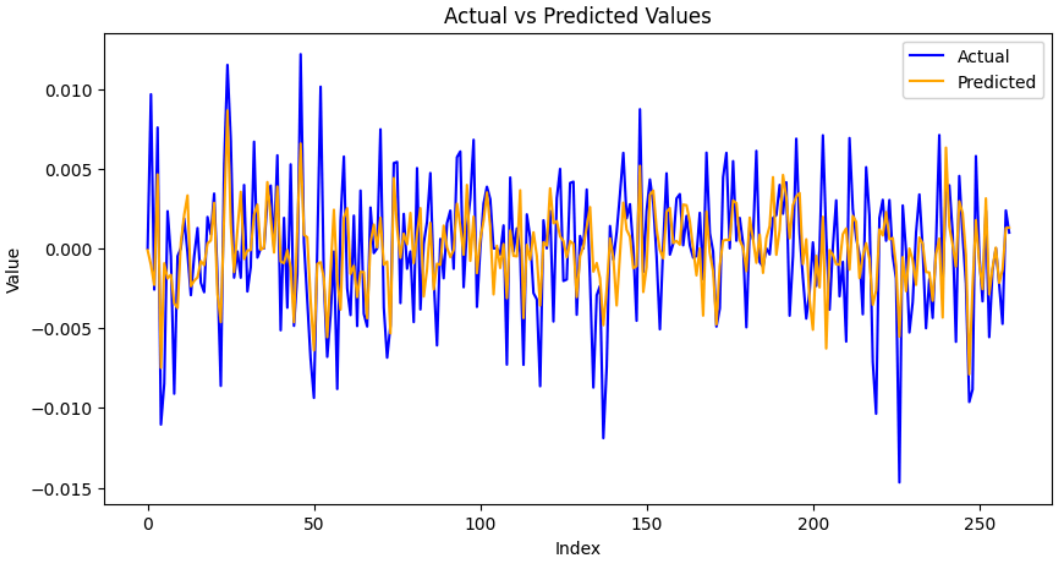


Figure: Year 2023 US Dollar Index daily return

Actual Performance vs. Predicted Performance

Based on the figure above, the curve activities can be captured by the model, however, the model's failure to deal with large values may be due to the presence of a penalty function.

- Journal: Fox, J., & Weisberg, S. (2002). Robust regression. An R and S-Plus companion to applied regression, 91, 6.

Category 2: Hierarchical Clustering

- **Advantages:**
  - ❖ No Need to Specify the Number of Clusters: Compared to k-means clustering, hierarchical clustering does not require the number of clusters to be specified beforehand under exploratory analysis. It enables researchers and analysts to observe the dendrogram and decide the number of clusters that best fit their data which can be argued that it provides flexibility in this instance.
  - ❖ Hierarchical structure: This is seen through nested clusters that provide insight into how the data points are grouped together at various levels, which can be useful for identifying sub-clusters within larger ones.

- Good visualization: Through the dendrogram, a comprehensive visualization of the clustering process is provided showing the distances at which clusters are combined and the sequence of merges. Aside from helping in understanding the data structure it also helps to know the number of clusters. The dendrogram helps analysts with their interactions as they explore many clustering solutions.

- **Disadvantages:**
  - Sensitive to outliers: Because of its sensitivity to outliers, formation of clusters is affected and this leads to the formation of less meaningful clusters. It does not have a built-in way of handling these outliers unlike some other clustering methods.
  - Difficulty in choosing the right linkage and Distance Metric: The choice of linkage criteria (e.g., single, complete, average, Ward's) can greatly influence the results. Different criteria can lead to different cluster structures, and it may not be clear which one is the best for a given dataset.

    Also, the choice of distance metric (e.g., Euclidean, Manhattan) affects the clustering outcome. Selecting the appropriate metric often requires experimentation.

  - Challenges during interpretation: For large datasets, dendrograms can become very complex and difficult to interpret. This complexity makes it challenging to extract meaningful insights from the hierarchical structure.

- **Equations**

  Measures of Distance
  - Euclidean Distance: $sqrt((x2 - x1)\text{^}2 + (y2 - y1)\text{^}2)$
  - Minkowski Distance: $D\_M(x, y) = (\sum\_\{i = 1\}\text{^}\{n\} |x\_i - y\_i|\text{^}p)\text{^}\{1/p\}$
  - Maximum Distance: $|a - b|\infty = max|ai - bi|$
  - Canberra Distance: $\Sigma \frac{|xi - yi|}{|xi + yi|}$
  - Manhattan Distance: $\Sigma |a\_i - b\_i|$

Linkage Criterion

- Single linkage: $d(u,v) = min(dist(u[i], v[j]))$
- Complete Linkage: $d(u,v) = max(dist(u[i], v[j]))$
- Average Linkage: $d(u,v) = \Sigma \frac{d(u[i], v[j])}{(|u| * |v|)}$
- Centroid Linkage: $d(u,v) = ||cu - cv||_2$

Features:

- Hierarchical clustering is able to handle various data types such as numerical and categorical data types.
- It does not require specifying the number of clusters beforehand unlike K-means clustering
- It works well with small and medium datasets
- Hierarchical clustering are robust to outliers.

Guide:

Inputs include the *Dataset, distance metric* and *linkage criteria*.

For instance, consider the dataset to be the daily closing prices of stocks like Apple (AAPL), Microsoft (MSFT), Google (GOOGL), and Amazon (AMZN) over the past year. With the distance metric,Euclidean distance computes the straight-line distance between two data points in the multi-dimensional space defined by the dataset. Correlation coefficient measures the strength and direction of a linear relationship between two variables. For example, using correlation coefficient, the similarity between the daily price movements of Apple and Microsoft can be quantified. With the linkage criteria, Complete linkage, which calculates the maximum distance between pairs of stocks in different clusters can be used in our example. The distance between clusters is determined by the maximum distance between any pair of stocks from each cluster.

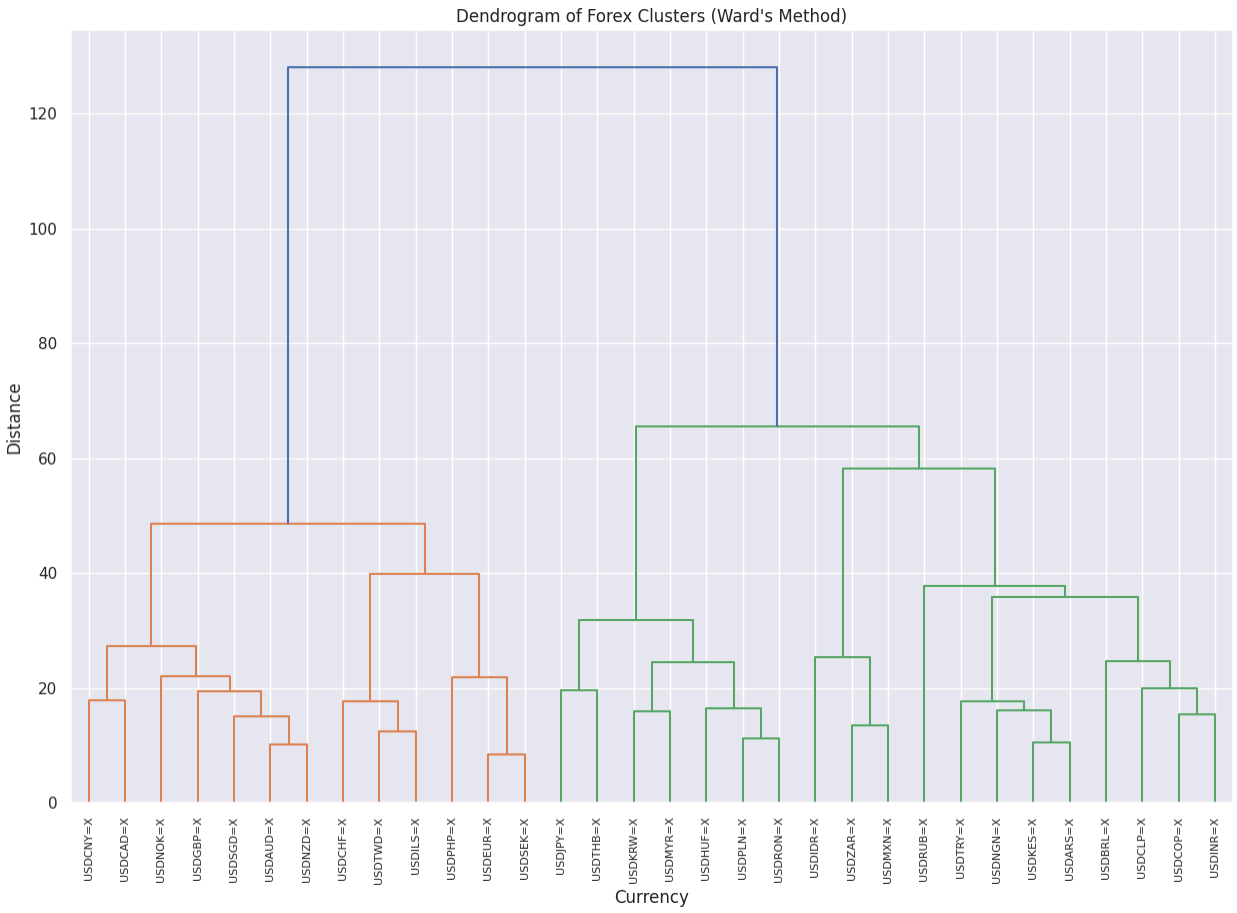
Outputs include Dendrogram, Cluster Assignment and Hierarchical Clusters.

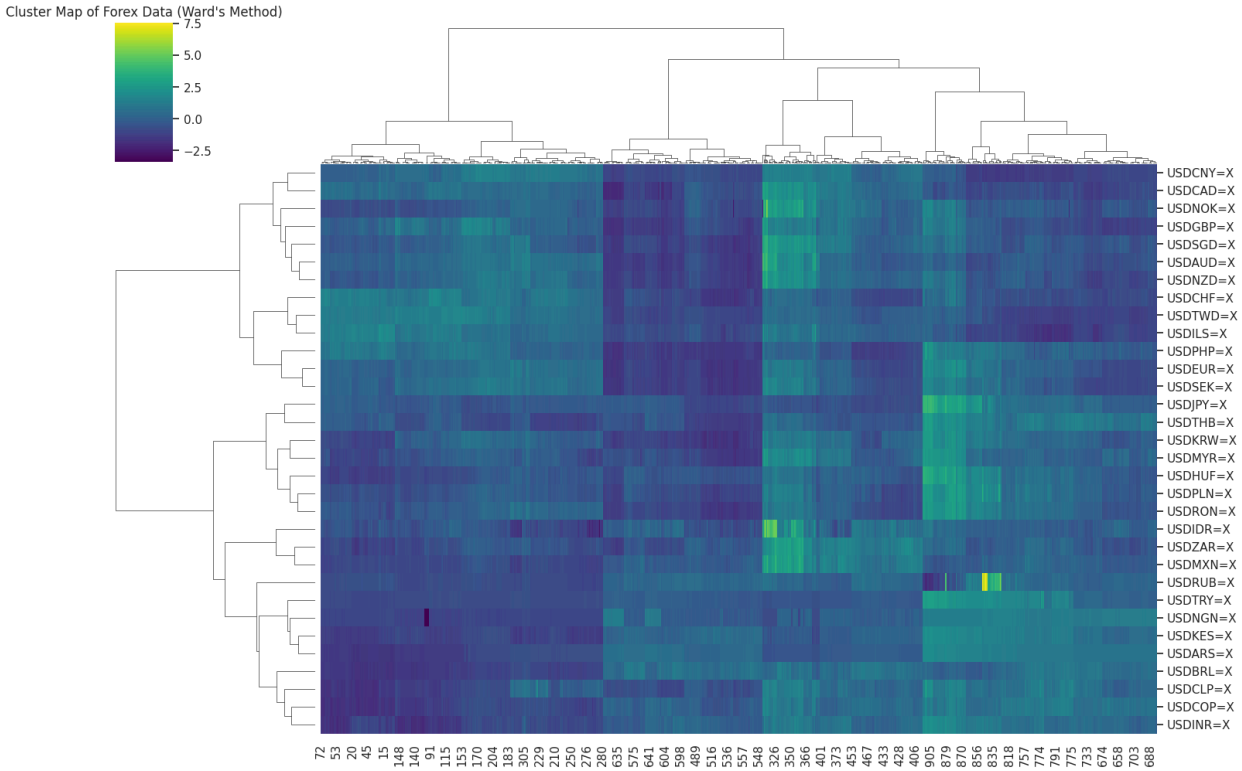
For instance, the dendrogram, which is a tree-like diagram that represents the hierarchical structure of the data can how clusters of technology stocks (e.g., AAPL, MSFT, GOOGL) and retail stocks (e.g., AMZN, Walmart) based on their historical price correlations. Under the cluster assignment, each stock will be assigned a cluster based on similarities or features. In our example, AAPL and MSFT might be assigned to a "technology" cluster, while AMZN and Walmart could be in a "retail" cluster.

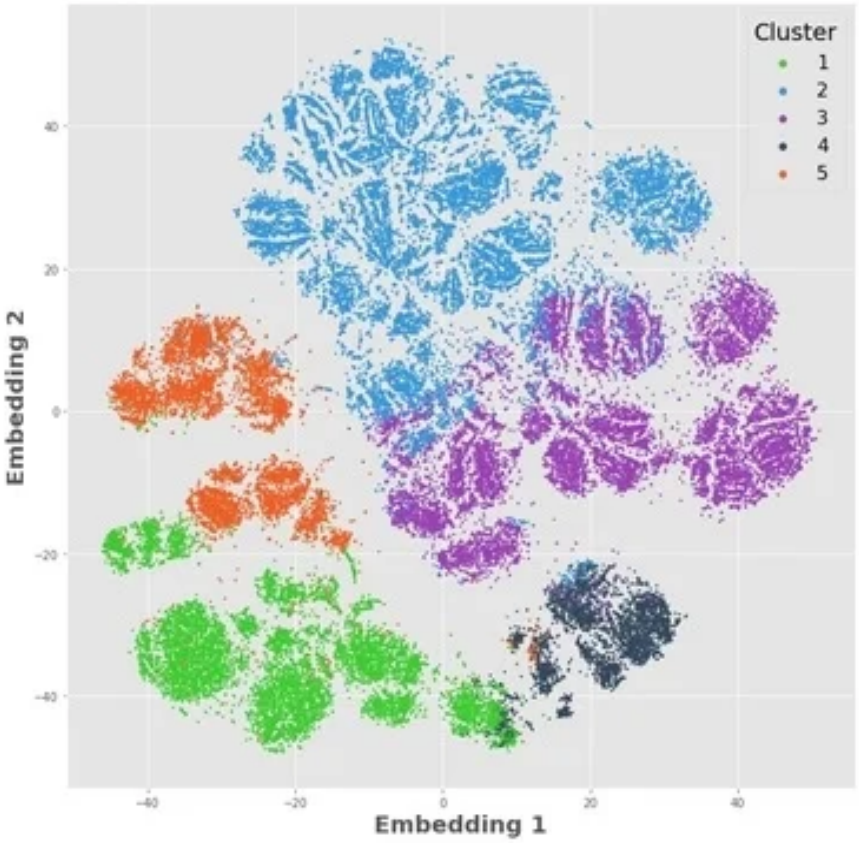
- **Hyperparemeters:**
  - Distance Metric
  - Linkage Criteria
  - Number of Clusters
  - Distance threshold

- Merging threshold

- **Illustration:**




Cluster Map of Forex Data (Ward's Method)
7.5
5.0
2.5
0.0
−2.5
USDCNY=X
USDCAD=X
USDNOK=X
USDGBP=X
USDSGD=X
USDAUD=X
USDNZD=X
USDCHF=X
USDTWD=X
USDILS=X
USDPHP=X
USDEUR=X
USDSEK=X
USDJPY=X
USDTHB=X
USDKRW=X
USDMYR=X
USDHUF=X
USDPLN=X
USDRON=X
USDIDR=X
USDZAR=X
USDMXN=X
USDRUB=X
USDTRY=X
USDNGN=X
USDKES=X
USDARS=X
USDBRL=X
USDCLP=X
USDCOP=X
USDINR=X
72 53 20 45 15 148 140 91 115 153 170 204 183 305 229 210 250 276 280 635 575 641 604 598 489 516 536 557 548 326 350 366 401 373 453 467 433 428 406 905 879 870 856 835 818 757 774 791 775 733 674 658 703 688


Cluster
1
2
3
4
5
Embedding 2
40
20
0
−20
−40
−40
−20
0
20
40
Embedding 1

- **Journal:** Thompson, John R.J., et al. "Know Your Clients' Behaviours: A Cluster Analysis of Financial Transactions." *Journal of Risk and Financial Management*, vol. 14, no. 2, 2021, p. 50. Available online at: Know Your Clients' Behaviours: A Cluster Analysis of Financial Transactions.

Category 3: principal components (PCs)

- Advantages: the main advantage of using PCs is the dimensionality reduction, as the number of variables representing a problem can be simplified to only a few PCs that can capture most of the variance of the problem
- Computation: code can be found in the 'Step 2' section of the associated jupyter notebook. PCs are applied to the daily returns of several stocks from the SP500 index.
- Disadvantages: the main difficulty with PC lies on how to interpret them, as they might not be related a priori with the variables of study. Relations linking the PCs and the problem variables should be found, but this might be problematic and, in some cases, even impossible. Also, it is sensitive to outliers.
- Equations: Given data which is represented by a set of features **X** = {X_1, X_2, … X_p}, which have previously been standardized and centered, principal components can be defined by finding the linear relations between the features and some variables (the principal components) which maximize the amount of variance. For example, the principal component 1 could be related as follows:

$$Y_1 = w_{11}X_1 + w_{21}X_2 + \ldots + w_{p1}X_{1p}$$

where w_{ij} represents the different weights, or loadings. In this case, the constraint on the loadings $\sum_{i=1}^{p} w_{i1}^2$ must hold. In matrix form, the linear relation could be represented as:

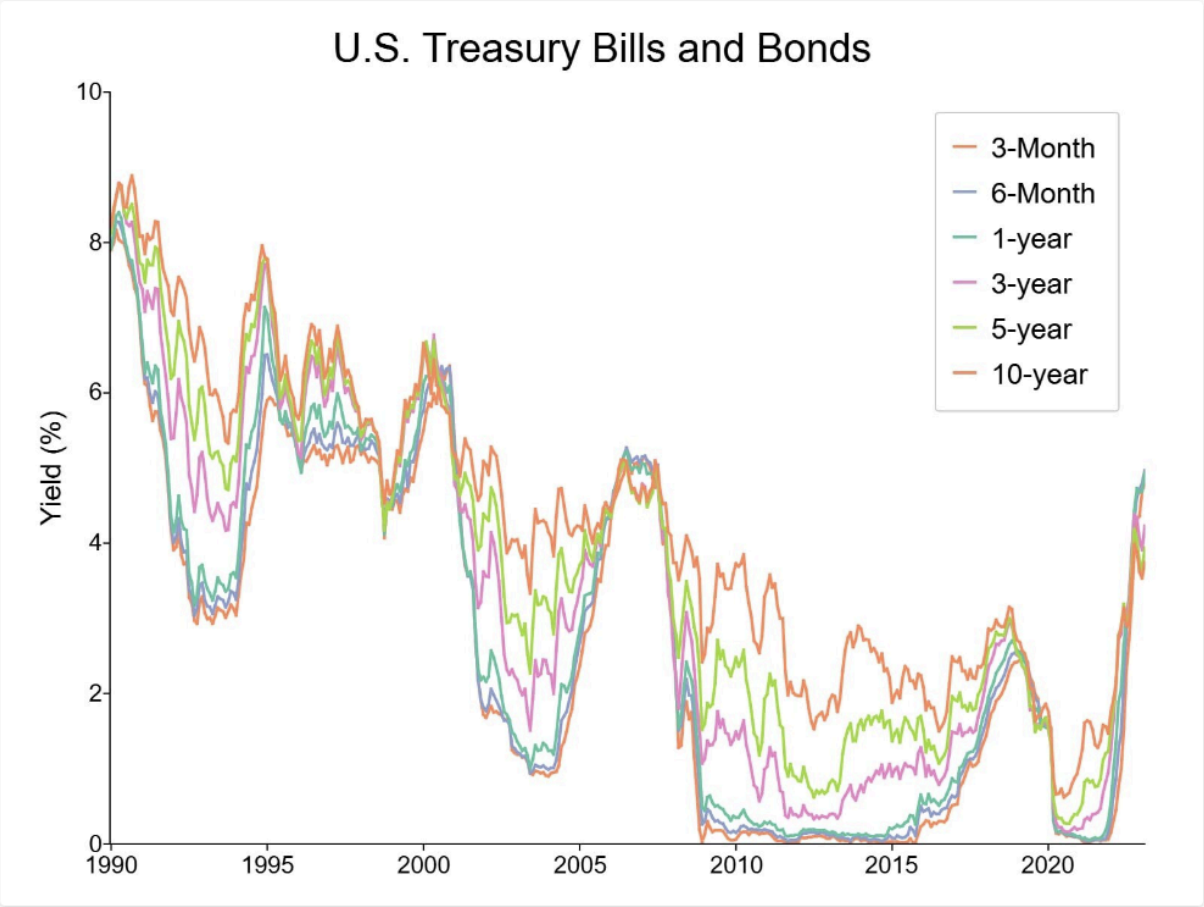
Y = X W

with **W** being the matrix containing the loadings, and **Y** is the matrix containing the principal components. The rows of matrix **W** are eigenvectors of the covariance matrix of the data.

- Features: This model works well with large quantities of data, since it allows to retrieve only a few components (directions, given by the principal components) that can represent most of the variance of the problem. Its computational cost is also acceptable, since it assumes linearity on the variables. This can also be a disadvantage, as PCA would be unable to capture non-linearity in data.
- Guide:
  - Inputs: feature data X, and variance threshold. For example, a few time series of several stock options, or yield curves of interest rates at several maturities
  - Outputs: the principal components, loading matrix, and variances by each component. The number of principal components to be retrieved (this is

dimensionality reduction) will be defined by the variance threshold (which is an input): the N of PCs chosen will be the ones that provide, at least, the total amount of variance chosen.

- Hyperparameters:
  - Number of components to keep (if none, all components will be kept)
  - Algorithm mode (for performing SVD)
- Illustration: Let's consider the following yield curve for US treasury bills and bonds (example taken from this webpage), at different time horizons:



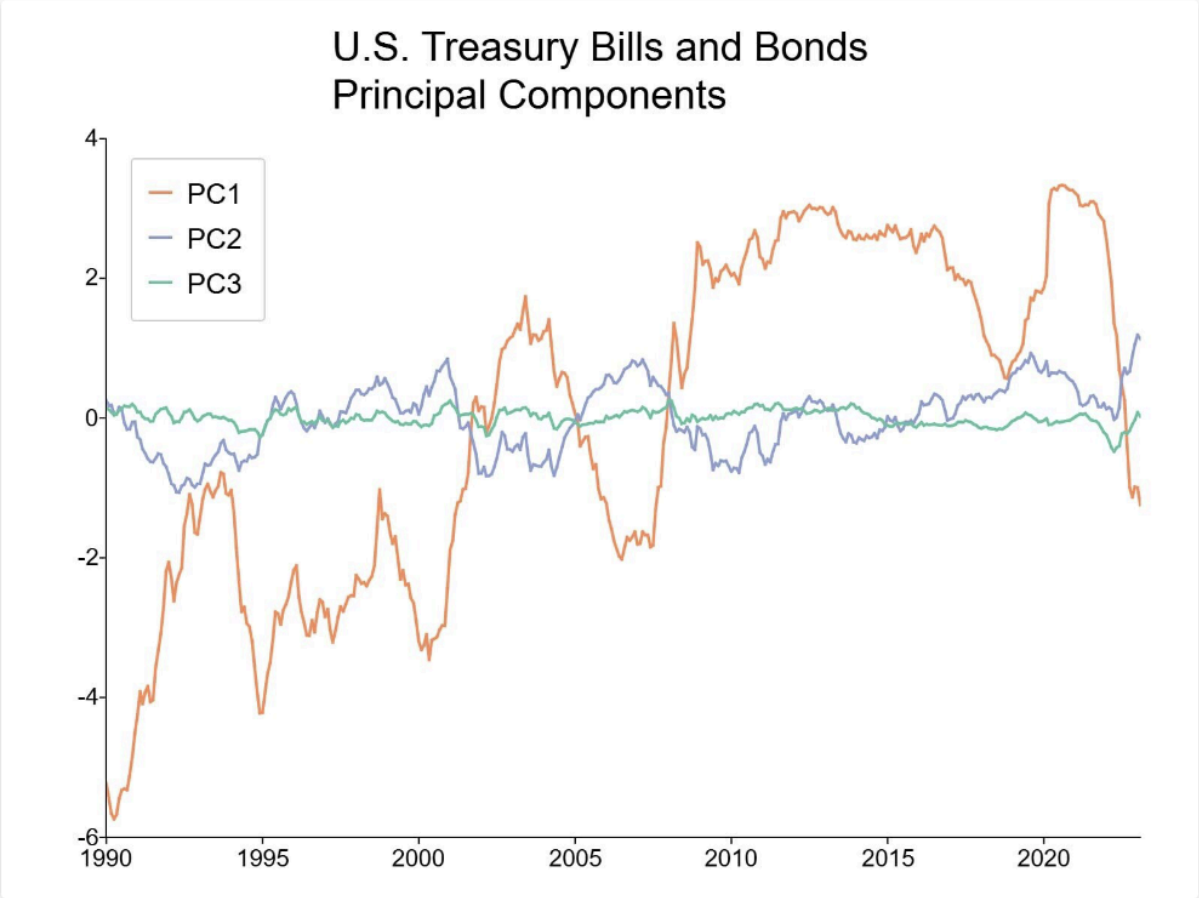
PCA can be applied to these data. First, the yield curves are standardized, to have zero mean and unit standard deviation. Then, PCA is applied with 6 components (that is, only up to 6 PCs are retrieved). The distribution of total variance among these PCs is as follows:

Component	Proportion Of Variance	Cumulative Proportion
PC1	0.960	0.960
PC2	0.038	0.997
PC3	0.002	1.000
PC4	0.000	1.000
PC5	0.000	1.000
PC6	0.000	1.000

So as observed, only the first component captures 96 % of the total variance, and the the first 3 components contain all the variance of the data. The loading matrix for this case would look as follows:

Principal components	PC1	PC2	PC3	PC4	PC5	PC6
GS3M	-0.4079	0.4111	0.4863	-0.5416	0.3029	0.2076
GS6M	-0.4094	0.3883	0.1535	0.2221	-0.5448	-0.5585
GS1	-0.4122	0.2970	-0.2404	0.6120	0.1557	0.5342
GS3	-0.4154	-0.0855	-0.5911	-0.1926	0.4744	-0.4567
GS5	-0.4102	-0.3607	-0.2806	-0.3932	-0.5725	0.3750
GS10	-0.3939	-0.6742	0.5040	0.3020	0.1856	-0.1024

Using these loadings, the data can then be transformed, and the 3 first PCs can be represented:



- Journal: Yu, H. Chen, R. Zhang, G., 2014. 'A SVM Stock Selection Model within PCA'. Procedia Computer Science, 31, pp .406-412. Available online.

Step 3: Technical section

- Lasso Regression

  With regularization, which controls the amount of shrinkage applied to the coefficients, can be tuned to find the optimal balance between the model's complexity and its balance since a larger value of λ increases the penalty leading to more coefficients being shrunk to zero.

- Hierarchical Clustering

  The choice of distance metric which helps to know the distance between the data point can affect the shape and tightness of the clusters. For instance, When clustering financial time series data, choosing the Euclidean distance might be appropriate for data with continuous variables.

  Therefore, there is a need to experiment with different distance metrics to see which one yields the most meaningful clusters. This can be done using cross-validation or silhouette analysis.

  The choice of the linkage criteria which helps compute the distance between the clusters during the merging process can result in different cluster shapes and structures. For example, using complete linkage might result in more compact clusters, which is useful when clustering stocks to form distinct market segments.

  Therefore different linkage criteria need to be tested or tried. Thereafter, an evaluation should be made on the dendrograms to determine the best linkage method for the data.

  The number of clusters can affect the granularity if the clustering results. To determine the optimal number of clusters, the silhouette score can help.

  Analyzing the dendrogram can help find the level at which clusters are formed.

- Principal Components

  The number of components to keep is very important for reducing dimensionality and simultaneously retaining most of the data's variance. For instance, in a dataset with 100 features, retaining 10 might capture a large percentage of the variance, probably about, 90 - 95%, significantly reducing dimensionality.

  With SVD, different algorithms can offer trade-offs between efficiency and accuracy. For instance, the use of the 'randomized' algorithm can be faster for large datasets with many features.

Step 4: Marketing Alpha

Three different machine learning strategies have been analyzed in this report: Lasso regression, hierarchical clustering and the obtention of Principal Components (PCs), which are the base of the Principal Component Analysis (PCA) technique for dimensionality reduction. The basics of the methods and their main advantages and drawbacks have been discussed. Some numerical examples have also been included, to show their possible application.

Even though application to more complex and realistic scenarios is still missing, the discussions and results shown in this report present an optimistic view of the applicability and usefulness of machine learning. Lasso regression, for example, has been shown to be a proper tool for supervised learning while being computationally affordable. Also, the combination of a linear regression with a L1 norm is simple to understand, which makes it a more interpretable algorithm as opposed to other machine learning tools, which act more like black boxes and hinder their understanding. The example of application (USD daily return) has shown that, for this case, the regression can properly retrieve the trend of the dollar. Even though it misses quantitative accuracy (due to the low R2 obtained), the qualitative behavior shows its capabilities.

The second method studied, hierarchical clustering, is an example of unsupervised clustering. Without the need of preselecting the number of clusters beforehand, this technique provides groups of data that can be used to understand relations and find patterns. Information is often presented in a graphical and easily interpretable way, such as tree structures and dendrograms. The algorithms and relations employed are also easily interpretable, such as the distance and linkage functions to be chosen. Finding clusters in the data through hierarchical clustering can then allow to differentiate among groups that can later be tackled particularly (for instance, groups of stocks, or currencies with similar characteristics). The relations found by hierarchical clustering might be difficult to be obtained *at first sight* (i.e. without applying algorithms), making it a powerful tool.

Finally, the foundations of PCs have been stated and applied to the returns of a selection of 9 stocks from the SP500 index. Even PCs being less interpretable than the previously discussed algorithms, it has been shown how they can recover most of the variance of the dataset with fewer variables than the original data. Relation between the PCs and the original data is linear, though the loadings of the PCs. Furthermore, PCs can also provide information on trends within the data. Further work could focus on using PCA for portfolio optimization.

All these examples have shown different algorithms of machine learning and their applications. It can therefore be concluded that machine learning tools are able to tackle different problems in a computationally efficient way, making it promising to be applied in financial applications.

Step 5: Learn more

According to the reference, the strength of machine learning techniques can be following:

- Capability of handling large and complex data sets: ML algorithms can process and learn from large volumes of data, making them suitable for big data applications. In addition, they can uncover the patterns or relationships within complex data sets that may be difficult to detect using traditional statistical methods.
- Automation: ML algorithms can automate processes to adjust their parameters and models to learn from the data, and make valid predictions accordingly.
- Flexibility: This portfolio includes several ML algorithms and techniques such as regression, classification, and principal components. Different cases can apply to different ML algorithms, and algorithms can generally fit the data.
- Accuracy: compared with non-ML algorithms, ML algorithms are more capable of interpreting large-size datasets, and they can interpret them much better. They can achieve high levels of precision in tasks such as image recognition, natural language processing, and speech recognition.
- Feature selection: ML algorithms can select features of data, or adjust features in the data, such as principal components and robust regressions.
- Nonlinear relationship capabilities: Huanhuan Yu, Rongda Chen, and Guoping Zhang (2014) stated that PCA, as an ML algorithm, can handle nonlinear classification, while Fox and Weisberg (2002) also mentioned that robust regressions can handle better in non-linear regression.

In conclusion, ML algorithms and techniques have better performance in data modeling and data handling. Also, ML models can be complicated to capture more features from the data or more flexible to find the data patterns. ML algorithms and techniques are much better to handle the real-world applications.