

DATA MINING (WQD7005)

Assignment 3

Instructor: Profesor Madya Dr. Teh Ying Wah

Group Members:

|  |  |
| --- | --- |
| **Name** | **Matric No.** |
| Choy Siew Wearn | WQD180053 |
| Teng Lung Yun | WQD180024 |
| Wo Choy Yun | WQD180054 |
| Xu Xiang | WQD180061 |
| Zhu Ting | WQD180077 |

**1.0 Introduction**

Stock analysis is an essential step for investors to search for potential profitable stocks and analyse stocks. As such, analysts can perform Principal Component Analysis (PCA) and algorithms such as Symbolic Aggregate Approximation (SAX) and Piecewise Aggregate Approximation (PAA). Both Symbolic Aggregate Approximation (SAX) and Piecewise Aggregate Approximation (PAA) are time series representation techniques that can be used with optimization algorithms to help analysts to identify hidden and relevant patterns in financial time series data.

**1.1 Principal Component Analysis And Covariance**

Principal Component Analysis (PCA) was invented by Karl Pearson in 1901. PCA uses orthogonal transformation to transform a set of data with possible correlated variables into linearly uncorrelated variables which is known as the principal components. In most cases, PCA is used an a exploratory tool for data analysis to make prediction models. In this assignment, PCA is done through covariance matrix where the results will be shown later in section 2.1.

**1.2 Symbolic Aggregate Approximation & Piecewise Aggregate Approximation**

Both Symbolic Aggregate Approximation (SAX) and Piecewise Aggregate Approximation (PAA) algorithms are invented by Eamonn Keogh and Jessica Lin. The main function of PAA is to reduce the dimensionality of the time-series by dividing the time-series into equal segments, then computing each of them by averaging the values in the segment. As for SAX, it is an extended based approach of PAA where it transforms the original time-series into PAA representation, then converts the PAA data into a string.

**2.0 Data Preprocessing**

In this assignment, we are using two dataset. The first dataset is ss.csv from previous milestone and the second dataset are crawled data for 13 days.

First dataset are be used to calculate PCA and covariance between attributes. Second datasets are used to calculate SAX and PAA, because second dataset is time-series data.

For the second dataset, 13 days of data are compiled into one file using R. In order for us to perform SAX and PAA, data preprocessing is required such as cleaning data. For data preprocessing, we did:

1. Created a sum column at the right known as “Total” to sum all the stock values
2. Remove rows with sum value “0.000”
3. Replace NA with “0.000” in dataset
4. Change all columns attribute of those containing values to numeric

Figure 2.0.1 below shows the dataset that has already been cleaned or preprocessed. In the first column shows the list of companies, where from columns named “1” to “13” contain the stock values for each company from day 1 to day 13. The column named “Total” is the sum of all stock values from day 1 to day 13.

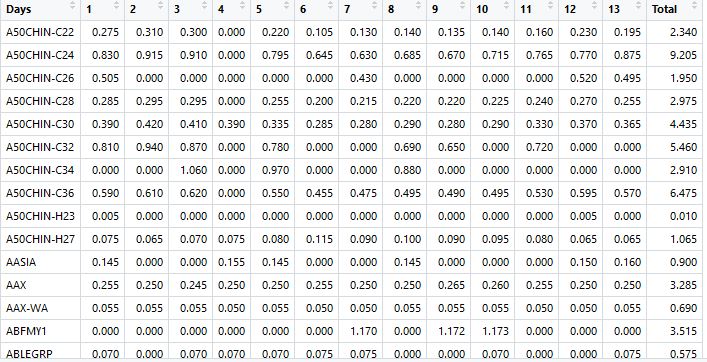


Figure 2.0.1: Second dataset Preprocessed Stock Data

**2.1 Covariance**

In this part of assignment, PCA and Covariance among attributes is done using Python. Using the first dataset, the results from the Python code determining covariance among attributes will be shown in Figure 2.1.1 below:

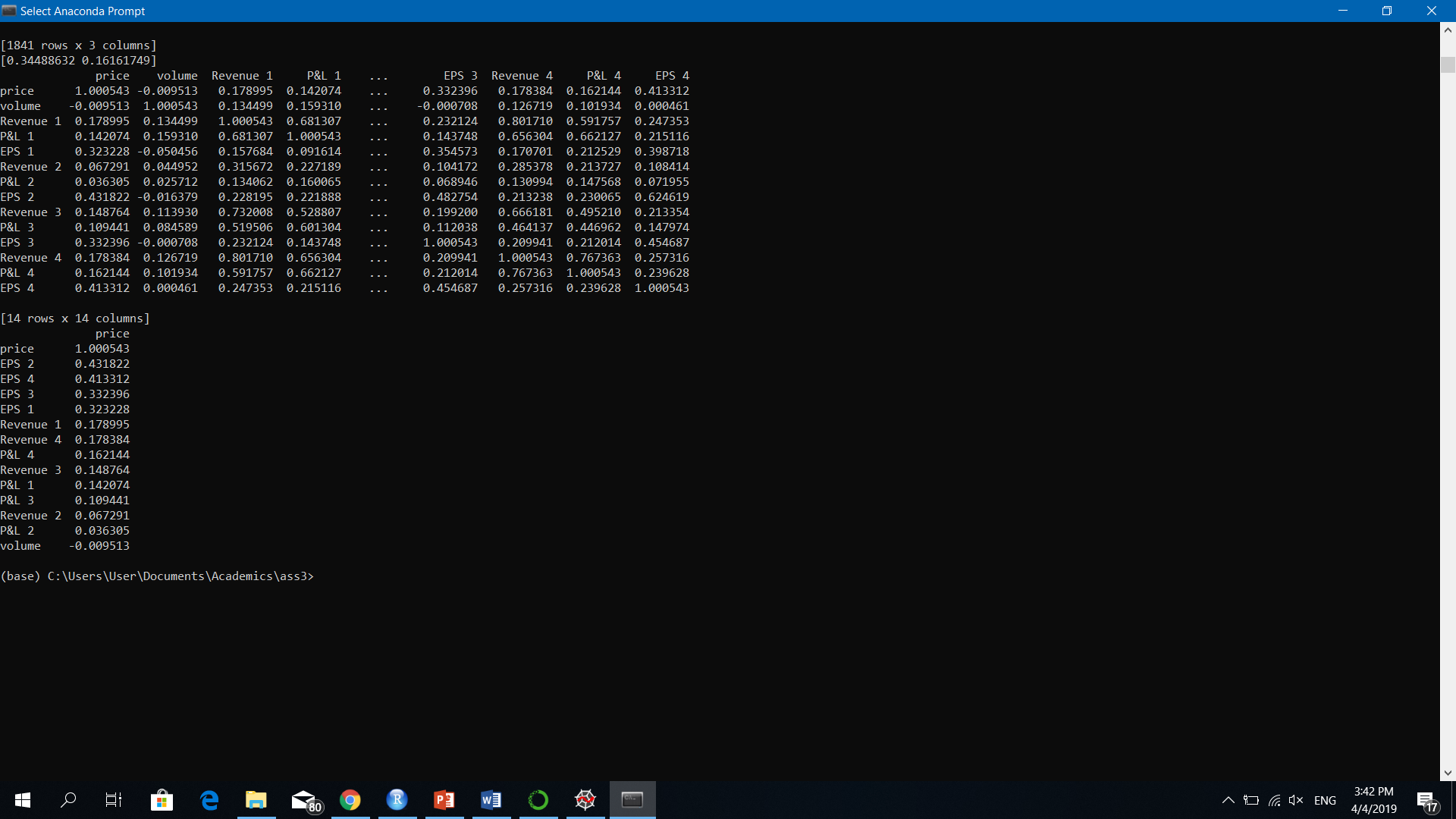


Figure 2.1.1: Covariance Results Among Attributes Using Python

Covariance calculations are used to find relationships between dimensions in high dimensional data sets where visualization is difficult. Figure 2.1.1 shows a covariance matrix between the target variable price and other variables and a reduced covariance matrix. The reduced covariance matrix shows that EPS 2, has the highest covariance in relative to price, followed by EPS 4, EPS 3 and EPS 1, etc. This shows that the price tends to be affected by EPS 2. Positive value of covariance indicates both dimensions increase or decrease together. A negative value indicates while one increases the other decreases, or vice-versa. The reduced covariance matrix shows all positive covariance values in relative to price except ‘volume’ variable.

The second dataset preprocessed data is transposed to below figure:



Figure 2.1.2: Stock with price for 13 days

Using data in Figure 2.1.2, covariance among stock price is determined using R. The results are shown in Figure 2.1.3 below.

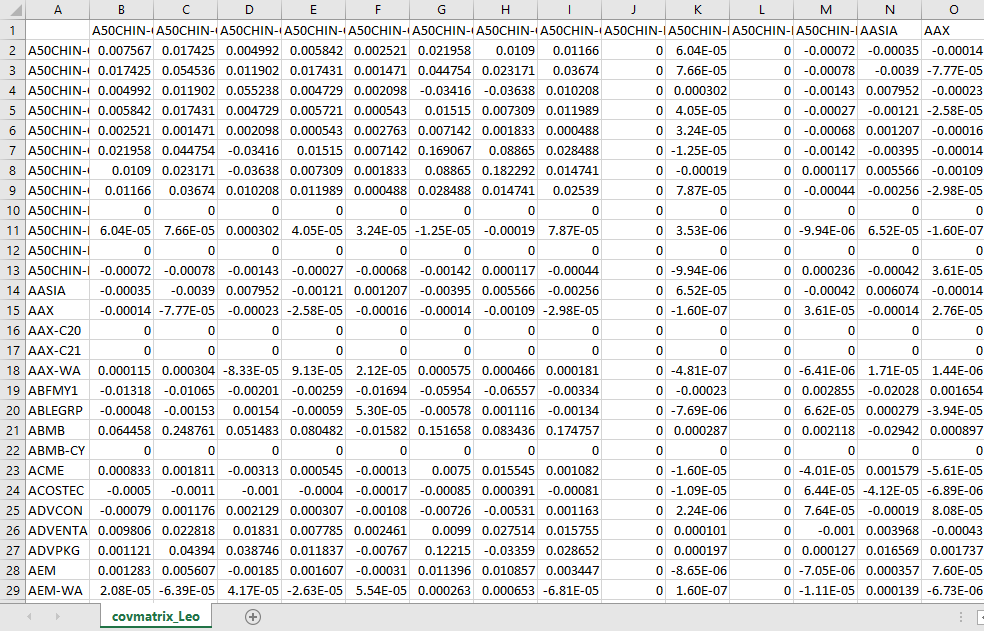


Figure 2.1.3: Stock Price Covariance

The readings above indicate the extent to which the volatility in prices move together. These readings are shown in covmatrix\_Leo.csv.

A variance-covariance matrix is particularly useful when it comes to analysing the volatility between elements of a group of data. For instance, a variance-covariance matrix has particular applications when it comes to analysing portfolio returns.

If several assets with a high covariance are included in a portfolio, then this represents high risk. This indicates that several assets with high volatility move together, which is what investors would typically want to avoid.

On the other hand, selecting assets that show negative covariance allows for greater diversification of the portfolio, since the volatility of the assets do not tend to move together.

**2.2 PCA**

PCA is a dimensionality reduction algorithm where new features are created which represents the original feature dimensions in a lower dimension with a little loss of the total information.

PCA is a linear transformation that chooses a new coordinate system for the dataset such that greatest variance by an y projection of the data set comes to lie on the first axis(then called the first principal component), the second greatest variance on the second axis, and so on.

The dataset used for PCA is the first dataset – ss.csv. The Features for PCA calculations for this assignment are Volume, Revenue 1, P&L 1, EPS 1, Revenue 2, P&L 2, EPS 2, Revenue 3, P&L 3, EPS 3, Revenue 4, P&L 4, EPS 4.

Volume refers to buy and sales volume of the day.

Revenue, Profit and Loss(P&L) and Earning per Share(EPS) are the quarter results for each month for previous consecutive 4 months.

Target variable is the price of the stocks.

The features are reduced to 2 dimensions as below Figure 2.2.1

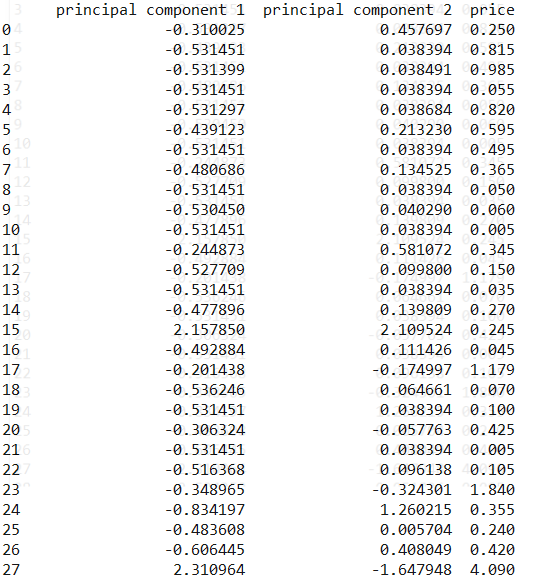


Figure 2.2.1: Principal Component Values with the Target Variable – Closing Price

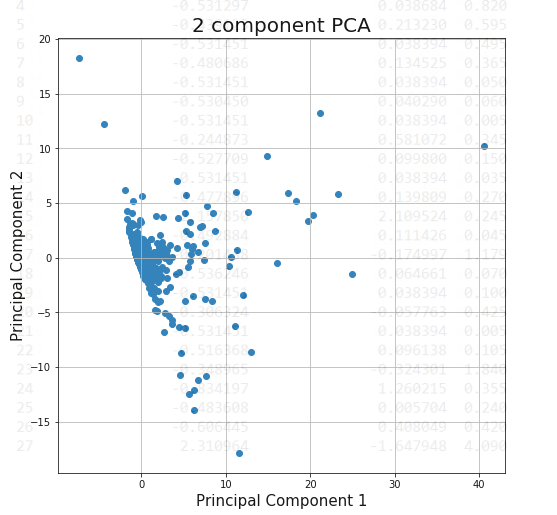


Figure 2.2.2: Graph of Principal Components

**2.2 Work Flow On Performing SAX and PAA**

In this assignment, SAX and PAA algorithms is performed using R. The steps and results will be shown below:

1. Perform Euclidean value to determine the distance relationship

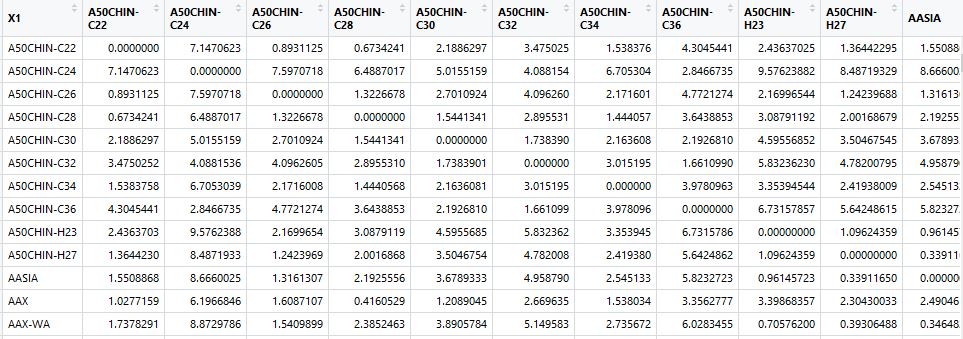


Figure 2.2.1: Stock Data with Euclidean Value Using R

1. Execute Z-Normalization on the data to reduce the value gap between multiple lines

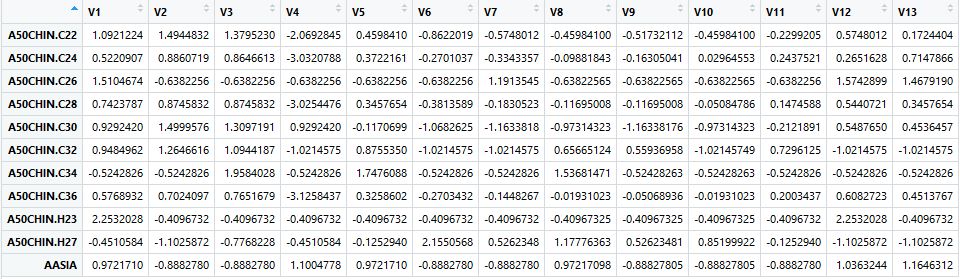


Figure 2.2.2: Stock Data With Z-Normalization Using R

1. Carry out PAA algorithm on the dataset and we will get the result as follows:

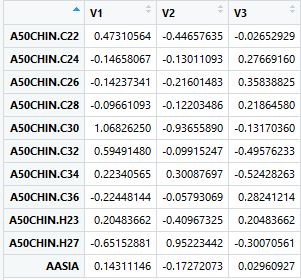


Figure 2.2.3: Stock Data With PAA Algorithm Results Performed Using R

**2.3 Visualisation**

Results from PAA is visualised as follows:

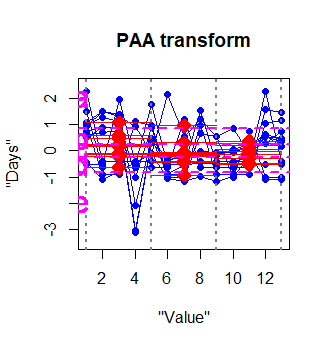


Figure 2.3.1: PAA Graph From R

Note that the PAA figure does not shows all data due to large amount of unique stock

**3.0 Discussion**

During the execution of SAX and PAA, we also did the PAA path output, which will be shown in Figure 4.1 below:

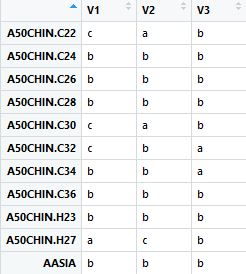
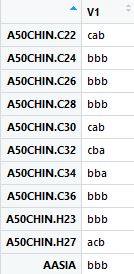


Figure 4.0.1: PAA Path Output Using R

Then SAX is the difference value, where SAX will transform the discrete variables into strings. An example of how to determine the SAX distance in this assignment will be shown below.

Determine the SAX distance between AASIA and A50CHIN:

1. SAX transform strings through 3 PAA points for AASIA : b b b, same goes to A50CHIN
2. SAX distance is calculated by adding the the gap distance between each PAA points in the y-axis as shown in Figure 4.2, C30 = 0.5 + 0.3 +0.3 = 0.5 where C30 is the gap distance.

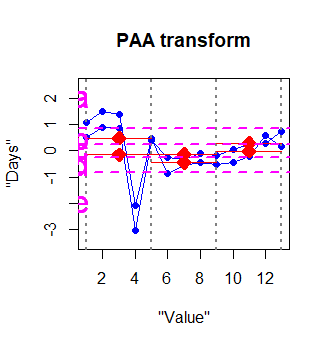


Figure 4.0.2: Visualised PAA Graph Using R

**4.0 Contribution**

|  |  |
| --- | --- |
| Group Member | Contribution |
| Teng Lung Yun | - Crawl data for 13 days, do PCA code and covariance |
| Choy Siew Wearn | - Clean data, do PAA and SAX, provide brief report of PAA and  SAX |
| Wo Choy Yun | - Compile, elaborate and completing full report, run through PAA,  PCA and covariance code |
| Zhu Ting | - Assist in report completion |
| Xu Xiang | - Assist in report completion |

**5.0 Reference**

1. Symbolic Aggregate Approximation. Retrieved from: <https://jmotif.github.io/sax-vsm_site/morea/algorithm/SAX.html>
2. Piecewise Aggregate Approximation of time series. Retrieved from: <https://jmotif.github.io/sax-vsm_site/morea/algorithm/PAA.html>
3. Symbolic Aggregate Approximation. Retrieved from: <http://www.cs.ucr.edu/~eamonn/SAX.htm>
4. A Beginner's Guide To Eigenvectors, Eigenvalues, Pca, Covariance and Entropy

Retrieved from: <https://skymind.ai/wiki/eigenvector>

1. Principal Component Analysis

Retrieved from:<http://www.cse.psu.edu/~rtc12/CSE586/lectures/pcaLectureShort_6pp.pdf>

1. Variance-Covariance Matrix: Stock Price Analysis in R (corpcor, covmat)

Retrieved from: <https://www.michael-grogan.com/variance-covariance-matrix-calculation-r/>