Financial Signal Processing Coursework

Tze Hei Tang
01221240

April 20, 2020

Contents

1 Regression Model **4**

1.1 Processing stock price data in Python 4

1.1.1 Time series data 4

1.1.2 Sliding window based first and second order statistics 4

1.1.3 log return time series 5

1.1.4 Suitability of log returns over simple returns 5

1.1.5 Log returns example 5

1.1.6 Shortcomings of log returns over simple returns 6

1.2 ARMA vs ARIMA Models for Financial Applications 6

1.2.1 ARMA vs ARIMA for S&P 500 7

1.2.2 Fitting ARMA(1,0) Model to the S&P 500 index time-series 7

1.2.3 Fitting ARIMA(1,1,0) Model to the S&P 500 index time-series 8

1.2.4 Necessity of taking the log of the prices for ARIMA analysis 8

1.3 Vector Autoregressive (VAR) Models 8

1.3.1 Concise representation of VAR 8

1.3.2 Optimal coefficients for B 9

1.3.3 VAR(p) stability 9

1.3.4 VAR(1) Model 10

1.3.5 Stocks in the same sector 13

2 Bond Pricing **14**

2.1 Examples of bond pricing 14

2.1.1 Compounding 14

2.1.2 2 15

2.1.3 3 15

2.2 Forward Rates 15

2.3 Duration of a coupon-bearing bond 15

2.4 Capital Asset Pricing Model (CAPM) and Arbitrage Pricing Theory (APT) 16

3 Portfolio Optimization **18**

3.1 Adaptive minimum-variance portfolio optimization 18

4 Robust Statistics and Non Linear Methods **20**

4.1 Data Import and Exploratory Data Analysis 20

4.2 Robust Estimators 28

4.3 Robust and OLS regression 29

4.4 Robust Trading Strategies 31

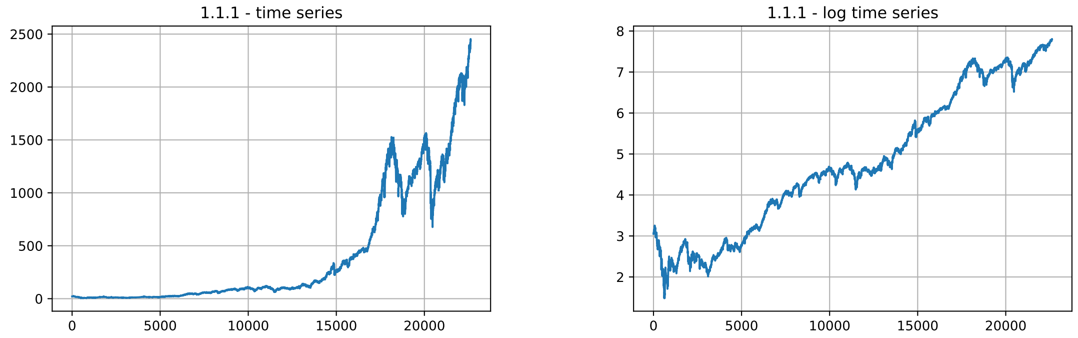
5 Graphs in Finance **32**

References **36**

1 Regression Model

1.1 Processing stock price data in Python

1.1.1 Time series data

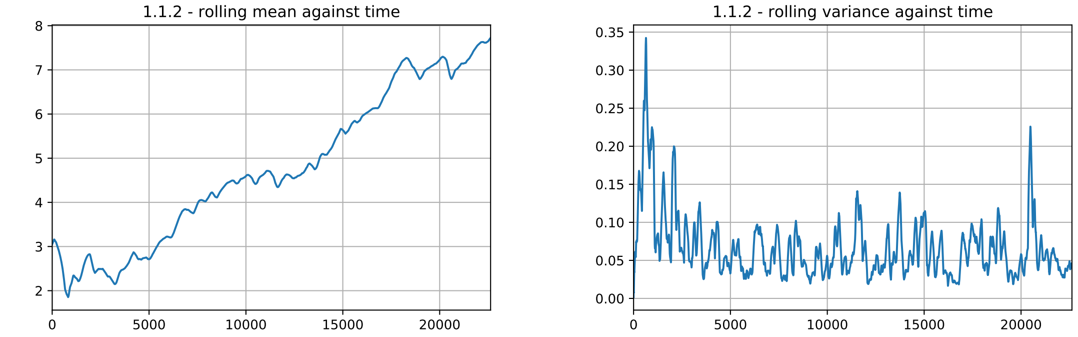


(a) time series

(b) log time series

Figure 1: 111

1.1.2 Sliding window based first and second order statistics



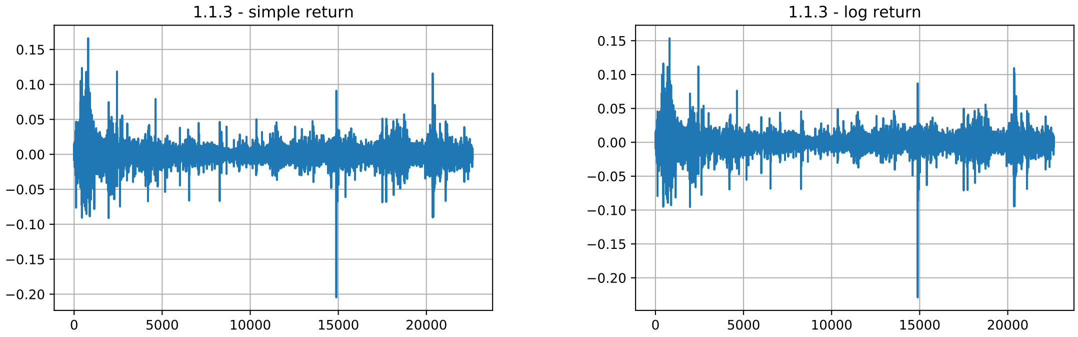
(a) Rolling mean

(b) Rolling variance

Figure 2: 112

The price time-series is not stationary, as shown in figure(2), the rolling mean is increasing against time.

1.1.3 log return time series



(a) Simple return

(b) Log return

Figure 3: 113

The mean and variance of the log and simple returns fluctuate much less than the S&P 500 time series share price in section 1.1.1. By converting the time series data from the price to return, the data is much more stationary, since return measures the change of price in time, it removes the constant increase in mean of the data.

1.1.4 Suitability of log returns over simple returns

By applying the Jarque-Bera test on the log and simple return, we obtain zero for both of the test results. The Jarque-Bera test tests whether the sample data has the skewness and kurtosis matching a normal distribution [1], and returns a p-value corresponding to the likeliness of the log and simple returns to have Gaussianity. Since the p-value of tests are zero, we can conclude that the log and simple returns time series does not have Gaussianity.

Log returns are also considered suitable due to the Log-normality of prices. Over short periods of time, prices are distributed log-normally.

1.1.5 Log returns example

Simple return is given by:

$$r_t = \frac{p_t}{p_{t-1}} - 1$$

Log return is given by:

$$\rho_t = log(\frac{p_t}{p_{t-1}})$$

We can calculate the returns over the following example:
"You purchase a stock for £1. The next day its value goes up to £2 and the following day back to £1."

t (day)	**0**	**1**	**2**
Share price (£)	1	2	1
Simple return	Null	1	-0.5
Log return	Null	log(2) = 0.30	log(1/2)=-0.30

Table 1: Simple and log return of example 1.1.5

The logarithmic return is more preferable over the simple returns on the basis of this example. Since overall over the entire period the share price has no net change, the net return should also be zero to represent that. The simple return gives a net return of $1 + (-0.5) = 0.5$, while the log return gives a net return of $0.30 + (-0.30) = 0$, the log return is a more accurate representation in this example, this is known as the time additivity property.

1.1.6 Shortcomings of log returns over simple returns

It is argued that log-return should not be used in the financial context, due to the fact that actual market returns is negatively skewed (large amount of decrease in times of panic). Whereas the log-returns over a longer term than daily is problematic because it is skewed positively. Simple return will not have this issue. Log-return is more suitable for shorter period of time such as 1-day returns.

1.2 ARMA vs ARIMA Models for Financial Applications

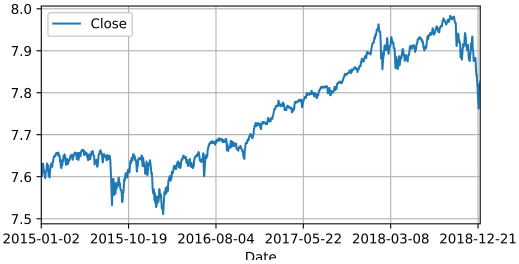


Figure 4: SP 500 index

1.2.1 ARMA vs ARIMA for S&P 500

Observing the S&P 500 index time-series plotting in figure(4), it is obvious that the index is non-stationary from the increasing trend. We can further validate this statement by using the Dickey–Fuller test (adfuller) to test for stationarity, the Null hypothesis of Dickey–Fuller test is that there is a unit root in the AR model fitted to the data. We obtain the p-value $p = 0.668$, hence, by having a significant level of 0.05, we cannot reject the null hypothesis and that we conclude that the data is non-stationary.

The use of ARIMA model will be preferred over ARMA, since ARIMA will be able to remove that non-stationarity in the time-series data through the use of differencing of raw observations.

1.2.2 Fitting ARMA(1,0) Model to the S&P 500 index time-series

Fitting the ARMA(1,0) model to the S&P 500 index, we can plot the prediction and the true signal, given in fig(5).

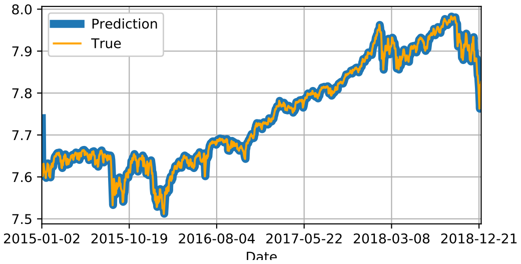


Figure 5: Modelling S&P 500 using ARMA(1,0)

We can see that the ARMA(1,0) model predicted the SP 500 signal fairly well, with the exception of the first data point, this is because the ARMA(1,0) only considers the previous data point, which is not available for predicting the first data point.

The SNR of the true signal against the error signal is 58.4 dB, computed via the SNR formula given by:

$$SNR = 10\log_{10}\left(\left(\frac{A_{Signal}}{A_{Error}}\right)^2\right)$$

Where A_{signal} is the root mean square of the signal.

The model parameters of the ARMA(1,0) model returned has the values [const] = 7.740 and [ar.L1.True] = 0.9974. Despite the good prediction performance, it is not very useful in practice, where the prediction can only be made not far from the current time frame, it does not translate well into real world applications.

1.2.3 Fitting ARIMA(1,1,0) Model to the S&P 500 index time-series

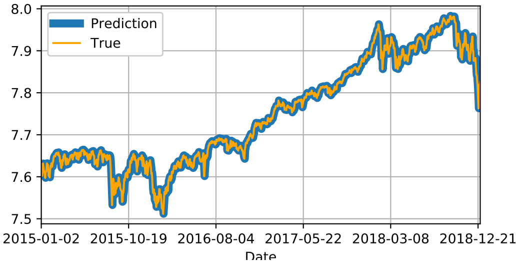


Figure 6: Modelling S&P 500 using ARIMA(1,1,0)

The model parameters of the ARMA(1,0) model returned has the values [const] = 0.000196 and [ar.L1.D.True] = −0.008752. The SNR of the true signal against the error signal is 59.1 dB, computed via the same method above. We can conclude that the ARIMA(1,1,0) model is better than the ARMA(1,0) model due to it having a better SNR with the error signal. ARIMA(1,1,0) is more physically meaningful compared to ARMA(1,0) since it also considers all of its past data.

1.2.4 Necessity of taking the log of the prices for ARIMA analysis

Taking the log of the prices removes stationarity of the data, while also compressing the range of the data, both of these factors combined can give a better model fitting and better parameters.

1.3 Vector Autoregressive (VAR) Models

1.3.1 Concise representation of VAR

The VAR(p) process given by

$$\mathbf{y_t} = \mathbf{c} + \mathbf{A_1 y_{t-1}} + \mathbf{A_2 y_{t-2}} + \cdots + \mathbf{A_p y_{t-p}} + \mathbf{e_t}$$

where the expanded form is given by

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \\ \vdots \\ y_{k,t} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_k \end{bmatrix} + \begin{bmatrix} a^1_{1,1} & a^1_{1,2} & \cdots & a^1_{1,k} \\ a^2_{2,1} & a^2_{2,2} & \cdots & a^1_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ a^1_{k,1} & a^1_{k,2} & \cdots & a^1_{k,k} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \\ \vdots \\ y_{k,t-1} \end{bmatrix} + \cdots + \begin{bmatrix} a^p_{1,1} & a^p_{1,2} & \cdots & a^p_{1,k} \\ a^p_{2,1} & a^p_{2,2} & \cdots & a^{p_1}_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ a^p_{k,1} & a^p_{k,2} & \cdots & a^p_{k,k} \end{bmatrix} \begin{bmatrix} y_{1,t-p} \\ y_{2,t-p} \\ \vdots \\ y_{k,t-p} \end{bmatrix} + \begin{bmatrix} e_{1,t} \\ e_{2,t} \\ \vdots \\ e_{k,t} \end{bmatrix}$$

1.3.2 Optimal coefficients for B

We try to get the form $\mathbf{Y} = \mathbf{BZ} + \mathbf{U}$ from the expanded form above, by letting

$$\mathbf{B} = \begin{bmatrix} \mathbf{c} & \mathbf{A_1} & \mathbf{A_2} & \dots & \mathbf{A_p} \end{bmatrix}$$

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}[p] & \mathbf{y}[p+1] & \cdots & \mathbf{y}[T] \end{bmatrix}$$

we can deduce that

$$\mathbf{Z} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \mathbf{y}[p-1] & \mathbf{y}[p] & \cdots & \mathbf{y}[T-1] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{y}[0] & v[1] & \cdots & \mathbf{y}[T-p] \end{bmatrix}$$

$$\mathbf{U} = \begin{bmatrix} \mathbf{e}[p] & \mathbf{e}[p+1] & \cdots & \mathbf{e}[T] \end{bmatrix}$$

through least-squares, we can estimate that the optimal set of parameters by assuming LS minimises the error vector to zero. Hence, we set the vector $\mathbf{U} = 0$

$$\mathbf{Y} = \mathbf{B_{opt}Z} + 0$$

because $\mathbf{Z}$ is not a square matrix

$$\begin{aligned} \mathbf{YZ^T} &= \mathbf{B_{opt}ZZ^T} \\ \mathbf{YZ^T}(\mathbf{ZZ^T})^{\mathbf{-1}} &= \mathbf{B_{opt}} \\ \mathbf{B_{opt}} &= \mathbf{YZ^T}(\mathbf{ZZ^T})^{\mathbf{-1}} \end{aligned}$$

1.3.3 VAR(p) stability

The VAR(1) process is given as

$$\mathbf{y}_t = \mathbf{A}\mathbf{y}_{t-1} + \mathbf{e}_t \tag{1}$$

with the previous time instant written as

$$\mathbf{y}_{t-1} = \mathbf{A}\mathbf{y}_{t-2} + \mathbf{e}_{t-1} \tag{2}$$

subbing (1) into (2)

$$\mathbf{y}_t = \mathbf{A}^2\mathbf{y}_{t-2} + \mathbf{e}_t + \mathbf{A}\mathbf{e}_{t-1} \tag{3}$$

$$\mathbf{y}_t = \mathbf{A}^k\mathbf{y}_{t-k} + \sum_{i=0}^{k-1} \mathbf{A}^i\mathbf{e}_{t-i} \tag{4}$$

We can see from eq(4), any shock from time instant k can have an effect on the present time, and the value of k extends to infinity. Therefore, for the VAR(1) model to be stable from any shock in the infinite past, we must require that the past shocks to have a decaying factor. Assuming that all of the eigenvalues of A have absolute value less than 1, the sum $\sum_{i=0}^{\infty} \mathbf{A}^i \mathbf{e}_{t-i}$ converges, and therefore is stable.

1.3.4 VAR(1) Model

The stocks with tickers [CAG, MAR, LIN, HCP, MAT] are selected.

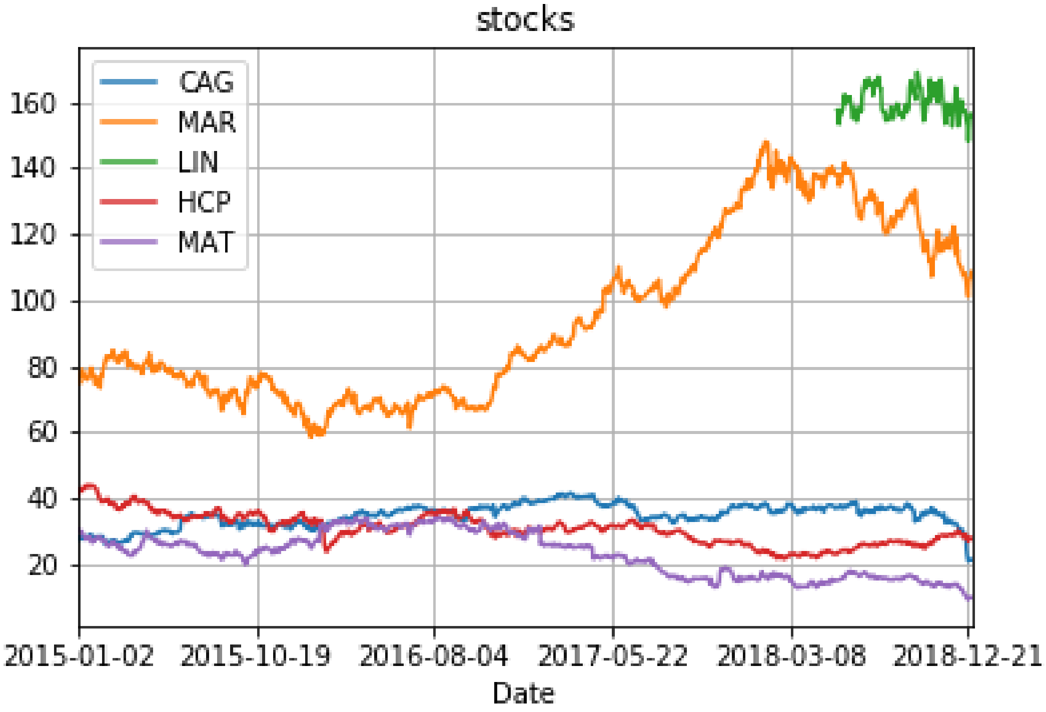


Figure 7: Stock prices

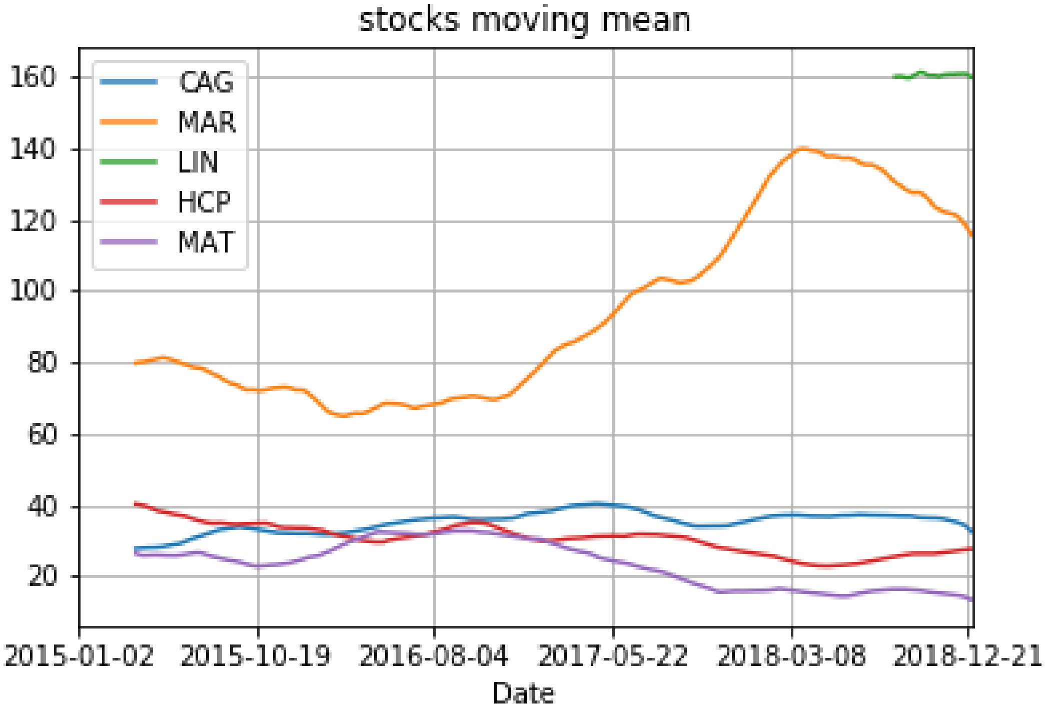


Figure 8: Stock price moving mean

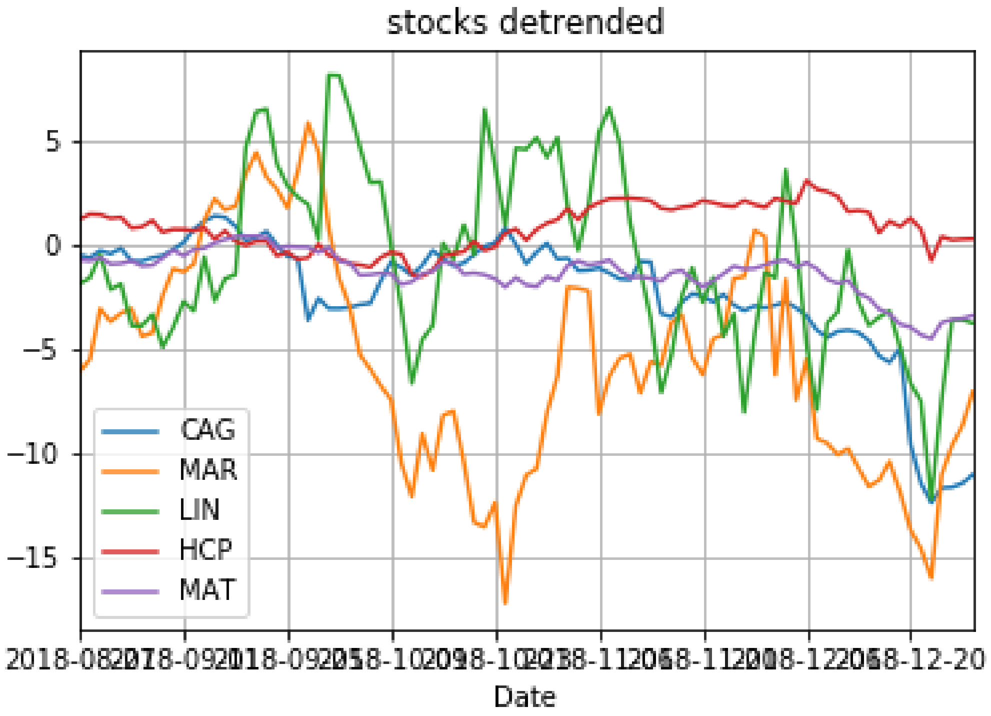


Figure 9: Stock price detrended

	eigA
CAG	0.72609393
MAR	0.72609393
LIN	1.00635964
HCP	0.86051894
MAT	0.91144512

Table 2: Stock eigenvalues

From table(2), LIN should not be chosen to be used to construct a portfolio, since its eigenvalue is slightly greater than 1, it is not considered stable as of the reasoning mentioned in part 1.3.3.

1.3.5 Stocks in the same sector

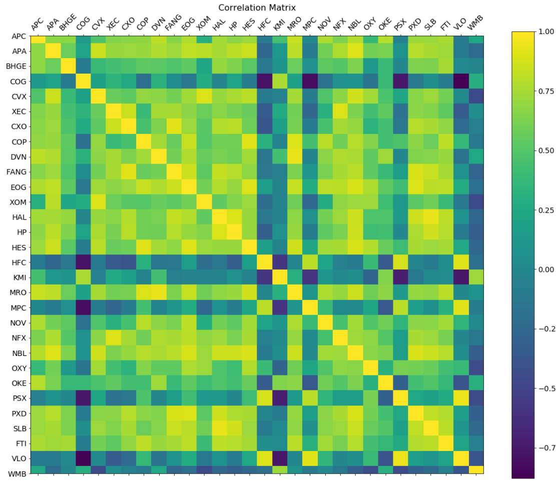


Figure 10: correlation matrix of the stocks in the GICS sector

The stocks are mostly positively correlated, so it is not advised to construct a portfolio entirely of stocks from the same sector, since it is best to diversify the portfolio with different types of stocks to reduce risk.

2 Bond Pricing

2.1 Examples of bond pricing

2.1.1 Compounding

The formula for calculating the return is given by:

$$
\begin{aligned}
FV &= PV(1+\frac{r}{m})^{nm} \\
\frac{FV}{PV} &= (1+\frac{r}{m})^{nm} \\
\frac{1}{nm}\ln\frac{FV}{PV} &= \ln(1+\frac{r}{m}) \\
1+\frac{r}{m} &= \exp\left\{\frac{1}{nm}\ln\frac{FV}{PV}\right\} \\
r &= m\Bigg\{\exp\left\{\frac{1}{nm}\ln\frac{FV}{PV}\right\}-1\Bigg\}
\end{aligned}
$$

where n = number of years, m = number of times compounded, r = annual interest rate. In all of the cases below, n = 1, PV = 1000, FV = 1100, and we require to find the value of r

a) annual compounding: m = 1

$$
\begin{aligned}
r &= 1\Bigg\{\exp\left\{\ln\frac{1100}{1000}\right\}-1\Bigg\} \\
r &= 10\%
\end{aligned}
$$

b) semiannual compounding: m = 2

$$
\begin{aligned}
r &= 2\Bigg\{\exp\left\{\frac{1}{2}\ln\frac{1100}{1000}\right\}-1\Bigg\} \\
r &= 9.762\%
\end{aligned}
$$

c) monthly compounding: m = 12

$$
\begin{aligned}
r &= 12\Bigg\{\exp\left\{\frac{1}{12}\ln\frac{1100}{1000}\right\}-1\Bigg\} \\
r &= 9.57\%
\end{aligned}
$$

d) continuous compounding

$$
\begin{aligned}
r &= \ln\{1.1\} \\
r &= 9.53\%
\end{aligned}
$$

2.1.2 2

$$
\begin{aligned}
ae^r &= a\big[1+\frac{0.15}{12}\big]^{12}\\
r &= \ln\big[1+\frac{0.15}{12}\big]^{12}\\
r &= 14.91\%
\end{aligned}
$$

2.1.3 3

$$
\begin{aligned}
ae^{0.12} &= a\big[1+\frac{r}{4}\big]^4\\
\frac{0.12}{4} &= \ln\big[1+\frac{r}{4}\big]\\
r &= 4\Big[\exp\big\{\frac{0.12}{4}\big\}-1\Big]\\
r &= 12.18\%
\end{aligned}
$$

an interest of $10000 \times \frac{12.18\%}{4} = \$304.5$ will be paid out in each quarter.

2.2 Forward Rates

A I would not take the 9% strategy. If I invested for a year, there is an opportunity that I can get a better return option in the second year by chance.

B The two route of investment strategies (5% → 7% or 9%) gives the same return. This is because of the principle of no arbitrage.

C The 9% forward rate option has the disadvantage of the interest rate in the second year being actually higher than what the forward rate predicted, hence the investor can potentially lose out by chance. But it has the advantage of protecting the investor against the interest rate dropping to a much lower level, this is a good choice if the investor is satisfied with the forward rate.

D ?

2.3 Duration of a coupon-bearing bond

A

$$Duration = \sum_t \frac{t \times PV(C_t)}{PV}$$

Using the table given, the Duration is calculated to be

$$
\begin{aligned}
&0.0124 + 0.0236 + 0.0337 + 0.0428 + 0.0510 + 0.0583 + 6.5377\\
=&6.7595
\end{aligned}
$$

B

$$
\begin{aligned}
Duration_{Modified} &= \frac{duration}{1 + yield} \\
&= \frac{6.7595}{1.05} \\
&= 6.4376
\end{aligned}
$$

C The modified duration is a measure of how much the the price of the bond can be changed with respect to changes in yield, ie. the measure of the sensitivity of the price of the bond. If the modified duration of a bond is high, then it may not be a good idea to purchase that bond.

2.4 Capital Asset Pricing Model (CAPM) and Arbitrage Pricing Theory (APT)

1. Estimate market returns

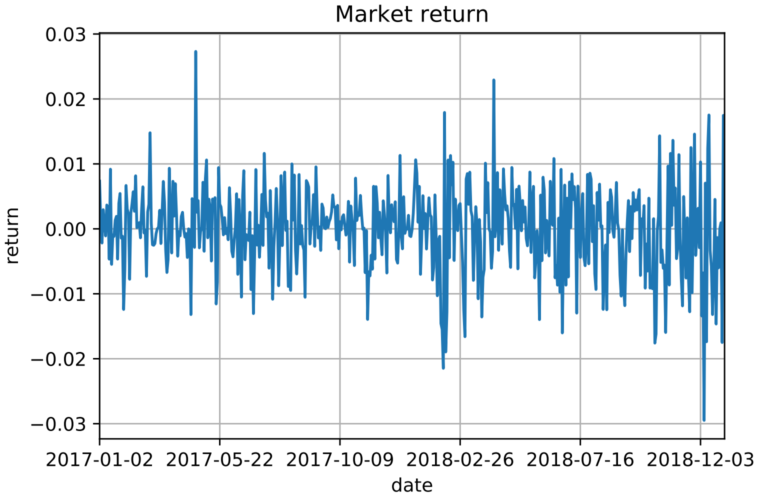


Figure 11: Market returns per day

2. Estimate rolling $\beta_{i,t}$ with rolling window of 22 days
   From CAPM, the Beta of a stock is given by [2]:

$$\beta = \frac{Cov[R_i, R_m]}{Var[R_m]}$$

where R_i denotes that return of the stock, R_m denotes the return of the market

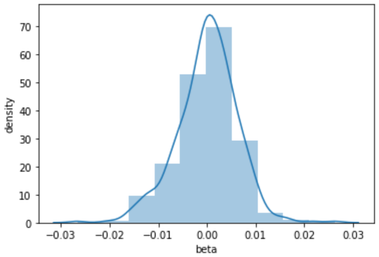


Figure 12: Rolling beta

Figure(12) shows the mean β values of all stocks of the 157 European companies given in the data. (use hist_fit) The histogram gives a distribution of μ 0. Most of the companies are between 0.5 and 1.5 times as volatile as the market return.

3. Estimate cap-weighted market return

   The cap-weighted market return given by

$$R_m = \text{ret(market)} = \sum_i \frac{\text{mcap}_i \times \text{ret}_i}{\sum_i \text{mcap}_i} \tag{5}$$

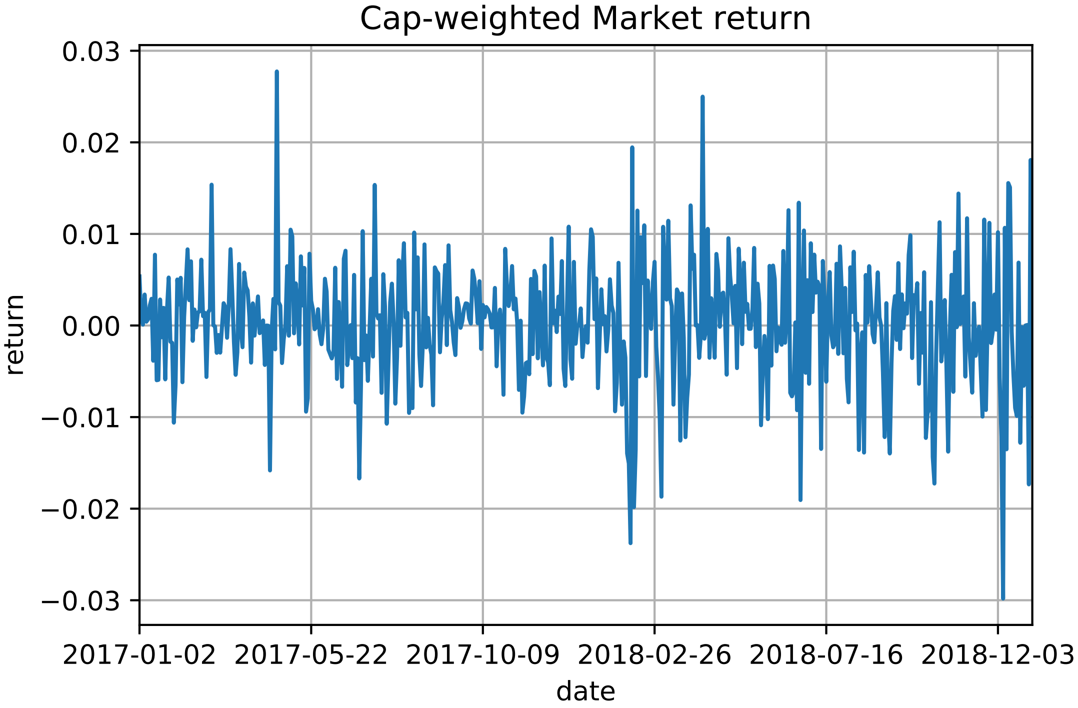


Figure 13: Caption

The cap-weighted market coefficients makes it such that the stocks with a higher market cap carry a heavier weighting percentage in the total market return. Arguably giving a more true representation of the market return.

4. Estimate the cap rolling $\beta_{m_{i,t}}$

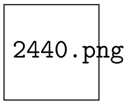


Figure 14: comparison of betas of normal market return and cap-weighted market return

We can see that by using cap-weighted market return, the histogram distribution is shifted to be centered around one.

5. Arbitrage pricing theory
   (a) —
   (b) —
   (c) —
   (d) —

3 Portfolio Optimization

3.1 Adaptive minimum-variance portfolio optimization

1. Derive optimal weights
   Solving the Lagrange optimization

$$\min_{\mathbf{w}} \quad J(\mathbf{w}, \mathbf{C}) = \frac{1}{2}\mathbf{w}^T\mathbf{C}\mathbf{w}$$

subject to $\quad \mathbf{w}^T\mathbf{1} = 1$

$$\begin{aligned} \tfrac{\partial J}{\partial \mathbf{w}} &= \mathbf{w}^T\mathbf{C} - \lambda\mathbf{e}^T = 0 \\ \tfrac{\partial J}{\partial \lambda} &= \mathbf{w}^T\mathbf{e} - 1 = 0 \end{aligned}$$

from $\frac{\partial J}{\partial \mathbf{w}}$

$$\mathbf{w}^T = \lambda\mathbf{e}^T\mathbf{C}^{-1}$$

$$\lambda\mathbf{e}^T\mathbf{C}^{-1}\mathbf{e} = 1$$

$$\lambda = \frac{1}{\mathbf{e}^T\mathbf{C}^{-1}\mathbf{e}}$$

Gives the optimal weight expression

$$w_{opt} = \frac{\mathbf{C}^{-1}\mathbf{e}}{\mathbf{e}^T\mathbf{C}^{-1}\mathbf{e}}$$

2. Splitting the data of the last 10 stocks into training and test data, with a 50/50 split. The training split is used to compute the minimum variance portfolio weights, compared against the performance obtained from the performance of an equally-weighted portfolio on the test data.

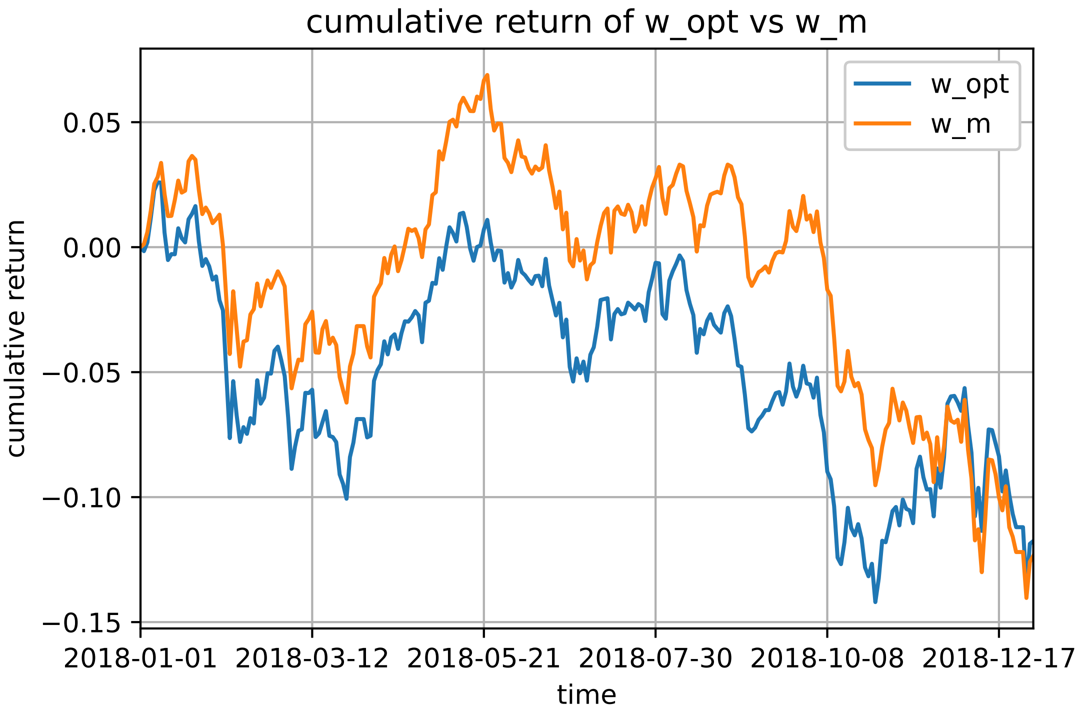
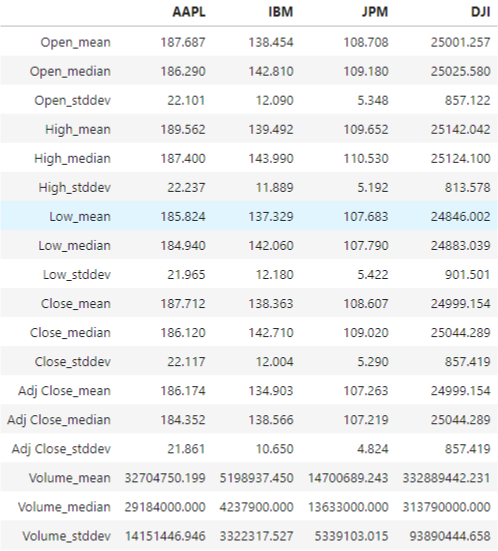


Figure 15: cumulative return over test data

The variance of $w_{opt} = 0.0014532270387243472$ and $w_m = 0.0008287808216735852$. The variance of both weights are similar, but w_m is slightly lower, this is likely due to the fact that w_{opt} is overfitted to the training data.

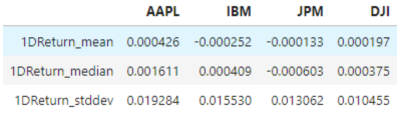
4 Robust Statistics and Non Linear Methods

4.1 Data Import and Exploratory Data Analysis



	AAPL	IBM	JPM	DJI
Open_mean	187.687	138.454	108.708	25001.257
Open_median	186.290	142.810	109.180	25025.580
Open_stddev	22.101	12.090	5.348	857.122
High_mean	189.562	139.492	109.652	25142.042
High_median	187.400	143.990	110.530	25124.100
High_stddev	22.237	11.889	5.192	813.578
Low_mean	185.824	137.329	107.683	24846.002
Low_median	184.940	142.060	107.790	24883.039
Low_stddev	21.965	12.180	5.422	901.501
Close_mean	187.712	138.363	108.607	24999.154
Close_median	186.120	142.710	109.020	25044.289
Close_stddev	22.117	12.004	5.290	857.419
Adj Close_mean	186.174	134.903	107.263	24999.154
Adj Close_median	184.352	138.566	107.219	25044.289
Adj Close_stddev	21.861	10.650	4.824	857.419
Volume_mean	32704750.199	5198937.450	14700689.243	332889442.231
Volume_median	29184000.000	4237900.000	13633000.000	313790000.000
Volume_stddev	14151446.946	3322317.527	5339103.015	93890444.658

Figure 16: key statistics of stocks and index



	AAPL	IBM	JPM	DJI
1DReturn_mean	0.000426	-0.000252	-0.000133	0.000197
1DReturn_median	0.001611	0.000409	-0.000603	0.000375
1DReturn_stddev	0.019284	0.015530	0.013062	0.010455

Figure 17: 1-Day return key statistics

1. Figure (16) and (17) shows the key statistics (mean, median, stddev) on the columns (Open, High, Low, Close, Adj Close, volume and 1-day Return) of the stocks AAPL, IBM, JPM and index DJI.

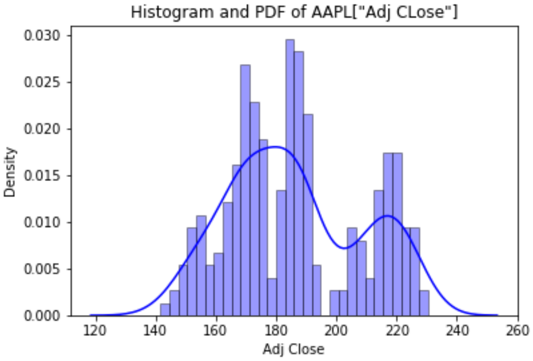


Figure 18: Histogram and pdf of AAPL

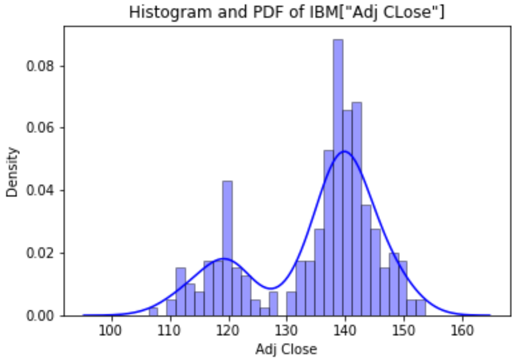


Figure 19: Histogram and pdf of IBM

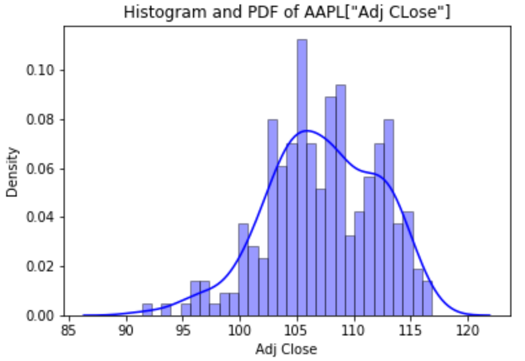


Figure 20: Histogram and pdf of JPM

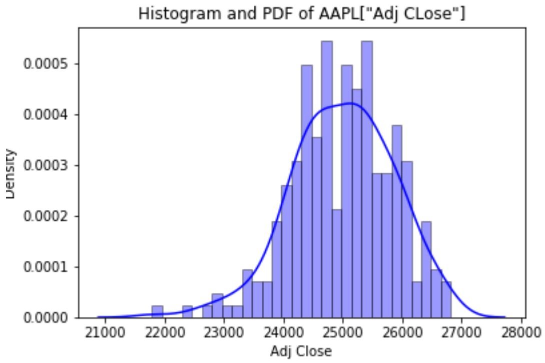


Figure 21: Histogram and pdf of DJI

2. Figure (18) to (21) shows the histogram and probability density function (pdf) of the adj. close of securities AAPL, IBM, JPM and DJI.

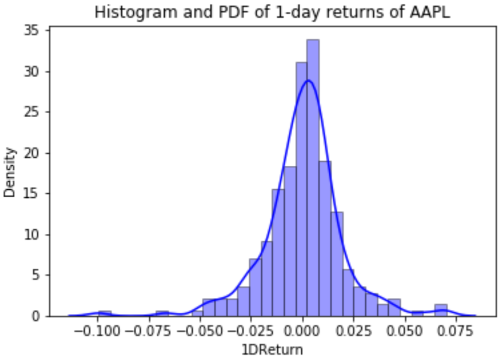


Figure 22: Histogram and pdf of 1-day returns of AAPL

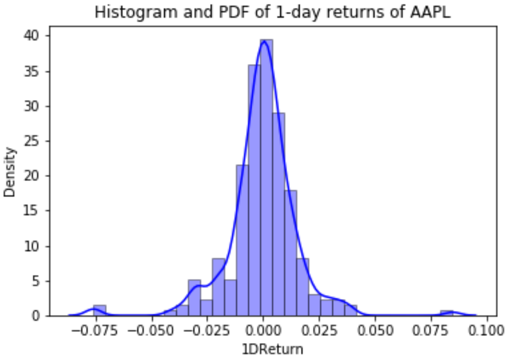


Figure 23: Histogram and pdf of 1-day returns of IBM

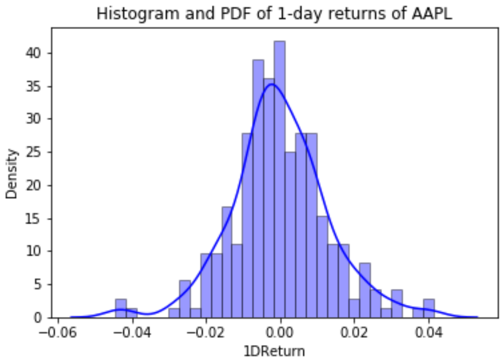


Figure 24: Histogram and pdf of 1-day returns of JPM

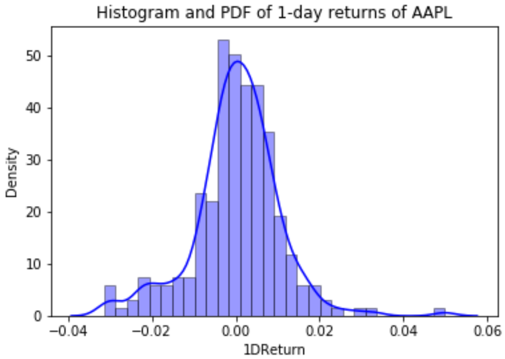


Figure 25: Histogram and pdf of 1-day returns of DJI

Figure (22) to (25) shows the histogram and probability density function (pdf) of the 1-day returns of securities AAPL, IBM, JPM and DJI.

It is observed that the histogram and pdf estimate of the 1-day returns is closer to a gaussian distribution, since it is mostly centered around zero. The adj. close histogram and pdf has a much higher variance and often have two peaks.

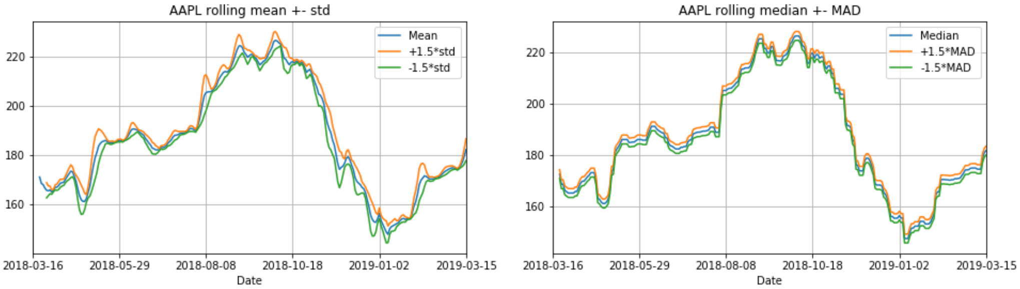


Figure 26: AAPL standard deviation median absolute deviation

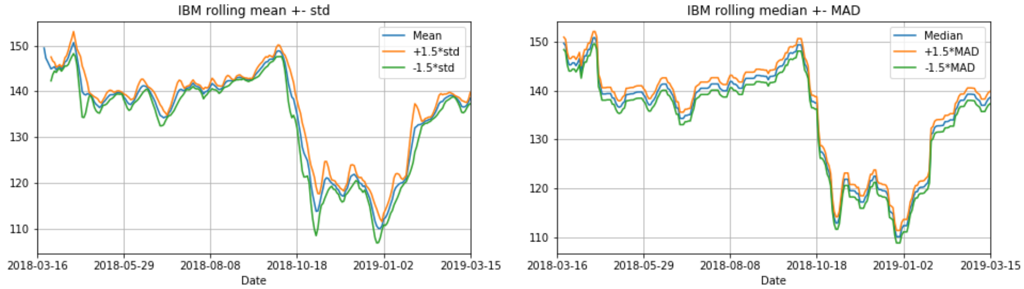


Figure 27: IBM standard deviation median absolute deviation

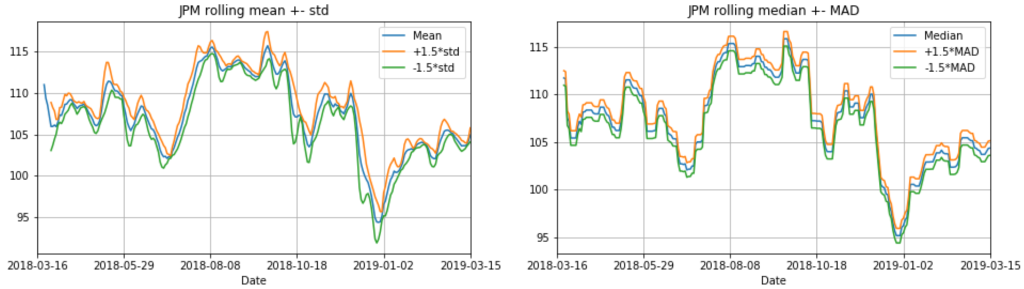


Figure 28: JPM standard deviation median absolute deviation

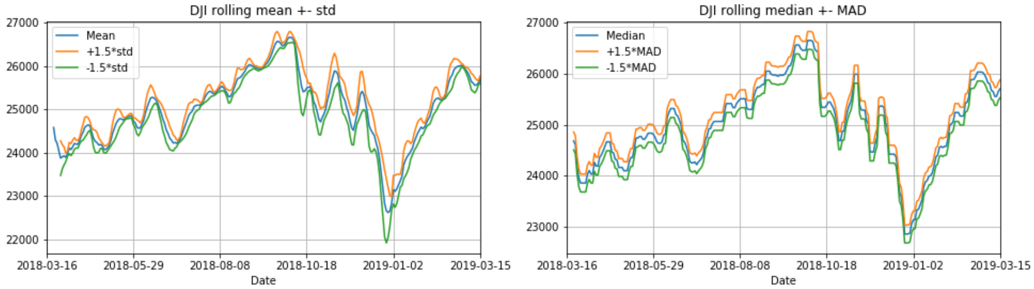


Figure 29: DJI standard deviation median absolute deviation

3. It is shown that with MAD, the upper and lower bound follows the shape of the original signal, while the standard deviation method creates values for the upper and lower bound that can be quite close to the original signal.

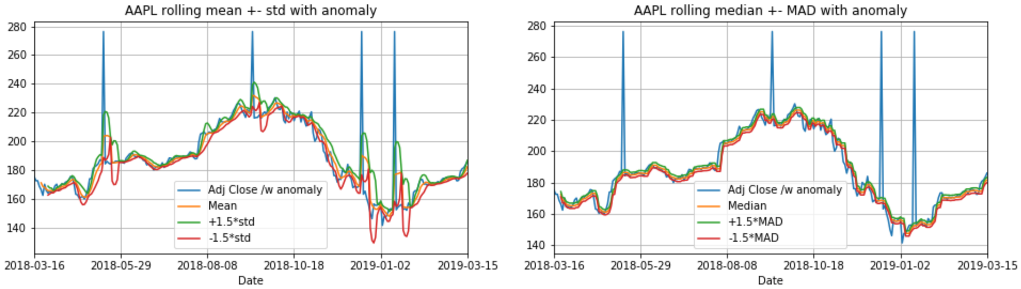


Figure 30: AAPL STDDEV & MAD w anomaly

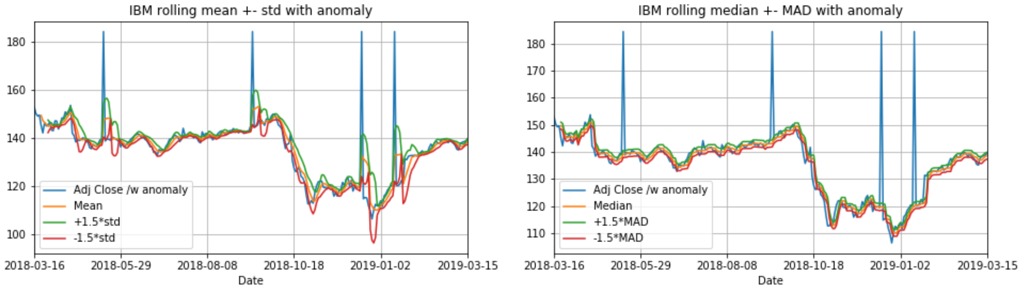


Figure 31: IBM STDDEV & MAD w anomaly

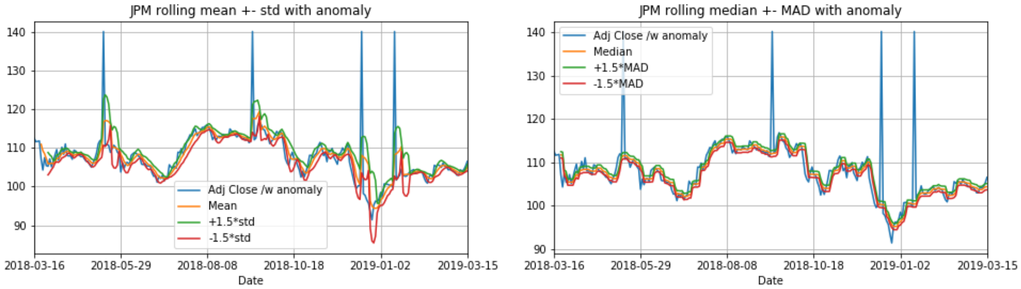


Figure 32: JPM STDDEV & MAD w anomaly

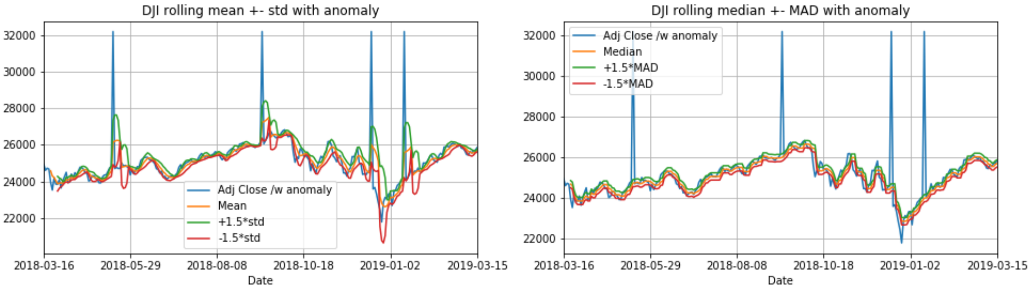


Figure 33: DJI STDDEV & MAD w anomaly

4. The figures above show the difference between the tolerance of the two methods against anomalies when introduced to the signal at dates 2018-05-14, 2018-09-14, 2018-12-14, 2019-01-14. For the case of using rolling mean and standard deviation, it is shown that the upper bound and lower bound is affected by the spike in the original signal, due to the fact that an average is applied on a 5 day window period. Conversely, the rolling median and MAD method is robust to the spike since such an exaggerated spike is never going to be the median value and will not affect the upper and lower bound.

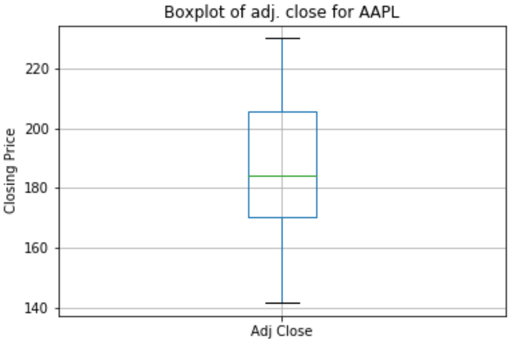


Figure 34: Boxplot of Adj. Close of AAPL

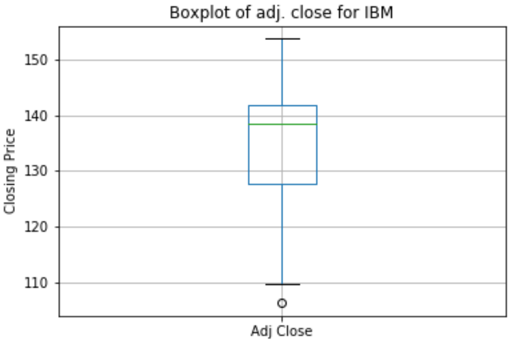


Figure 35: Boxplot of Adj. Close of IBM

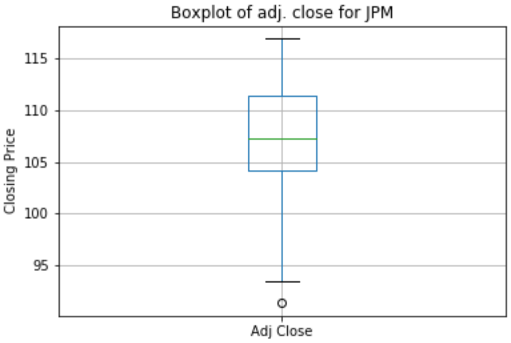


Figure 36: Boxplot of Adj. Close of JPM

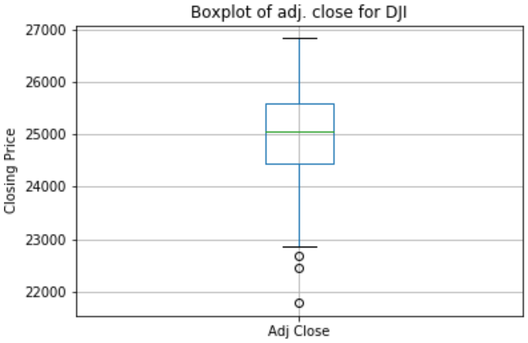


Figure 37: Boxplot of Adj. Close of DJI

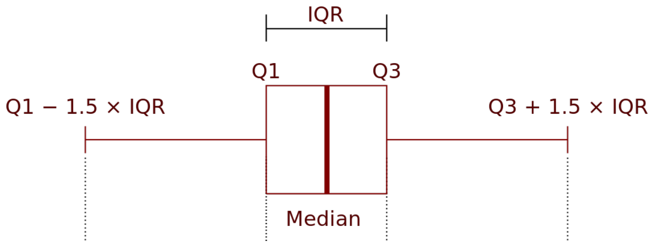


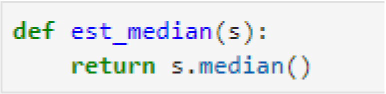
Figure 38: Information conveyed by a boxplot

5. The information the box plot conveys are shown in figure(38), where $Q1, Q3$ denotes the lower and upper quarter percentile (25% and 75% respectively). The points beyond $Q1 - 1.5 \times IQR$ and $Q3 + 1.5 \times IQR$ are considered outliers. Using the information from the box plots, we can get an insight of the distribution of the data.

   For example, it is shown that AAPL and JPM are slightly positively skewed, shown by the median (green line) is closer to the lower percentile, while IBM is greatly negatively skewed, and DJI showing almost no skewness. It is also shown that DJI has the most outliers, while IBM and JPM has some outliers and AAPL having almost none.

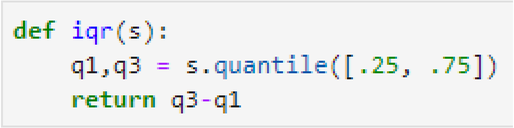
4.2 Robust Estimators

1. Python function for robust location and scalar estimators



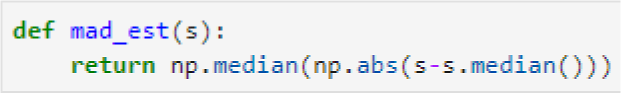
```python
def est_median(s):
    return s.median()
```

Figure 39: Median estimator python function



```python
def iqr(s):
    q1,q3 = s.quantile([.25, .75])
    return q3-q1
```

Figure 40: IQR estimator python function



```python
def mad_est(s):
    return np.median(np.abs(s-s.median()))
```

Figure 41: MAD estimator python function

2. Computational efficiency of the estimators are $O(n)$, $O(n \log n)$, and $O(n \log n)$. The most efficient algorithm for finding median is median of medians which has a worst case of $O(n)$ and the IQR and MAD algorithms are mostly dominated by unordered search algorithms, will have a worst case of $O(n \log n$.

3. The breakdown point is the point of which the estimator fails. For median as a robust location estimator, the breakdown point is where the outliers exceeded 50% of the data points before the outlier can be considered as the median. And since MAD uses median as well, they have the same breakdown point.

4.3 Robust and OLS regression

OLS and Huber regression is used to do linear regression to estimate the relationship between the DJI index and the other 3 stocks AAPL, IBM and JPM. Each of the stock is fitted to a model with the DJI index, and the model is used to give a prediction on the 1-day return.

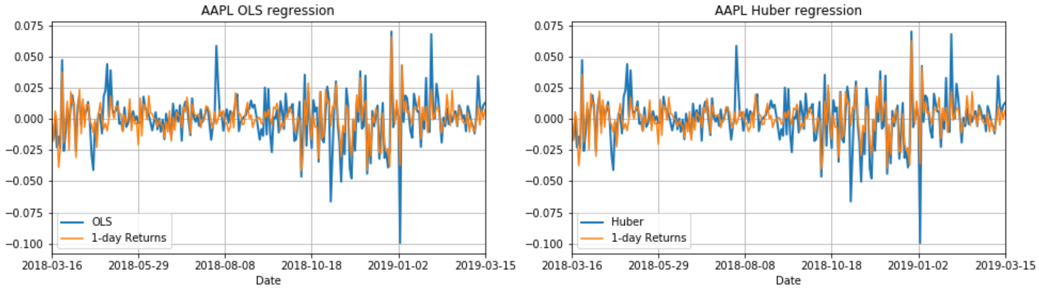


Figure 42: AAPL OLS and Huber regression against DJI

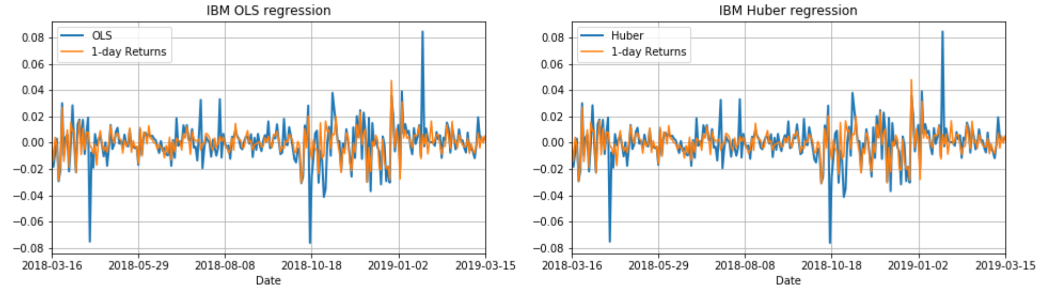


Figure 43: IBM OLS and Huber regression against DJI

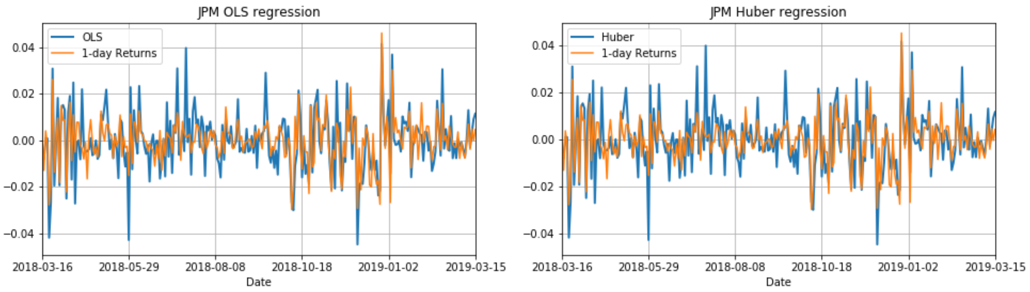


Figure 44: JPM OLS and Huber regression against DJI

We can see that both OLS and the Huber regression gives similar numerical predictions. We repeat the procedure of adding outlier to the data and apply OLS and Huber regression again.

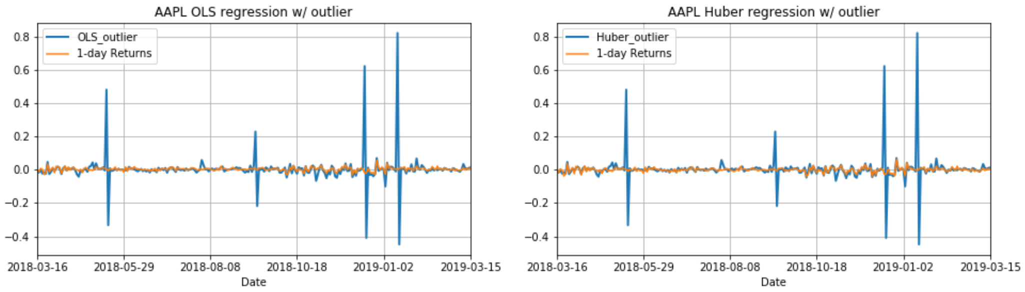


Figure 45: AAPL OLS and Huber regression against DJI with outlier

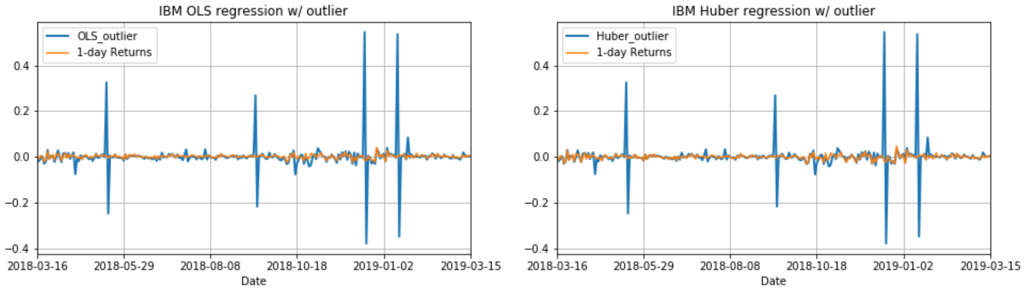


Figure 46: IBM OLS and Huber regression against DJI with outlier

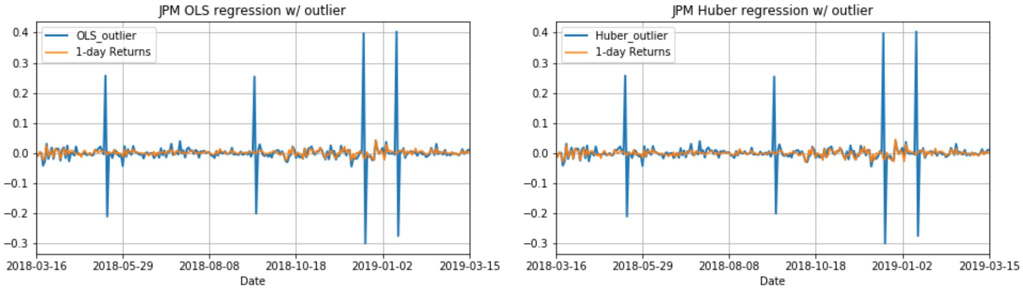


Figure 47: JPM OLS and Huber regression against DJI with outlier

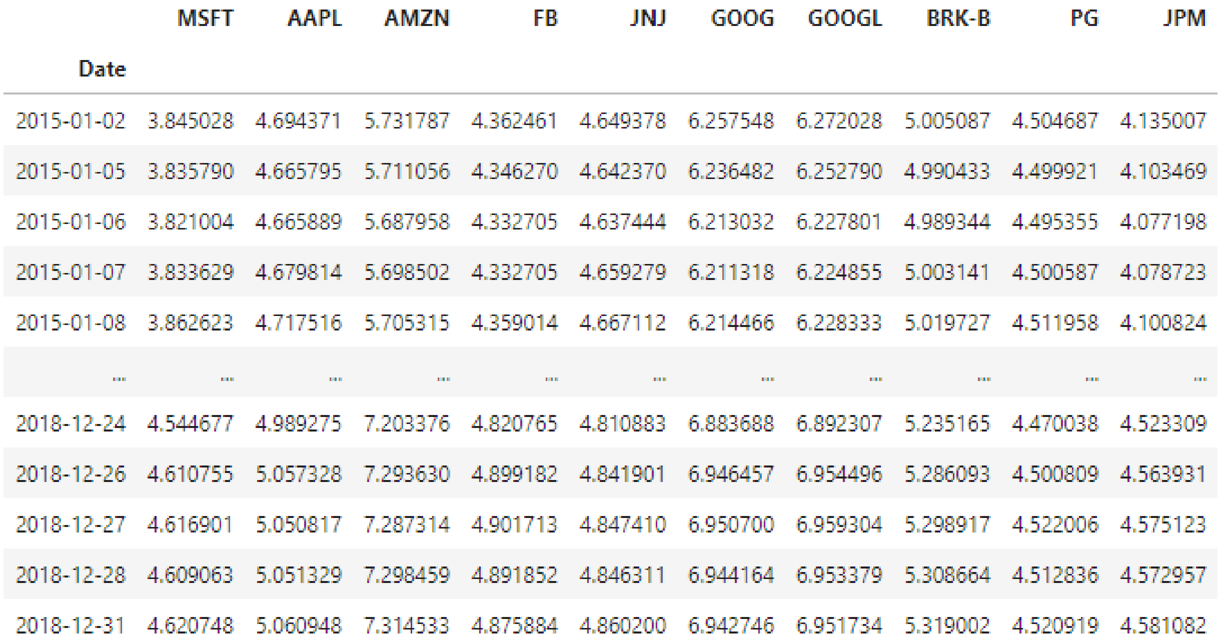
The OLS and Huber regression prediction gives similar answer as above and is robust against the outliers introduced.

4.4 Robust Trading Strategies

1. ——

2. ——

5 Graphs in Finance



	MSFT	AAPL	AMZN	FB	JNJ	GOOG	GOOGL	BRK-B	PG	JPM
Date										
2015-01-02	3.845028	4.694371	5.731787	4.362461	4.649378	6.257548	6.272028	5.005087	4.504687	4.135007
2015-01-05	3.835790	4.665795	5.711056	4.346270	4.642370	6.236482	6.252790	4.990433	4.499921	4.103469
2015-01-06	3.821004	4.665889	5.687958	4.332705	4.637444	6.213032	6.227801	4.989344	4.495355	4.077198
2015-01-07	3.833629	4.679814	5.698502	4.332705	4.659279	6.211318	6.224855	5.003141	4.500587	4.078723
2015-01-08	3.862623	4.717516	5.705315	4.359014	4.667112	6.214466	6.228333	5.019727	4.511958	4.100824
...	...	...	...	...	...	...	...	...	...	...
2018-12-24	4.544677	4.989275	7.203376	4.820765	4.810883	6.883688	6.892307	5.235165	4.470038	4.523309
2018-12-26	4.610755	5.057328	7.293630	4.899182	4.841901	6.946457	6.954496	5.286093	4.500809	4.563931
2018-12-27	4.616901	5.050817	7.287314	4.901713	4.847410	6.950700	6.959304	5.298917	4.522006	4.575123
2018-12-28	4.609063	5.051329	7.298459	4.891852	4.846311	6.944164	6.953379	5.308664	4.512836	4.572957
2018-12-31	4.620748	5.060948	7.314533	4.875884	4.860200	6.942746	6.951734	5.319002	4.520919	4.581082

Figure 48: Log returns of 10 companies from the S&P 500 index

1. The 10 companies chosen from the S&P 500 index are

   (a) Microsoft Corporation - MSFT

   (b) Apple Inc. - AAPL

   (c) Amazon.com Inc. - AMZN

   (d) Facebook Inc. Class A - FB

   (e) Johnson & Johnson- JNJ

   (f) Alphabet Inc. Class C - GOOG

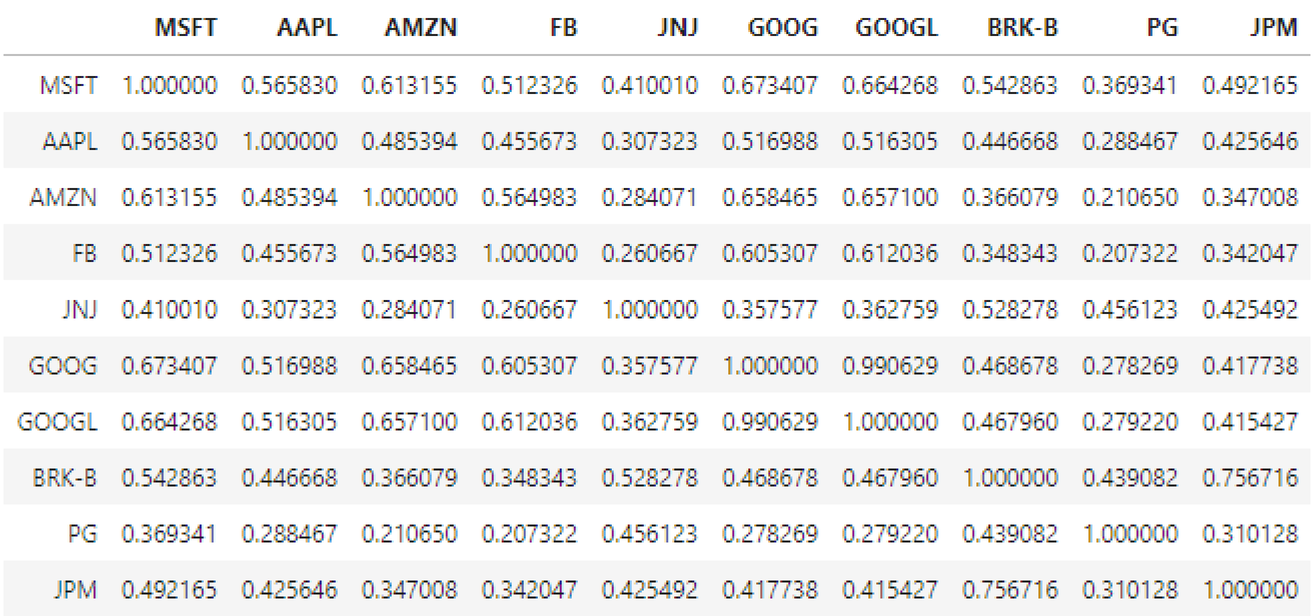
   (g) Alphabet Inc. Class A - GOOGL

   (h) Berkshire Hathaway Inc. Class B - BRK-B

   (i) Procter & Gamble Company - PG

   (j) JPMorgan Chase & Co. - JPM

   These companies have the largest market capitalization in the S&P 500 index. [3]



	MSFT	AAPL	AMZN	FB	JNJ	GOOG	GOOGL	BRK-B	PG	JPM
MSFT	1.000000	0.565830	0.613155	0.512326	0.410010	0.673407	0.664268	0.542863	0.369341	0.492165
AAPL	0.565830	1.000000	0.485394	0.455673	0.307323	0.516988	0.516305	0.446668	0.288467	0.425646
AMZN	0.613155	0.485394	1.000000	0.564983	0.284071	0.658465	0.657100	0.366079	0.210650	0.347008
FB	0.512326	0.455673	0.564983	1.000000	0.260667	0.605307	0.612036	0.348343	0.207322	0.342047
JNJ	0.410010	0.307323	0.284071	0.260667	1.000000	0.357577	0.362759	0.528278	0.456123	0.425492
GOOG	0.673407	0.516988	0.658465	0.605307	0.357577	1.000000	0.990629	0.468678	0.278269	0.417738
GOOGL	0.664268	0.516305	0.657100	0.612036	0.362759	0.990629	1.000000	0.467960	0.279220	0.415427
BRK-B	0.542863	0.446668	0.366079	0.348343	0.528278	0.468678	0.467960	1.000000	0.439082	0.756716
PG	0.369341	0.288467	0.210650	0.207322	0.456123	0.278269	0.279220	0.439082	1.000000	0.310128
JPM	0.492165	0.425646	0.347008	0.342047	0.425492	0.417738	0.415427	0.756716	0.310128	1.000000

Figure 49: Correlation matrix of log-returns

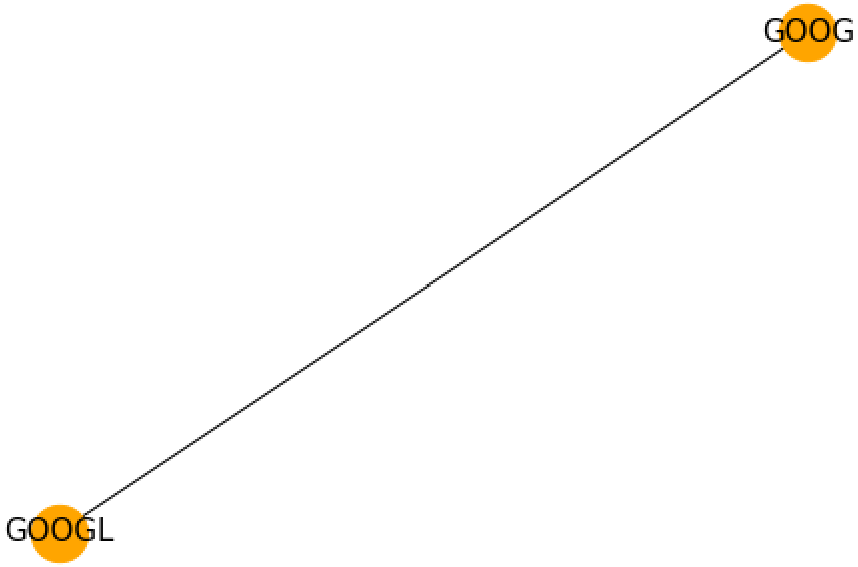


Figure 50: graph with threshold = 0.8

2. Computing the correlation matrix of the log-returns of the 10 stocks, we obtain the matrix in figure(49). The graph shown in figure(50) is obtained by drawing some relationship between attributes with correlation greater than a certain threshold value, in this particular case, the correlation threshold value is 0.8 as followed in the networkx tutorial. We can see that there is

only a single connection between GOOG and GOOGL, which is obvious since both of the tickers belongs to Google. If we relax the threshold value to 0.6, we get the following graph.

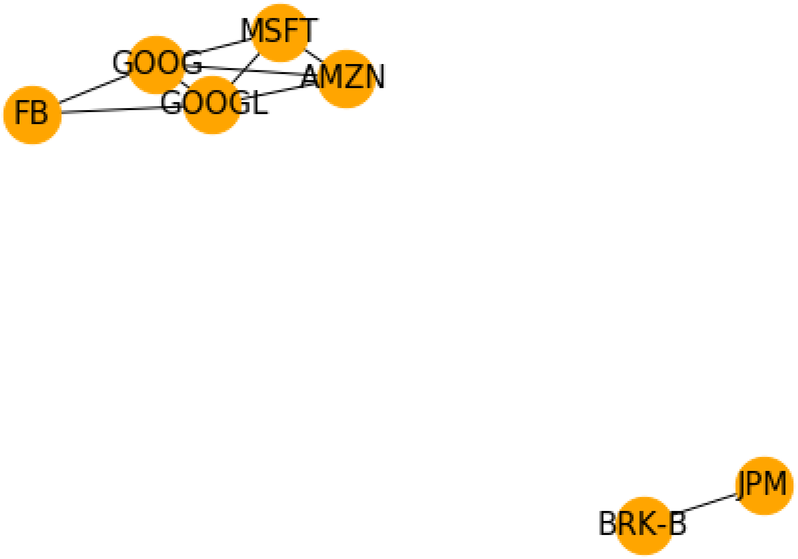


Figure 51: graph with threshold = 0.6

With a more relaxed correlation threshold, more correlation between companies are shown, particularly tech companies has high correlation, which makes sense. This implies that there might be something inducing the change in returns for the companies in the same tech sector, eg. perhaps there is a breakthrough in hardware technology which improves the performance of products produced by these companies.

3. The graph below is generated in the same way as above, except that the log-return data is shuffled. A correlation threshold of 0.6 is used.

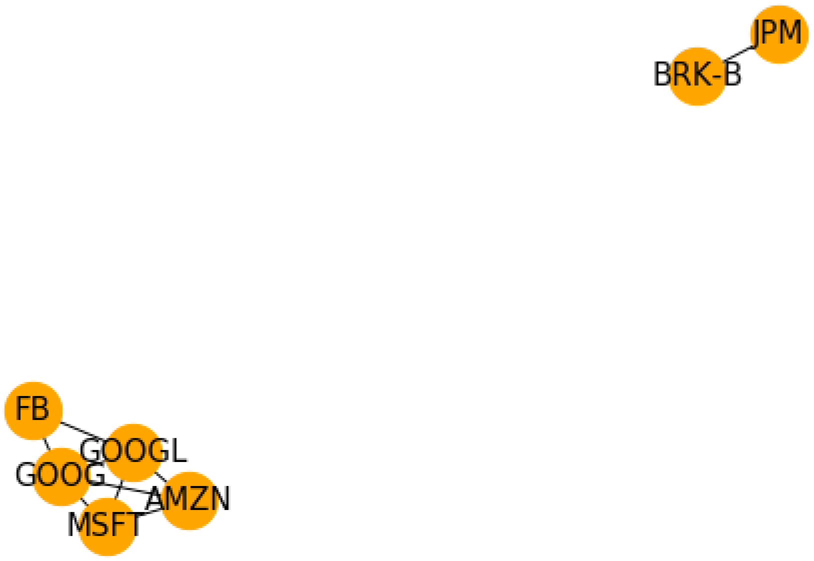


Figure 52: Shuffled log-returns

We can see from figure(52) it is the same, this makes sense since simply shuffling the data used to produce the correlation matrix will not change the relationship. However, if we shuffle the time series data directly, we get

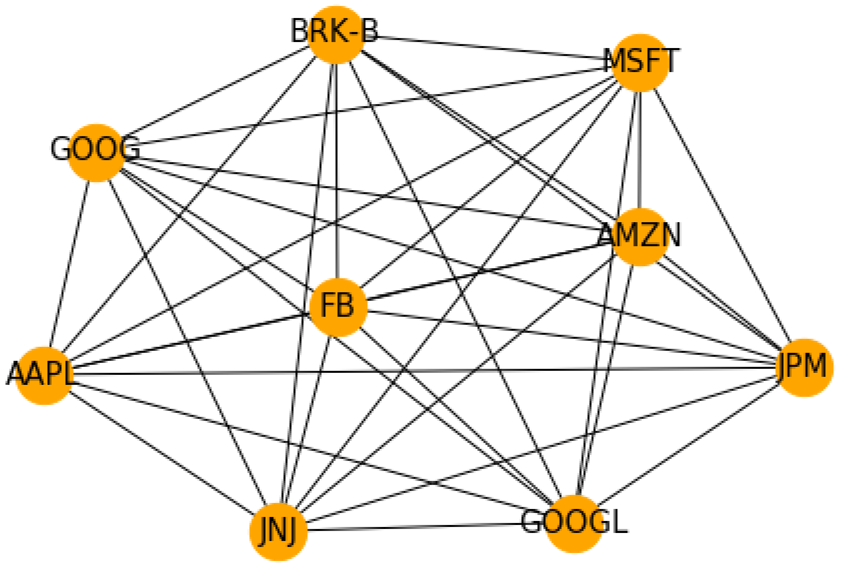


Figure 53: Shuffled time-series data

This is because the log-return function applied to the time-series before producing the correlation matrix takes into the account of sequence of the time-series data, shuffling will give a completely different result.

4. —

5. —

References

[1] "scipy jarque-bera documentation." https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.jarque_bera.html. Accessed: 2020-02-27.

[2] S. Nickolas, "The formula for calculating beta," Feb 2020.

[3] "Sp 500 companies by weight."