EE4-69: Financial Signal Processing Coursework

April 18, 2019

Prof. Danilo P. Mandic (2018-2019)
Author: Divyansh Manocha (CID: 01053537)

Please note that this a PDF generated directly from the notebook, and has no additional information. While most of the formatting errors from the generation have been fixed, it is recommended that the HTML be read as the formatting of the report has been corrected with that format in mind.

Abstract

The coursework primarily focuses on providing practical insights into common financial and econometrics theory, using signal processing techniques. The notebook is split into four sections - focusing on regression methods, bond pricing, portfolio optimisation, robust statistical methods and graph theory. It makes use of the S&P 500 stocks dataset, daily returns of European counties, as well as some regular stocks of individual companies for a specified time frame.

A table of contents for the structure of this notebook is provided below:

Contents

Notation:

Symbol	Description
x, $\boldsymbol{x}$, $\boldsymbol{X}$, χ	Scalar, vector, matrix and tensor
x[n]	Elements of a vector or scalar realisations of random variables

1 Regression Methods

The purpose of this part of the coursework is to produce an exploration of regression methods on financial data.

1.1 Processing stock price data

Loading Dataset
Consider the dataset. There are 22629 entries from 1930-02-01 to 2017-06-21. If the dataset were a true time series, with equal intervals, there should be 31947 days. If holidays and weekends are exclueded, the dataset was expected to have 22375 entries. The small discrepency may be because the trading industry may have different definitions of such holidays to the tool used to estimate this. This therefore cannot be considered a time series dataset as it is, and should not be plotted with date.

Hence in this report the x-axis is left as indices of events, assumed to be equally spaced - by a working day.

```
The number of entries: 22629, the number of unique days: 22629
```

Ensure that all days are unique - thereby allowing the treatement of the dataset as a time series of working days.
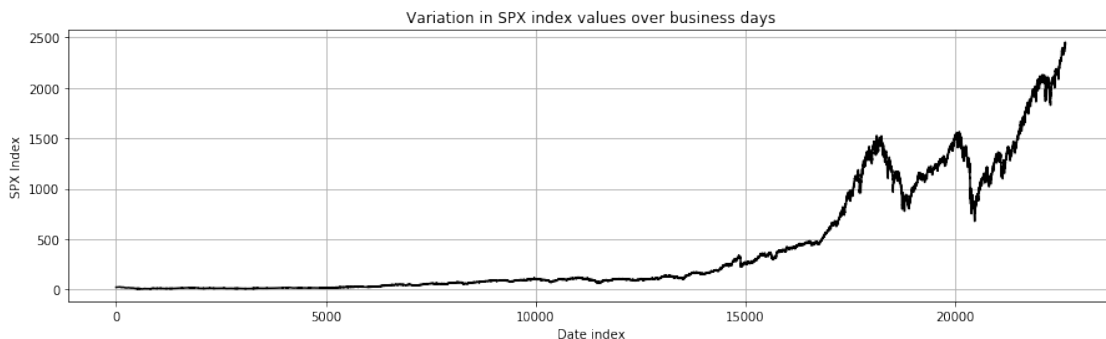
1.1.1 Log time plot
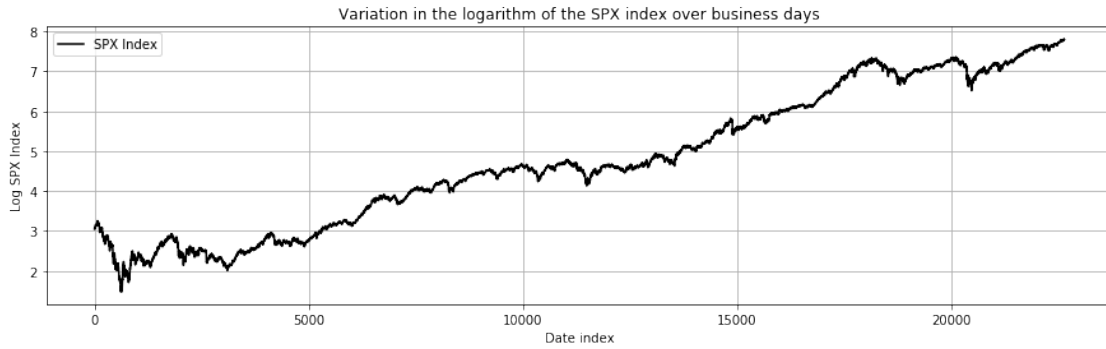
1.1.2 Sliding windows on working year
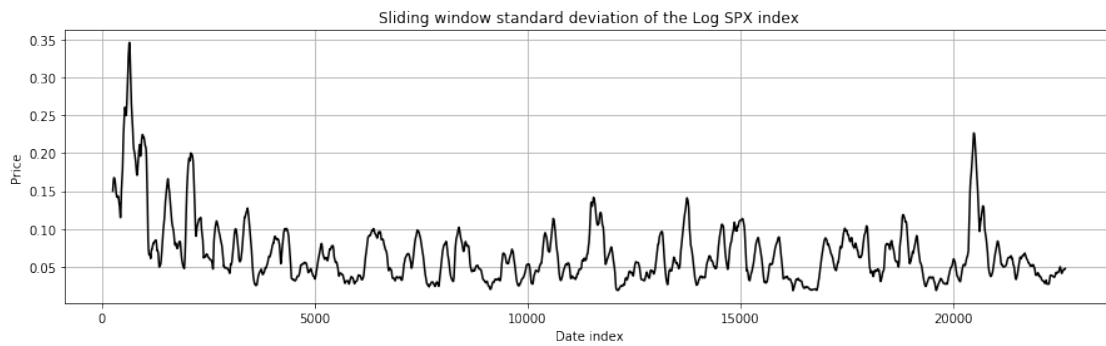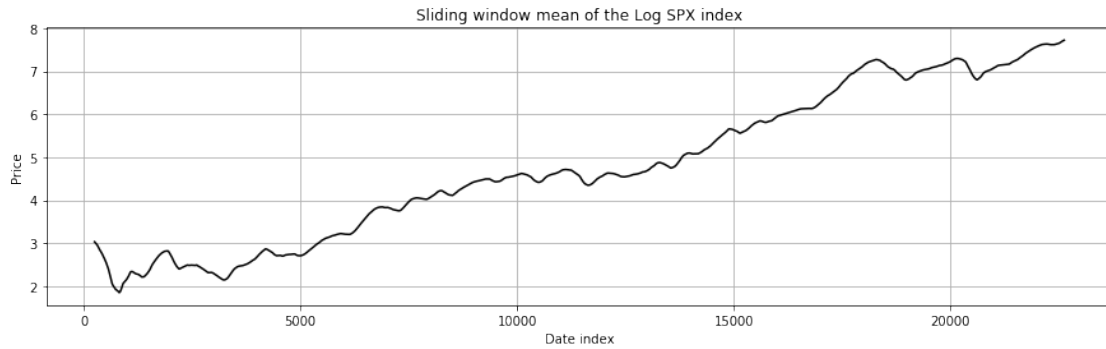
Discussion

A stationary time series is one Section **??** that exhibits statisticla properties that are invariant to a shift in the time index. For a first order strict sense stationarity $f_x(x,t)f_x(x,t+c)$. For wide sense stationarity $E\{x(t)\} = \mu$ and $E\{x(t_1)x(t_2)\} = R_{xx}(t_1 - t_2)$.

As can be seen the plots for the sliding windows of the log time series do not appear to be stationary.

1.1.3 Simple and log returns
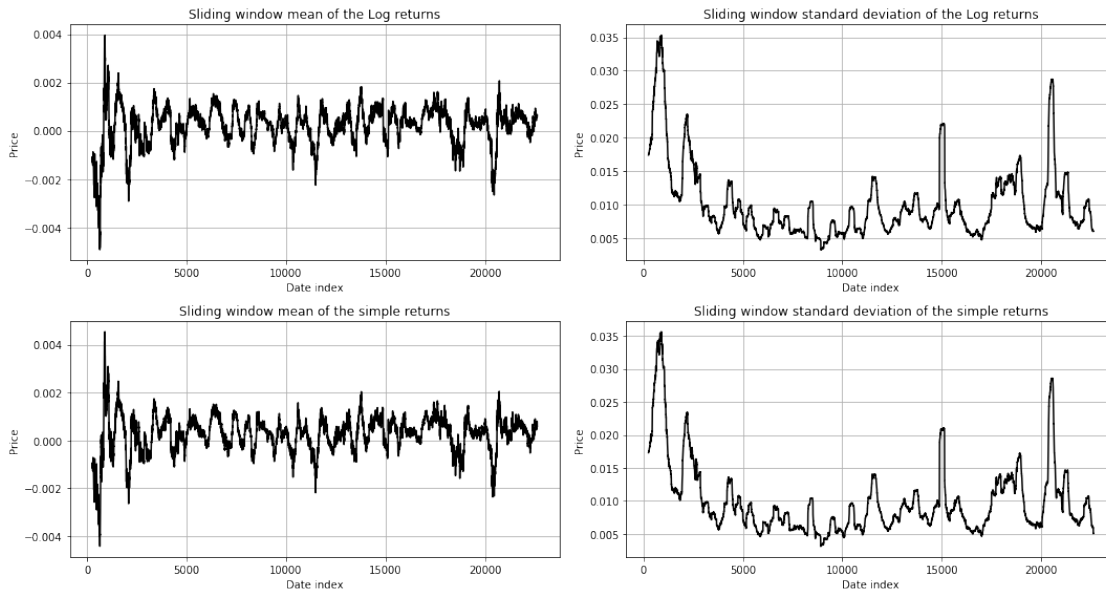
Discussion

Continuing from the earlier discussion of section 1.1.2, by visual inspection it can be seen that the plots obtained from differencing the log returns or conducting a percentage change for the simple return are more likely to be stationary. Changes in logarithmn are almost equivalent to percentage changes of the original datatset.

Trends and seasonal effects are easier to discern than changes to the variance. For this purpose we use the Augmented Dickey Full Test.

Hypothesis

Hypothesis	
Null Hypothesis H_0	Time series has a unit root - non stationarity
Alternative Hypothesis H_1	Rejected the Null Hypothesis

As a rule of thumb, an obtained p-value less than n (usually 0.05) is usually considered to reject the null hypothesis

```
ADF P-value of the price: 1.0
ADF P-value of the simple returns: 0.0
ADF P-value of the log: 0.9863263713847654
ADF P-value of the log returns: 0.0
```

Clearly differencing the time series data has produced, what can be considered, a stationary result

1.1.4 Comparing log returns to simple returns

Discussion

The purpose of using returns as opposed to prices is the normalisation of data. Most moving average models rely on the dataset being stationary, which can be achieved through differencing as shown above.

Simple returns can be considered to be defined as $R^s_{[0,T]} = \frac{P_T - P_0}{P_0}$ for one period.

Log returns for one period are defined as $R^L_{[0,T]} = ln\left(\frac{P_T - P_0}{P_0}\right) = ln\left(1 + R^s_{[0,T]}\right)$. If these returns are invariant, it can be easily projected to k steps with Fourier transforms.

The most useful property of log returns exploits the additive property of logarithms. Consider a sequence of n events/trades, and the compounding return is given by $\Pi_i(1 + r_i)$. For uncorrelated datasets, the sum of normally distributed variables is quite simple to calculate as normally distributed.

$\sum_i log(1 + r_i) = log(1 + r_1) + log(1 + r_2) + ... + log(1 + r_n)$

This is particularly useful for a time series dataset.

Log returns are approximately invariant, and therefore allow better estimation of distributions. The distribution of time series, prices for example, are log-normal. Unlike simple arithmetic returns, log returns will therefore result in normal returns. This normal distribution allows for easier time aggregation methods.

The Jarque Bera test can be used to quantitatively examine when the dataset is normal.

Hypothesis	
Null Hypothesis H_0	Samples from a normal distribution. Zero skewness and zero excess kurtosis
Alternative Hypothesis H_1	Rejected the Null Hypothesis

```
The Jarque-Bera test for the pricesand log prices are as follows:
 (9841.00501701649, 0.0) (1393.703385565961, 0.0)


The Jarque-Bera test for the log returnsand simple returns are as follows:
(309277.64126696053, 0.0) (284160.67730312835, 0.0), for comparison purposes
```

We have said that log returns are useful in the event that the time series is log-normal. The Jarque-Bera test clearly fails both the datasets, and it cannot be concluded that it is actually log-normal. This is different to the theoretical expectation. One possible explanation for this is the long time scales used - the JB test diverges in distribution as the sample size becomes larger. Therefore the null hypothesis is rejected even if the errors are normally distributed, due to nonstationarity.

Clearly the p value is larger than 0.05, meaning the null hypothesis is not rejected.

1.1.5 Simple and Log Returns Example

```
Simple returns for the example:
     0
1  1.0
2 -0.5
Log returns for the example:
```

```
        0
1   0.693147
2  -0.693147
```

Discussion

Log returns in this case indicate the same amount of increase as decrease, which simple returns do not. Log returns do not give a direct measure of the change in wealth. Further if the lower bound of a simple return is R_t, then the lower bound of the log return will be $ln(1 + R_t)$. Therefore the log return has a lower bound, whilst a normal random variable does not. Therefore log returns are more suitable to manage risks.

### 1.1.6 When should Log returns not be used

As was found above, log-returns do not give a direct measure of the change in wealth. This is particularly a problem for stochastic systems, as investigated by Meucci et al.Section **??** Section **??**. In the same paper it is proven that mean log returns are directly related to the mean and variance of simple returns. Therefore the mean of the log returns cannot be used to infer the mean of the simple returns, in cases where it is required. This difference therefore incerases with the variance of the data. Therefore they cannot be used to obtain arithmetic portfolio return.
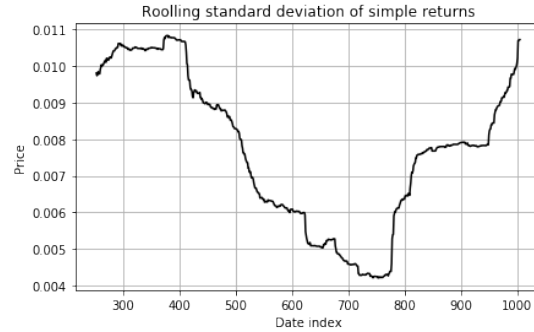
1.2 ARMA vs. ARIMA Models for Financial Applications

An *ARMA(p,q)* model is defined to be as follows:

$$x[t] = \sum_{i=1}^{p} a_i x[t-i] + \sum_{i=1}^{q} b_i \eta[t-i] + \eta[t] \tag{1}$$

1.2.1 Closing prices of S&P 500
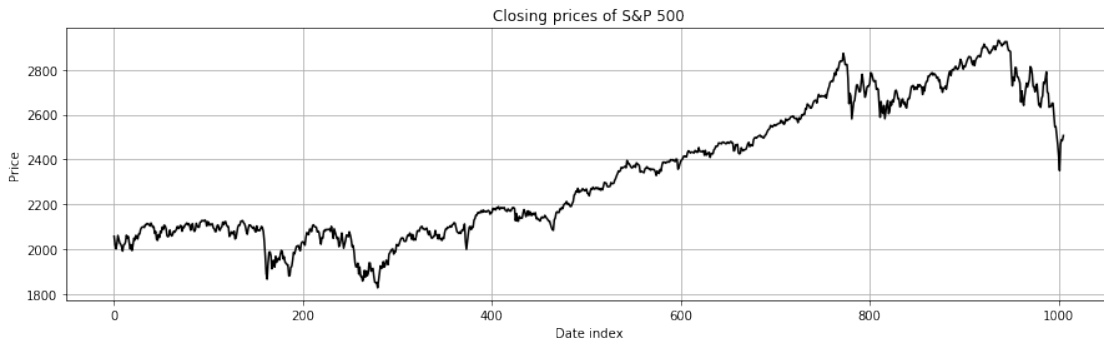
```
ADF P-value of the price: 0.7025879191711838
ADF P-value of the simple returns: 0.0
ADF P-value of the log: 0.6675983326694176
ADF P-value of the log returns: 0.0


The Jarque-Bera test for the log and normal data series are as follows:
 (80.04678574015176, 0.0) (86.55065452871632, 0.0)
```
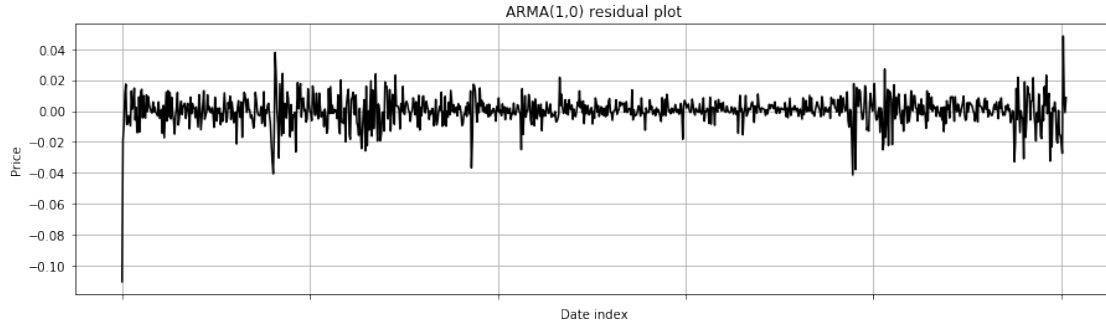
Since ARMA models assume stationarity, they are more apppropriate when no differencing is involved. If differencing us needed, ARIMA would be more appropriate. The tests above clearly shows that differencing provides a stationary time series.

1.2.2 Fitting an ARMA(1,0)

ARMA(1,0) residual plot
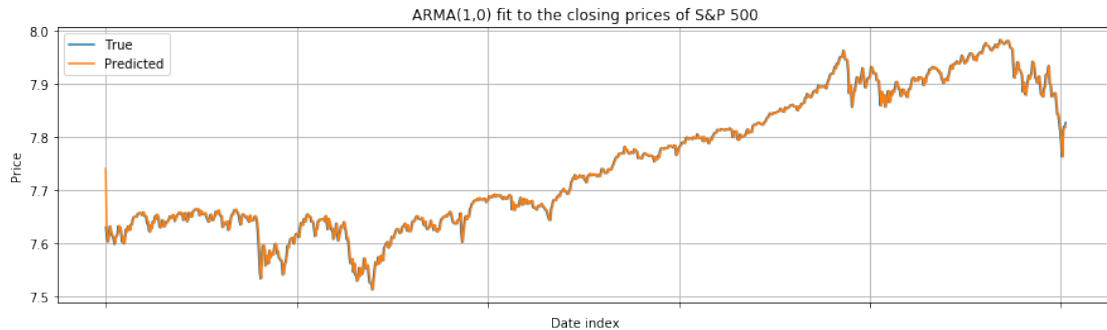
An ARIMA(1,0,0) is first order autoregressive model. The lag order is 1, the degree of differencing is 0 and the order of moving average is also 0. Upon the assumption that the time series is stationary and autocorrelated, it can be represented as:

$$Y_t = c + \phi_1 Y_{t-1}$$

It can be stated that this is Y regressed on itself with a lag of one. This is precisely an $ARIMA(1,0,0) + c$ model, where c is a constant derived from the mean of Y. Since the current value is only determined by its previous, it can be said to be a Markov process.

ϕ_1 may be considered as the slope. The next value will be ϕ_1 times more distant from the mean value than the current period. In the case that it is positive and $|\phi_1| < 1$, i.e. stationary and stable, the model shows mean reverting behaviour. If it is negative, alternating behaviour will be observed - Y will be greater than the mean in the subsequent period if it is less in the current.

In this context, we observe that the parameters of ARIMA(1,0,0) are close to one, suggesting that the price of the stock at t_1 will be very similar to that at t. In predicting stocks, this is not very useful.

1.2.3 Fitting an ARIMA(1,1,0)

ARIMA(1,1,0) fit to the closing prices of S&P 500

ARIMA(1,1,0) residual plot
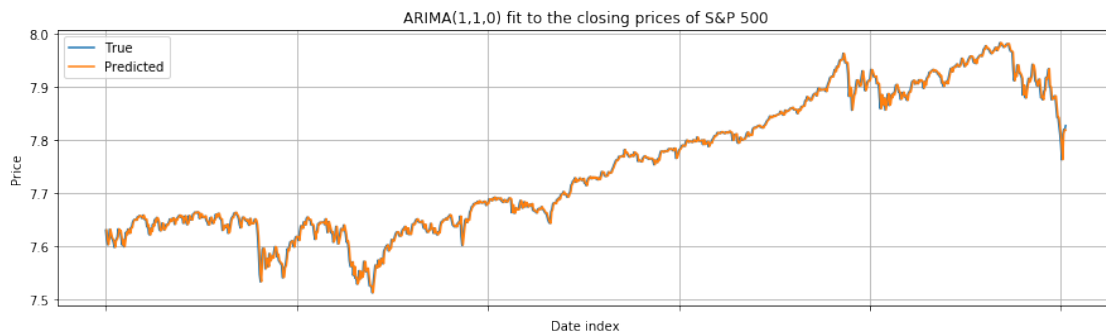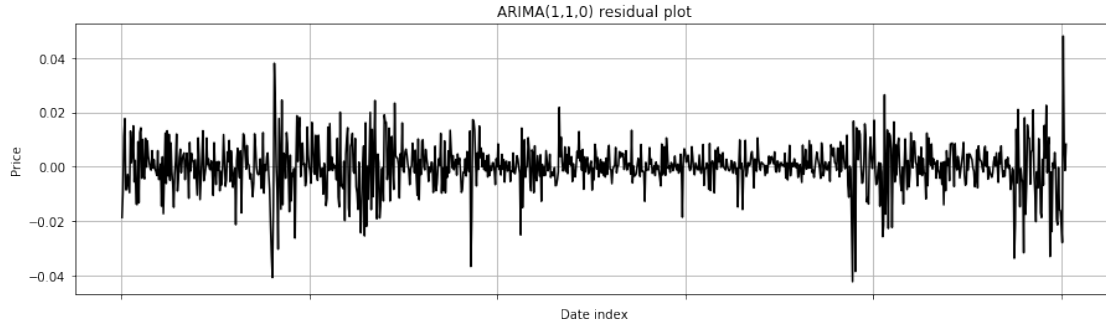
A first order autoregressive model with a degree of differencing is 1 and the order of moving average is also 0. In the case that the errors are autocorrelated, adding a lag of the dependant variable may provide a better estimate. This can be represented in the following equations:

$$
\begin{aligned}
_t - Y_{t-1} &= \mu + _1(Y_{t-1} - Y_{t-2}) \\
_t - Y_{t-1} &= \mu \\
\therefore _t &= \mu + Y_{t-1} + m_1(Y_{t-1} - Y_{t-2})
\end{aligned}
\tag{2}
$$

Discussion

In both ARMA and ARIMA models autocorrelation over time is assumed to be constant, and we are therefore trying to model signal with noise. From the residual plots, ARMA seems to be better fit the dataset by average value of the residuals. However it does suffer from spikes at the beginning and end. As the lag order is only 1, predictions of only the very near future can be made with some reasonable accuracy. If the objective is only to predict what might happen in the following event, then the models may be of use. However they cannot be used for any long term prediction.

1.2.4 Log of the prices for ARIMA

Necessity of logarithms:

This section adds on to the discussion of Section 1.1.4, which is relevant for this question. As was discussed earlier, log returns depend on simple return and the variance of the time series. Therefore logs are useful if the variance varies significantly (usually increases) with time in the dataset. Models that use differencing, such as ARIMA, assume little.

Taking the log of a series maintains the trend of the time series, however it reduces the heteroscedacity by flattening exponential growth patterns. This makes fitting linear models more appropriate. Unlike a deflating method, it maintains trends.

In situations where the marignal effect of one variable affects the expected value of the other, percentage changes are more appropriate than absolute changes. Therefore in such cases taking the log before perfroming a moving average will be appropriate.

Logs are also used in the case that a multiplicative regression model is being used, that implements weighted geometric mean $Y(t) = X_1^{\alpha 1}(t)...X_k^{\alpha k}(t)\epsilon(t)$. This is because they ensure the parameters are linear. [Section **??**]

1.3 Vector Autoregressive (VAR) Models

1.3.1 Concise representation of VAR processes

$$Y = BZ + U$$

$$Y \in \mathbb{R}^{K \times T}, B \in \mathbb{R}^{K \times (KP+1)}, Z \in \mathbb{R}^{(KP+1) \times T} \text{ and } U \in \mathbb{R}^{K \times T} \tag{3}$$

$$\therefore \mathbf{B} = [c, A_1, A_2, ... A_p]$$

Consider the matrix notation:

$$\mathbf{Y} = \begin{bmatrix} (y_{11}...y_{1T}) \\ (y_{21}...y_{2T}) \\ \vdots \\ (y_{k1}...y_{kT}) \end{bmatrix} = \begin{bmatrix} \mathbf{y_t} \\ \mathbf{y_{t-1}} \\ \vdots \\ \mathbf{y_{t-p}} \end{bmatrix} = \mathbf{BZ} + \mathbf{U} \tag{4}$$

By Comparison it can be deduced that:

$$Z = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ y_{p-1} & y_p & \cdots & y_{T-1} \\ y_{p-2} & y_{p-1} & \cdots & y_{T-2} \\ \vdots & \vdots & \ddots & \vdots \\ y_0 & y_1 & \cdots & y_{T-p} \end{bmatrix} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ y_{1,p-1} & y_{1,p} & \cdots & y_{1,T-1} \\ y_{2,p-1} & y_{2,p} & \cdots & y_{2,T-1} \\ \vdots & \vdots & \ddots & \vdots \\ y_{k,p-1} & y_{k,p} & \cdots & y_{k,T-1} \\ y_{1,p-2} & y_{1,p-1} & \cdots & y_{1,T-2} \\ y_{2,p-2} & y_{2,p-1} & \cdots & y_{2,T-2} \\ \vdots & \vdots & \ddots & \vdots \\ y_{k,p-2} & y_{k,p-1} & \cdots & y_{k,T-2} \\ \vdots & \vdots & \ddots & \vdots \\ y_{1,0} & y_{1,1} & \cdots & y_{1,T-p} \\ y_{2,0} & y_{2,1} & \cdots & y_{2,T-p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{k,0} & y_{k,1} & \cdots & y_{k,T-p} \end{bmatrix} \tag{5}$$

$$U = \begin{bmatrix} e_p & e_{p+1} & \cdots & e_T \end{bmatrix} = \begin{bmatrix} e_{1,p} & e_{1,p+1} & \cdots & e_{1,T} \\ e_{2,p} & e_{2,p+1} & \cdots & e_{2,T} \\ \vdots & \vdots & \ddots & \vdots \\ e_{k,p} & e_{k,p+1} & \cdots & e_{k,T} \end{bmatrix} \tag{6}$$

Since a constant multiplied something is still a constant

1.3.2 Deriving optimal B

Consider the ordinary least squares estimation for the optimal B: B_{opt}, where the goal would be to minimise the residual. Therefore it can be given as follows:

$$B_{opt} = argmin_B ||\mathbf{y} - B\mathbf{Z}||^2$$

$$\frac{\partial}{\partial B}(\mathbf{Y} - B\mathbf{Z})^T(\mathbf{Y} - B\mathbf{Z}) = 0$$

$$\implies 2(\mathbf{Y} - B\mathbf{Z}) = 0 \tag{7}$$

$$\implies \mathbf{Y} = B\mathbf{Z}$$

$$YZ^T = BZZ^T$$

$$(YZ^T)(ZZ^T)^{-1} = B$$

The OLS estimate gives $\mathbf{Y} \approx \mathbf{B}\mathbf{Z}$ and $\therefore B_{opt} = (YZ^T)(ZZ^T)^{-1}$ as required

1.3.3 Stability of VAR

Considering impulse responses in the first order case of the VAR process:
$y_t = Ay_{t1} + e_t$
Induction is used to explore the effect that the jth element of the vector e_t has on the ith element of a state y_t n periods later.

$$y_t = Ay_{t1} + e_t$$

$$y_t = A\left(Ay_{t2} + e_{t-1}\right) + e_t = A^2 y_{t2} + Ae_{t-1} + e_t \tag{8}$$

$$y_t = A\left(A(Ay_{t3} + e_{t-2}) + e_{t-1}\right) + e_t = A^3 y_{t3} + A^2 e_{t-2} + Ae_{t-1} + e_t$$

In the case of two periods, this will be given by the matrix A^2. From these steps it can be seen that the effect of any one impulse e_t will effect y at all times in the future.

In order for this to be a stable process, this effect needs to be diminishing. Therefore $|A^k| <= 1$ for stability, which means that the eigenvalues are less than 1 in absolute value.

1.3.4 Fitting a VAR model

Loading Dataset

Visualising detrend
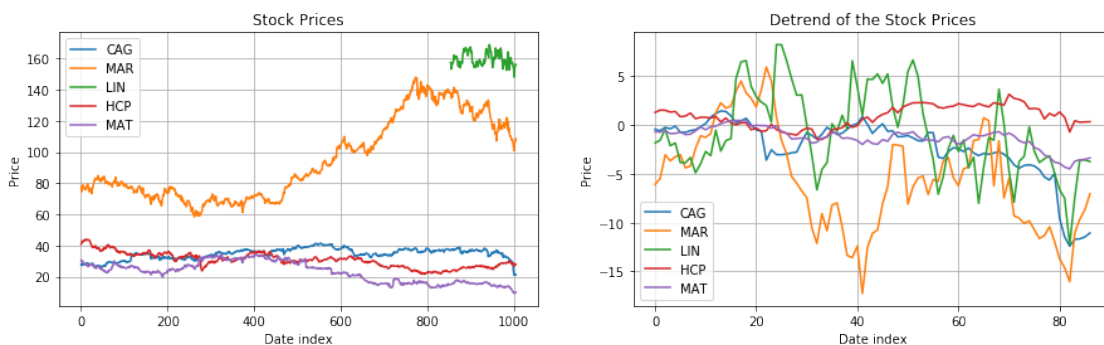
```
Eigenvalues: [0.72609393 0.72609393 1.00635964 0.86051894 0.91144512]
P-value of the Jarque-Bera-test: 7.554199683120128e-48
The correlation matrix:
```

Discussion

Firsly, consider the model itself with the dataset.

The Jarque-Bera test clearly shows the errors are normally distributed, one of the weak assumptions of VAR. However as can be seen from Section A and Section B, the chosen model may not be the most appropriate one for this dataset. Section **??** As all of the eigenvalues are not less than one, VAR may not be an appropriate model for this particular stock data due to instability.

Now consider whether it is a good stock portolio. Please note a complete portfolio optimisation analysis is out of the scope of this coursework. With the full dataset, MontCarlo simulations would also be used.

To be a good stock portfolio, a diverse range of stocks are needed. This is to ensure that the downfall of one does not directly mean the downfall of others. For such few tickers, they will have to be uncorrelated or a good balance of positive and negative correlations to be considered an appropriate portfolio. Through the correlation matrix, clearly stocks MAR AND MAT are highly correlated negatively, which is also reflected in the covariance matrix of their residuals found in Section A. Similarly HCP has a high positive correlation index with MAR and MAT. Therefore it is my conclusion that a porfolio using these stocks will not be a good one.
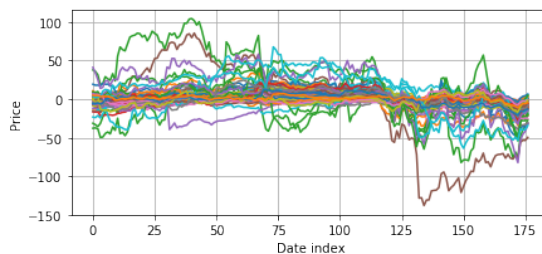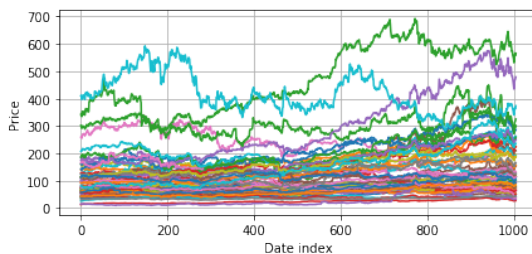
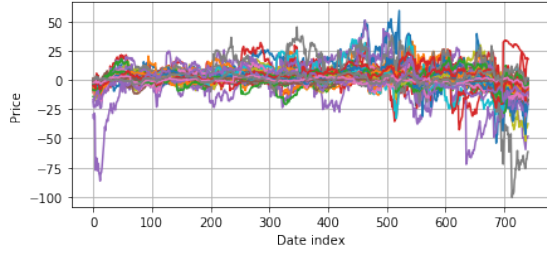### 1.3.5 Grouping the stocks by sectors

```
The industries present:
['Industrials' 'Health Care' 'Information Technology'
 'Communication Services' 'Consumer Discretionary' 'Utilities'
 'Financials' 'Materials' 'Real Estate' 'Consumer Staples' 'Energy']
```
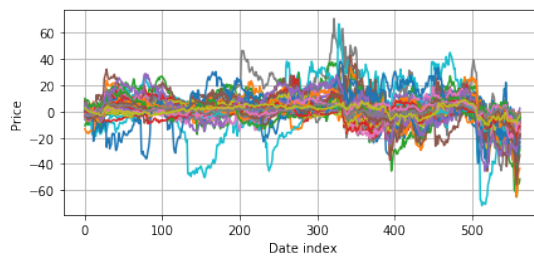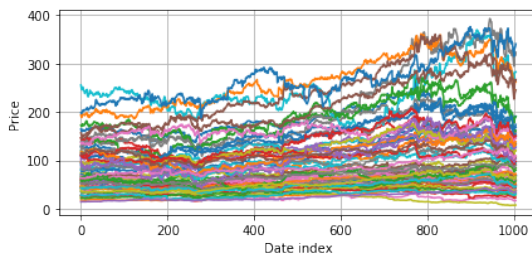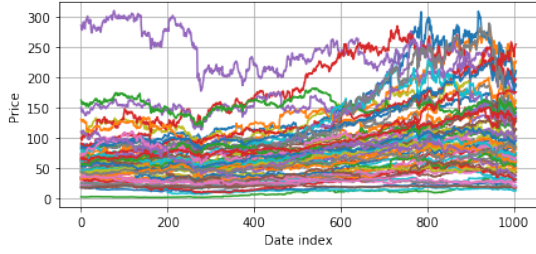


```
There are eigenvectors greater than 1: False
```

There are eigenvectors greater than 1: False

There are eigenvectors greater than 1: False

There are eigenvectors greater than 1: False

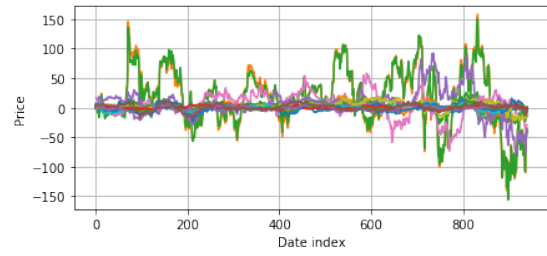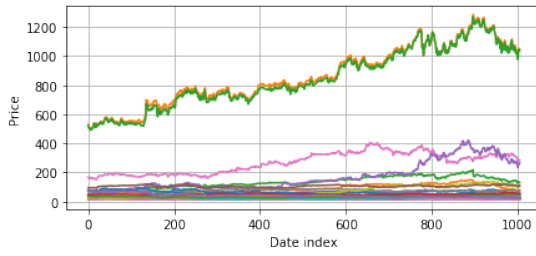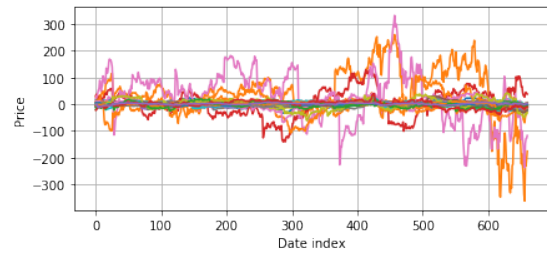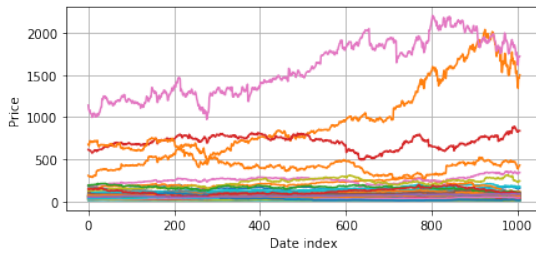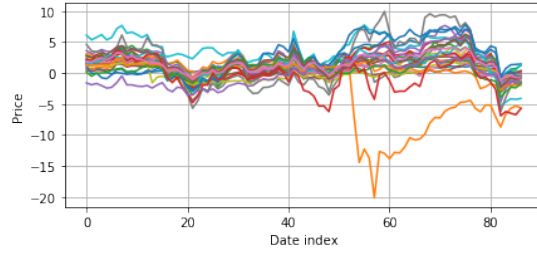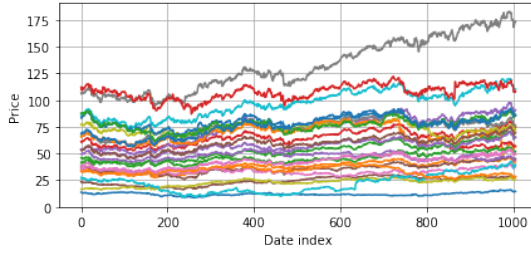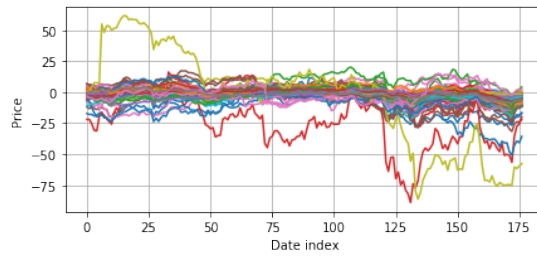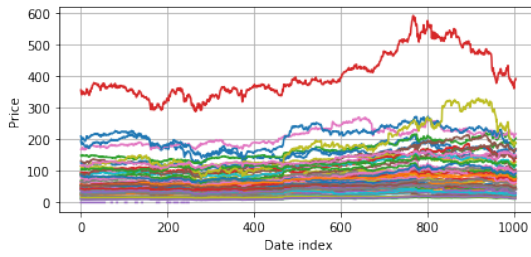There are eigenvectors greater than 1: False

15

There are eigenvectors greater than 1: False

There are eigenvectors greater than 1: True
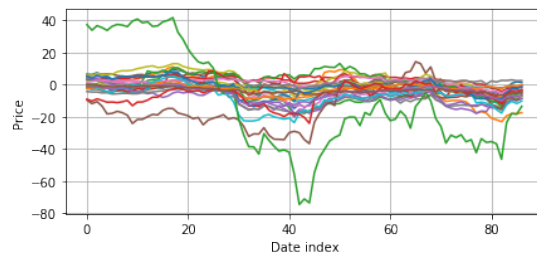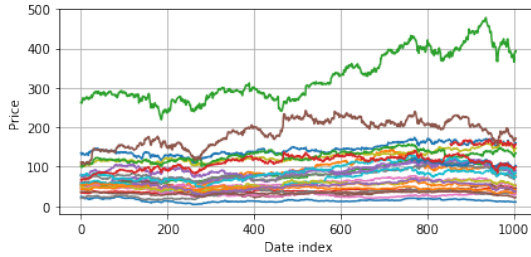
There are eigenvectors greater than 1: False
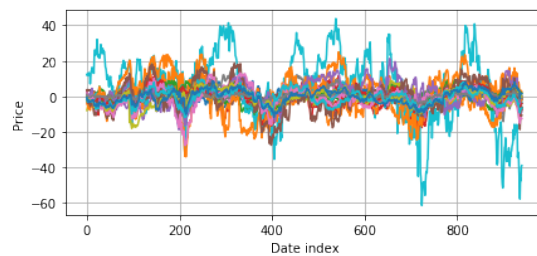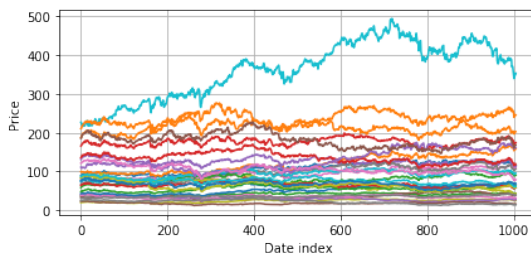
```
There are eigenvectors greater than 1: False
```



```
There are eigenvectors greater than 1: False
```
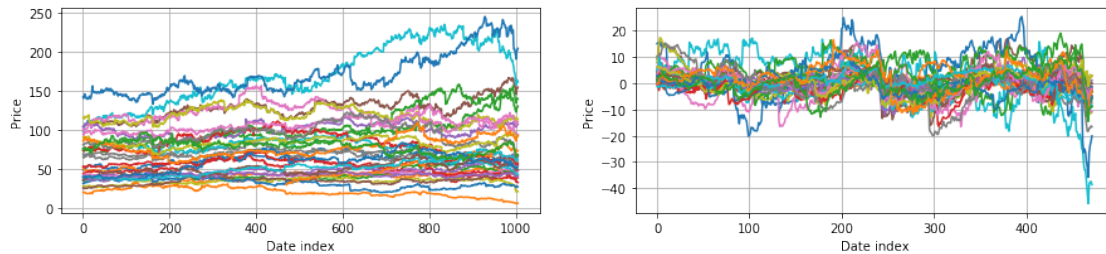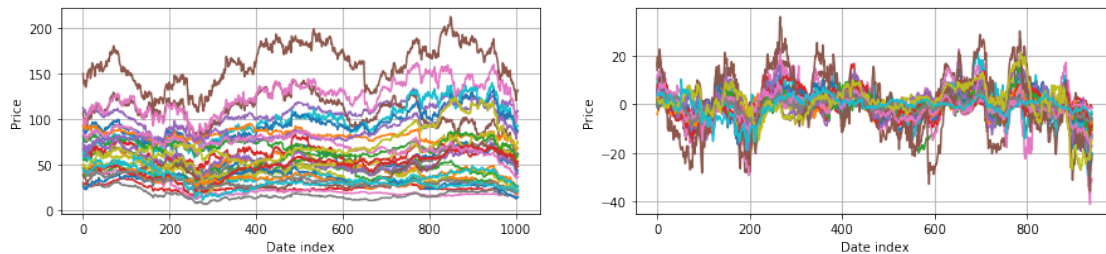


```
There are eigenvectors greater than 1: False
```

Discussion

As can be seen from the dataset, only the *Financials* industry has an absolute eigenvalue greater than 1. For all other industries, a VAR process could be suitable. If predictions are to be made of the returns, better results are likely to be achieved if the stocks are grouped into clusters.

In general it would be a good idea to group stocks by their sectors when calculating returns of a portfolio. This is because companies of different industries tend to behave similarly. Two terms often used in this context are the **defensive sector** and **cyclical sector**. The defensive sector includes companies that have a low price elasticity of demand, and include *Utilities* and *Computer Staples*. The remaining nine industries are grouped into cyclical sectors, as they tend to move up and down during different business cycles.

For modelling purposes it may be more convenient to group stocks together in such a way, for example by sector, such that the constructed model is stable to analyse. However as stated previously, a good portfolio is one that has some diversity to minimise risks. While grouping by sector allows easier matching to the objectives, it also increases the risk of investment.

2 Bond Pricing

This section explores fundamental techniques to set an appropriate valuation on bonds, including CAPM(Capital Asset Pricing Model) and APT(Arbitrage Pricing Theory).

2.1 Examples of bond pricing

1. Rate of returns were calculated as follows:

$$A = P \times (1 + \frac{r}{n})^{(nt)} \tag{9}$$

   Where P is the initial investmment, n is the number of times a year and r the annual interest rate

Compounding	Description	Answer
Annual	$1000 \times (1+r)^1$	$r = 0.1 \therefore r = 0.1$ or 10%
Semi-Annual	$1000 \times (1+\frac{r}{2})^2$	$\frac{r}{2} = \sqrt{1.1} - 1 \therefore r = 0.0976$ or 9.76%
Monthly	$1000 \times (1+\frac{r}{12})^{12}$	$\frac{r}{12} = \sqrt[12]{1.1} \therefore r = 0.0957$ or 9.57%
Continous	$1000 \times e^r$	$r = ln(1.1) \therefore r = 0.0953$ or 9.53%

2.

$$P \times (1 + \frac{0.15}{12})^{12} = P \times (e^r)$$
$$r = ln((1 + \frac{0.15}{12})^{12}) \approx 0.149 \tag{10}$$

3. The interest paid in **each quarter** can be calculated as follows

$$P \times e^{0.12} = P \times (1 + \frac{r}{4})^4$$
$$\therefore \frac{r}{4} = \sqrt[4]{e^{0.12}} - 1 = 0.03045 \tag{11}$$

2.2 Forward rates

Forward rates depict the interests earned on an agreement today to invest x from t to $t+1$:

$$(1 + y_{t-1})^{t-1}(1 + f_{t-1}) = (1 + y_{t-1})^t \tag{12}$$

This is therefore a contract today, at $t = 0$, of what the return must be between time $t - 1$ and t.

a) It depends. The received yeild from the first option is more variable if a forward contract has not been signed, as the interest rates may change. However if one has been agreed on, the invester should be netural about their decision in this aspect. In addition, it depends if the individual or company requires more liquidity - in which case the former option would be better.

b) What is initially known are the values of the short-term 5% per annum investment returns, and the longer term 7% per annum returns. The principal difference between the two lies in the uncertainty of the estimation, as the future estimate is not known. In the case that a forward rate agreement is used, the rate of interest for the following year with the one year investment would be the equivalent of 9% if the market sport rate is lower.

c) The liquidity preference hypothesis states that investors require a risk-premium to invest in long-term bonds. The main advantage of using the forward rates from a forward contract would be to hedge against risk. However there is a risk of default, and the originally set price for delivery could change. Also if the interest rates turn out to be more favourable (i.e. >9%) for the following year, it may not be possible to take advantage of that.

d) To match the one year investment, $r_1^2 = 1.05 \therefore r_1 = 1.025 \; or \; 2.5\%$

2.2.1 Duration of a coupon-bearing bond

a)

$$Duration = \sum_{t=1}^{7} t \times \frac{PV(C_t)}{PV} \tag{13}$$

$$Duration = 0.0124 + 0.0236 + 0.0337 + 0.0428 + 0.0510 + 0.0583 + 6.5377 = 6.7595$$

b)

$$Modified \; bond \; prices = volatitility(\%) = \frac{duration}{1 + yield} = -\frac{dPV}{dy}$$
$$Modified \; bond \; prices = volatitility(\%) = \frac{6.7595}{1 + 0.05} = 6.44 \tag{14}$$

Discussion

As the maturity increases, duration increases and the bond becomes more volatile in both cases. Duration is often used with bond portfolio risk with immunization strategies. However modified duration provides a measure of risk as it estimate the decline of the price of a bond if the interest rates increase.

c)

They are convinient measures because they measure the sensitivity of the bond price with respect to a change in the interest rates. The larger the value, the greater the interest rate risks. In the case of a pension, it usually has a long time to maturity and during this time interest rates are likely to change - therefore it has a long duration. If interest rates increase, the value of the bond will decrease and vice versa. Investors would demand higher yields for the risks associated with buying long term bonds.

This allows one to create an appropriate duration strategy (long-duration strategy) and take convenient measures to protect the pension plan against unexpected changes in interest rates. For example if it is likely that interest rates rise above what is expected in a yield curve, hedging duration would be appropriate.
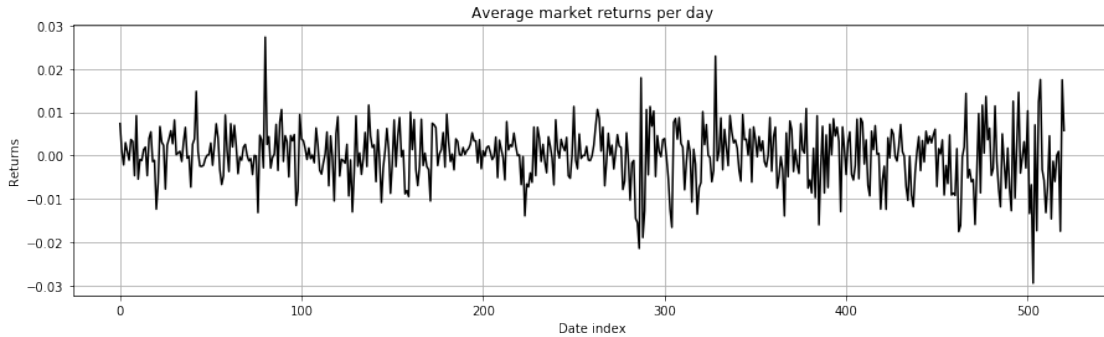
2.3 Capital Asset Pricing Model (CAPM) and Arbitrage Pricing Theory (APT)

The dataset has mcap and ret as keys, each with a shape of (521,157)

2.3.1 Estimating the market returns per day

```
All days are unique
There are 4220 nans in this dataset. But pandas aggregates ignores these
```


2.3.2 Estimating the the rolling beta: $\beta_{i,t}$

Estimating a rolling (sliding) beta, $_{i,t}$, for every company i, with the rolling window of 22 days

Beta measures the stocks of i in relation to the overall market. It is therefore commonly used as a measure of risk for individual assets in a portfolio.

$$\beta_{i,t} = \frac{cov(R_{i,t}, R_{m,t})}{var(R_{m,t})} \tag{15}$$

If, for example, there is no correlation between company i, and the market M, then $\beta_i = 0$.

Assuming the calculations are given by a 22 day rolling window, the following results are obtained.

Discussion

Some companies clearly have a much larger volatility than others (for example G_AALLN as compared to G_ACFP). It is also observed that negative values of $\beta_{i,t}$ are obtained for some companies (e.g. G_VIEFP). As the variance cannot be negative, this means that for such companies one observes a negative covariance and thus relationship of their returns with the market returns: $cov(R_{i,t}, R_{m,t})$.

2.3.3 Estimating the cap-weighted market return: R_m

$$R_m = \sum_i \frac{mcap_i \times ret_i}{\sum_i mcap_i} \tag{16}$$

```
Dates are the same, no need to apply where functions
```

Important to note: There are numerous NaNs in this dateset. For sum, Numpy considers these as zeros. Although they could be removed, it has been preferred in this report to use the default handling, until manual intervention is required.

Discussion

The weighting coefficient is the sum of the market caps for a particular day of all the companies: $\sum_i mcap_i$. One way that this can be interepreted is as the market value at that particular time. Market capitalisation is given by the stock price, of a listed company, multiplied by its outstanding shares. Therefore the values change in proportion to the price changes of each component.

2.3.4 Recalculating the rolling beta: $\beta_{i,t}$

Difference between the rolling beta and its weighted equivalent

Discussion

The purpose of calculating β would be to gather many uncorrelated assets (w.r.t M) and thereby giving a risk free asset with a deterministic rate of return. As it is correlated with the market it can be said to be a measure of systematic/market risk, one that cannot be reduced by diversifying.

Comparing the equally weighted β_m to the cap-weighted β_m, it can be seen that the same risk relationship are not observed in both. Not only do the calculated values differ - suggesting a different risk, they also do not maintain the same sign throught time consistently - suggesting a different relationship of systematic/market risk.

2.3.5 Arbitrage Pricing Theory

APT (Arbitrage Pricing Theory) assumes that an asset's returns are linearly variable to the asset's expected return and other variables. However, it does not assume markets to be perfectly efficient and some mispricings are assumed before the market corrects itself.

$$r_i = a + b_{m_i} R_m + b_{s_i} R_s + \epsilon_i \tag{17}$$

Therefore one estimates the beta coefficients using linear regression.

a) Estimating a, R_m, R_s

The multiple linear regression has been solved below, but a library implementations from Statsmodel (OLS) has alo been given below.

$$\mathbf{y} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & b_{m_1} & b_{s_1} \\ 1 & b_{m_2} & b_{s_2} \\ \vdots & \vdots & \vdots \\ 1 & b_{m_n} & b_{s_n} \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{bmatrix} a \\ R_m \\ R_s \end{bmatrix}}_{\beta} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} = \mathbf{X}\mathbf{fi} + \mathbf{ffl} \tag{18}$$

Using least squares, we would like to minimise the error:
$\sum_{i=1}^{n} \epsilon_i^2 = (\mathbf{y} - \mathbf{X}\mathbf{fi})'(\mathbf{y} - \mathbf{X}\mathbf{fi})$
Taking the derivative of this and setting it to zero by the Orthogonal Decomposition Theorem:

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{fi}_*) = \mathbf{0}$$
$$\implies \mathbf{X}'\mathbf{X}\mathbf{fi}_*) = \mathbf{X}'\mathbf{y} \tag{19}$$

Which gives the solution:

$$\beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{20}$$

The fitted values are given by:

$$\mathbf{y}_* = \mathbf{X}\mathbf{fi}_* = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$
$$\therefore \epsilon = \mathbf{y} - \mathbf{y}_* = (\mathbf{I} - \mathbf{H})\mathbf{y} \tag{21}$$

```
Removing nan columns, we now have 10 less columns in the dataset
Size of r_i (141, 1), b_mi: (141, 1), b_si: (141, 1)


The shapes of A, Rm and Rs are (500,), (500,), (500,) - as expected
```

b) Discussion: magnitude and variance

```
Regression param: a
        Mean: -0.004234460317229239, Var: 0.0016947602650338371
Regression param: R_m
        Mean: -0.00028848015311193397, Var: 6.182772556956221e-05
Regression param: R_s
        Mean: 0.0001933727660867127, Var: 2.9979687694277568e-06
```
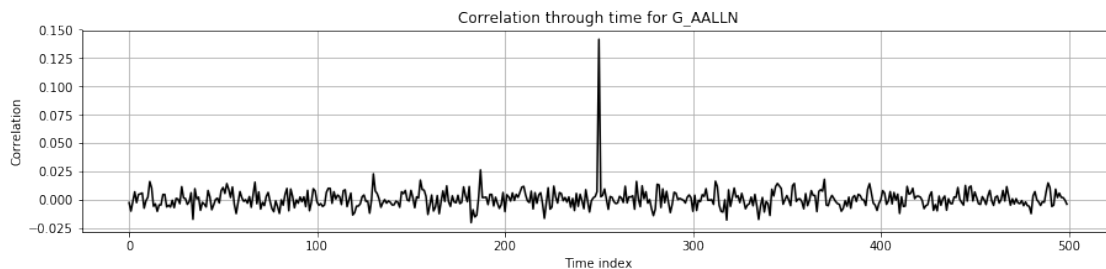
Evidently the betas have a very small value and variance, suggesting that they are not statistically significant in explaining the cross-sectional variation of average bond returns.

c) Temporal domain: finding the correlation through time for every company

d) Calculating the covariance from R_m and R_s
Rolling covariance with a window of 22 days. Here each matrix is of size 2×2
An aggregate covariance matrix of size 2×2

```
The determinant of the aggregate matrix is 1.7895522560516342e-10, and has eigenvalues: [6.207
```

Discussion

The determinant is almost close to zero, indicating that the two returns: R_m and R_w are almost perfectly correlated. The covariance matrix is stable as the eigenvalues are $<< 1$.

Analysing the covariance of the specific returns

For this question, please note that some companies have been dropped that contained NaNs beyond those that occurred due to the rolling windows. Therefore we deal with 141 companies.

$$\mathbf{E} = \begin{bmatrix} \epsilon_{1,t} & \epsilon_{1,t+1} & \cdots & \epsilon_{1,t=500} \\ \epsilon_{2,t} & \epsilon_{2,t+1} & \cdots & \epsilon_{2,t=500} \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_{k,t} & \epsilon_{k,t+1} & \cdots & \epsilon_{k,t=500} \end{bmatrix}_{500x141} \tag{22}$$

```
The size of the matrix is (141, 141) due to the reasons discussed above
```

The variances are the diagonals of the covariance matrix. When performing PCA, eigenvalues are used as the variances

```
The first component gives 7.3436811440454886% as the percentage of the variance explained
The first five components gives 23.41641161132465% as the percentage of the variance explained
```

Discussion

The amount of variance of the errors explained by the first component is 7.3%. This suggests that not a significant proportion of the data can be reduced through one principal component.

One further analysis that can be done is to examine the percentage of the variance explained by the errors for the returns. Due to the small beta values obtained earlier, it is hypothesised that a high percentage will be obtained.

```
The errors explain 66.10735579824356% of variance
```

As expected, a very large proportion of the variance of the returns was explained by the errors. This corroborates our earlier findings.

3   Portfolio Optimization

The purpose of the section is to select the best combinations of assets and equities out of the subset being considered.

3.1   Section 3.1: Adaptive minimum-variance portfolio optimization

3.1.1   Lagrangian optimization

For the purpose of all matrix derivatives used in this report, we refer to the matrix cookbook Section **??**.

Consider the optimisation problem formed when naively minimising the variance:

$$min_{\mathbf{w}}J(\mathbf{w}, \mathbf{C}) = \frac{1}{2}\mathbf{w^T C w}$$
$$st.\mathbf{w^T 1} = 1 \tag{23}$$

This gives the Lagrangian

$$min_{\mathbf{w},\lambda}J(\mathbf{w}, \mathbf{C}, \lambda) = \frac{1}{2}\mathbf{w^T C w} - \lambda(\mathbf{w^T 1} - 1) \tag{24}$$

In order to solve for the optimal weights, we first consider the partial derivates of the lagrangian.

$$(1)\frac{\partial L}{\partial w} = Cw - \lambda\mathbf{1} = 0$$
$$(2)\frac{\partial L}{\partial \lambda} = w^T\mathbf{1} - 1 = 0 \tag{25}$$

Now rearranging them for substitution:

$$(1)Cw = \lambda\mathbf{1}$$
$$\implies w = \lambda C^{-1}\mathbf{1} \tag{26}$$
$$(2)w^T\mathbf{1} = 1$$

Using the rearranged equations and combining (1) and (2):

$$\lambda\mathbf{1}^T C^{-1}\mathbf{1} = 1$$
$$\therefore \lambda = \frac{1}{\mathbf{1}^T C^{-1}\mathbf{1}}$$
$$\textit{Using the rearranged version of}(1) \tag{27}$$
$$\therefore w^* = \frac{C^{-1}\mathbf{1}}{\mathbf{1}^T C^{-1}\mathbf{1}}$$

Now consider the variance: $var(w_*)$, given by $w^T C w$

$$w_*^T C w_* \implies \frac{1}{(\mathbf{1}^T C^{-1}\mathbf{1})^2}\mathbf{1}^T C^{-1}CC^{-1}\mathbf{1}$$
$$\textit{Because } \mathbf{1}^T C^{-1}\mathbf{1} \textit{ is a scalar} \tag{28}$$
$$\therefore var(w^*) = \frac{1}{\mathbf{1}^T C^{-1}\mathbf{1}}$$

3.1.2 Minimum Variance Portfolio

Calculating w^*
   Note the expected shapes are highlighted below: * train is 261 x 157
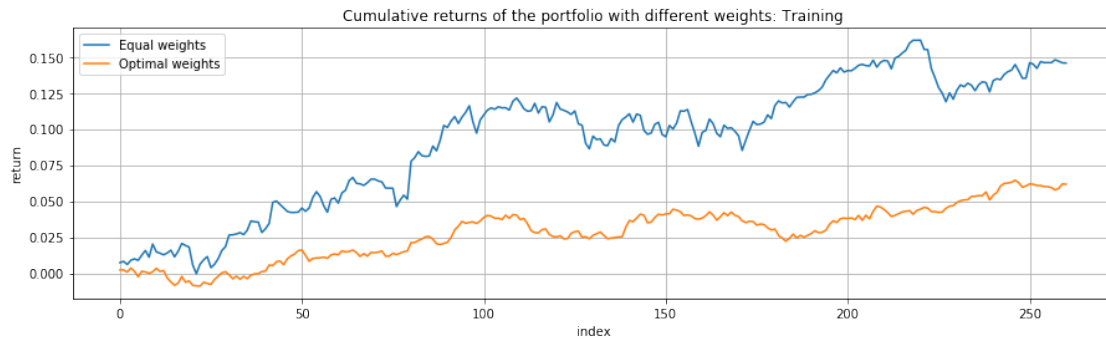* w is expected to have one per company, which is 157 x 1
* C will be a 157 x 157 matrix as a result
* the size of vector one must be 157 x 1
   Comparing the performance of the optimal weights with equally an weighted portfolio

The variance of the portfolio using the the minimum variance portfolio weights was [[4.0826258
and using the equally weighted weights was [[2.56503088e-05]]

Performance on the test data

The variance of the portfolio using the the minimum variance portfolio weights was [[1.1712231
and using the equally weighted weights was [[1.56125946e-05]]
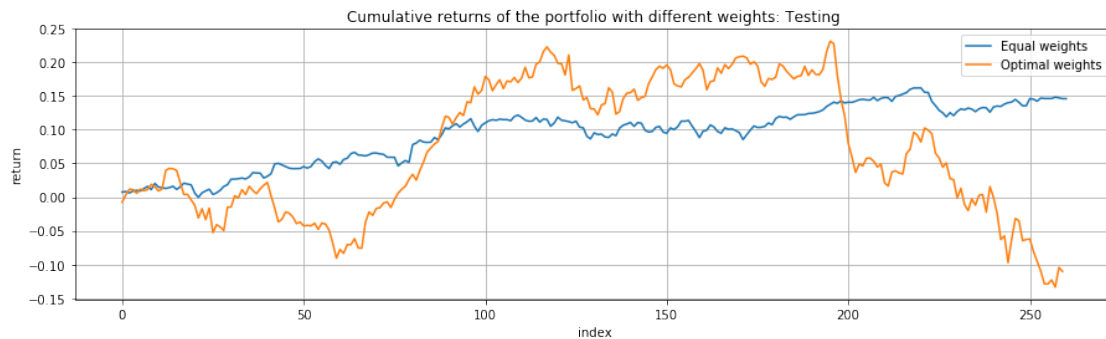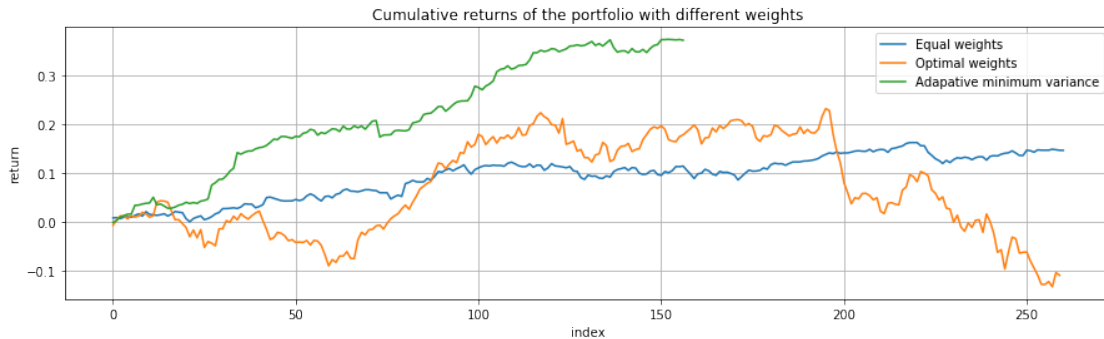

Discussion

From the calculated variance above, it can be seen that a smaller variance was achieved using the weights calculated from the training data. This is precisely the cost functions that we were minimising. However another measure of performance would be the total returns achieved by the portfolio. From the graph above, it can be seen (unexpectedly) that higher cumulative returns are achieved using equal weights as compared to optimal weights.

One possible explanation of this considers the components of excess returns more closely. These returns may be broken down into residual and systematic components. The systematic part is what depend on the benchmark excess return, as it is computed by multiplying beta with the benchmark returns. In the study Section **??**, it was found that a portfolio constructed by equal weights has higher systematic returns due to its "relatively higher exposure to the market, size and value factors". The α is also higher, as can be observed in the graph above, due to a monthly rebalancing strategy that exploits the time-series and cross-ssectional properties of the returns of

stocks. A more complete discussion of the reasons as to why an equally weighted portfolio per-
forms better than the minimum-variance optimised weights can be found in Section **??** and Section
??.

Adaptive time-varying minimum variance portfolio   One choice that had to be made in the
calculation of a time-varying minimum variance portfolio was an appropriate window size.

Discussion

The model can be thought of as one that adapts by observation. Approaches considered be-
forehand assumed that the disturbances: ϵ_t are independent, however here a full correlation be-
tween the disturbances is used here. Therefore information of volatility and correlation should
propogate through the model.

From the variance graph of this part of the question, it is observed that this portfolio approach
has the same variance as that of equal weights. However the returns and excess returns are larger,
as is seen in the two following graphs.

A robust portfolio is then one that is designed to optimize the worst-case per- formance within
the set of values for the mean and for the covariance matrix in the corresponding uncertainty set.
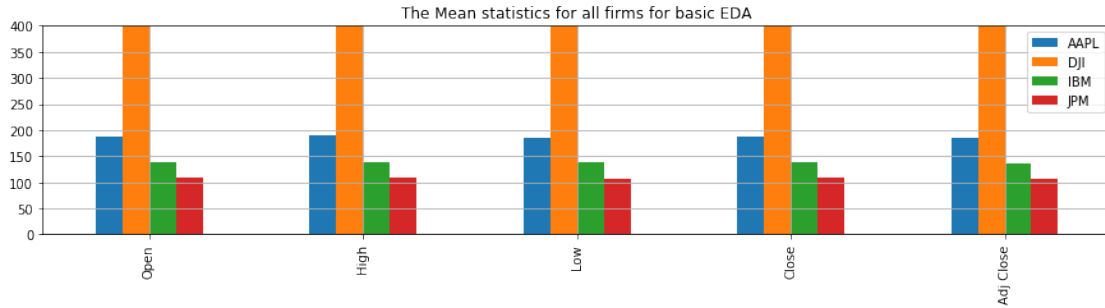
Global Minimum Variance portfolio

equally weighted data, it does not exhibit the current state of the market. It reflects market
conditions which are no longer valid by assigning equal weights to the most recent and the most
distant observations.
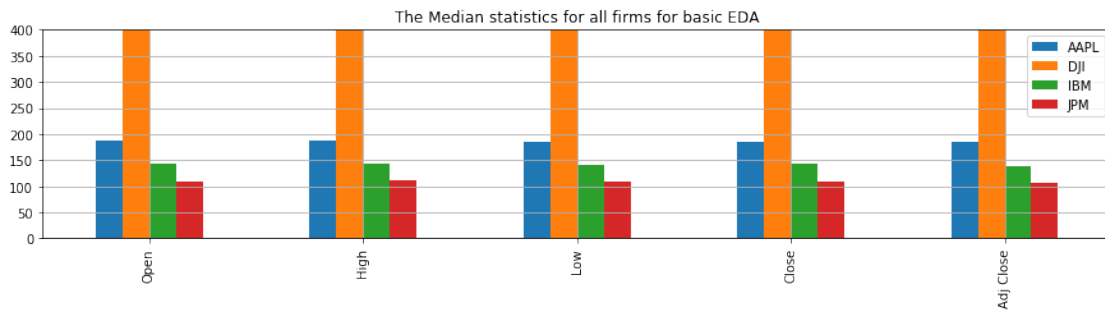
4   Robust Statistics and Non Linear Methods

This section of the notebook explores the models and techniques used to increase the strength of
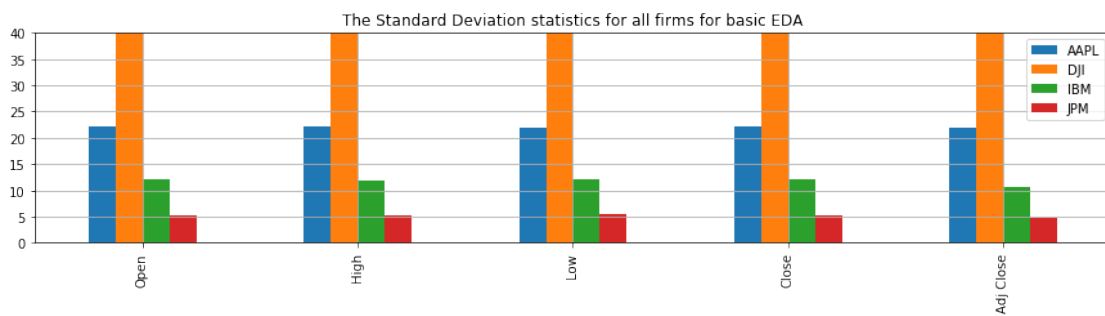robust statistics.

4.1   Exploratory Data Analysis

Load the datasets

The Mean statistics for all firms for basic EDA

Warning, DJI has not been fully plotted

The Median statistics for all firms for basic EDA

Warning, DJI has not been fully plotted

The Standard Deviation statistics for all firms for basic EDA

Warning, DJI has not been fully plotted

|| EDA for AAPL||

```
---- Basic Dataset checks ----
Number of missing data: 1

---- Specific to stocks statistics ----
The maximum closing price for the firm: 232.07000699999998

|| EDA for IBM||

---- Basic Dataset checks ----
Number of missing data: 1

---- Specific to stocks statistics ----
The maximum closing price for the firm: 160.91000400000001

|| EDA for JPM||

---- Basic Dataset checks ----
Number of missing data: 1

---- Specific to stocks statistics ----
The maximum closing price for the firm: 118.629997

|| EDA for DJI||

---- Basic Dataset checks ----
Number of missing data: 1

---- Specific to stocks statistics ----
The maximum closing price for the firm: 26828.390625
```
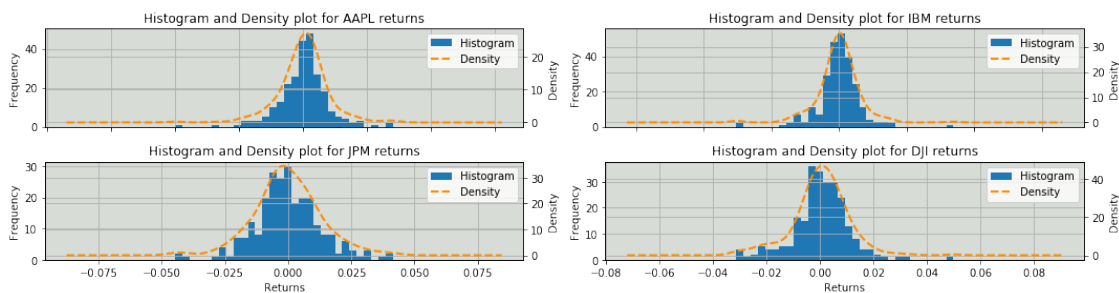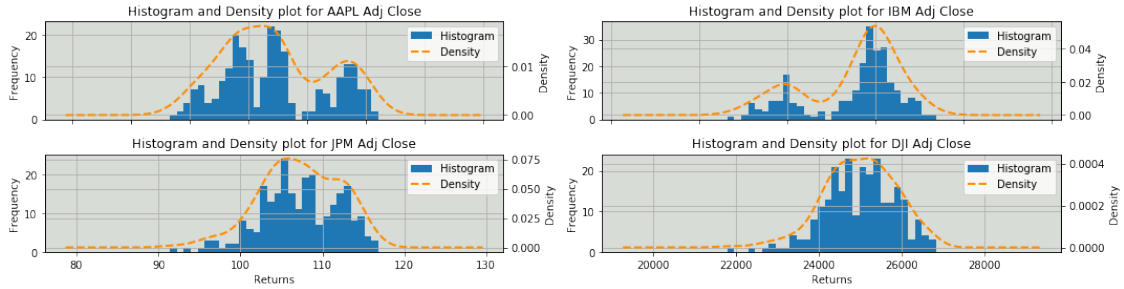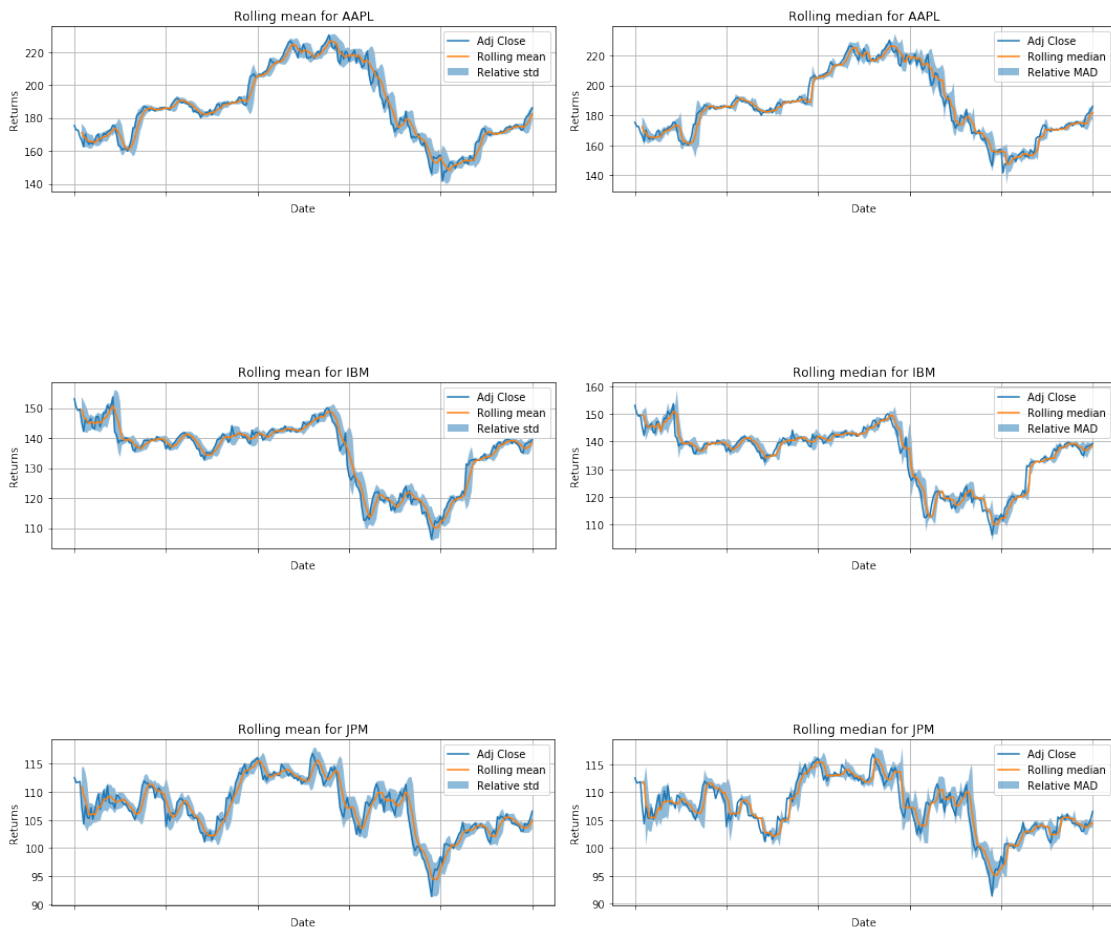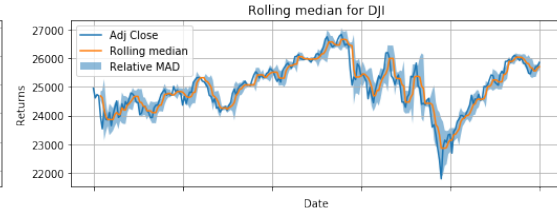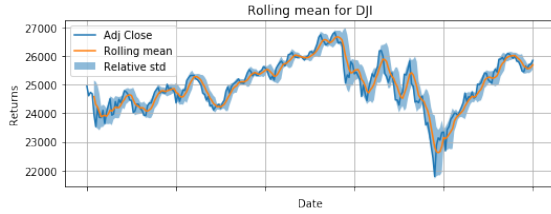
Histogram and Density plot for AAPL Adj Close | Histogram and Density plot for IBM Adj Close
Histogram and Density plot for JPM Adj Close | Histogram and Density plot for DJI Adj Close

Discussion

It is observed from the analysis that the density of the Adj Close is much wider than that of the returns. Therefore the standard deviation for the former is larger than the latter. Returns also appear to be more closely resembled by a normal distribution than Adj Close, although the tails are fatter, suggesting randomness. Interestingly Adj Close shows a linear combination of two such distributions. Most returns are observed to be skewed slightly towards a positive return.
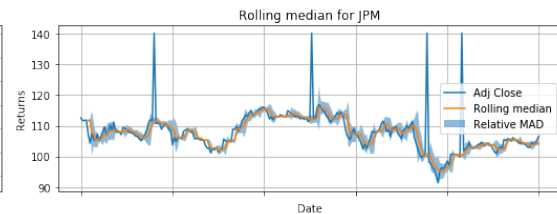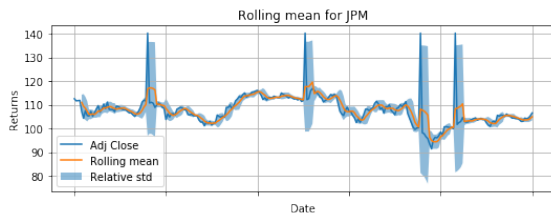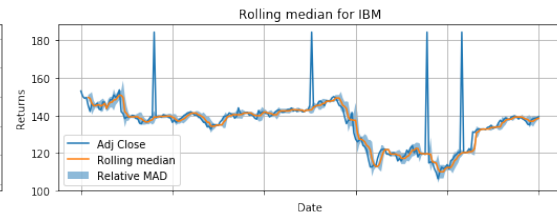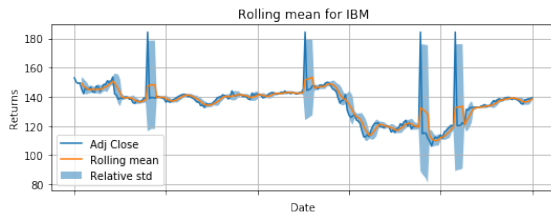

Rolling mean for AAPL | Rolling median for AAPL
Rolling mean for IBM | Rolling median for IBM
Rolling mean for JPM | Rolling median for JPM

Rolling mean for DJI / Rolling median for DJI

Discussion

Mean Absolute Deviation was first introduced by Konno & Yamazaki (1991) as a risk measure. The idea behind it was to eliminate the assumptions of gaussian returns, thereby minimising complexity. Quartiles do not take into account every score in the dataset.

In most cases similar values were obtained from MAD and std. However it can be seen that in the case that there is a large deviation at some point, such as that observed DJI, MAD is less effected than Std.

Note: it will not be possible to use MAD when the units of the prediction errors differ to the original series, unlike std

Introducing outliers

Rolling mean for AAPL / Rolling median for AAPL

Rolling mean for IBM / Rolling median for IBM

Rolling mean for JPM / Rolling median for JPM

30
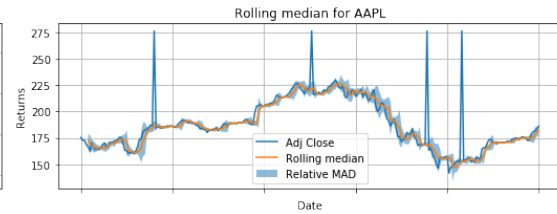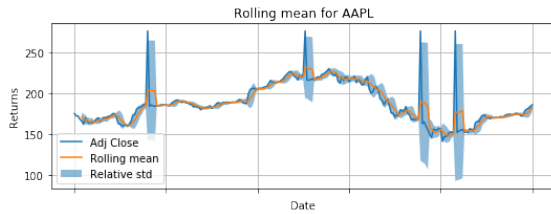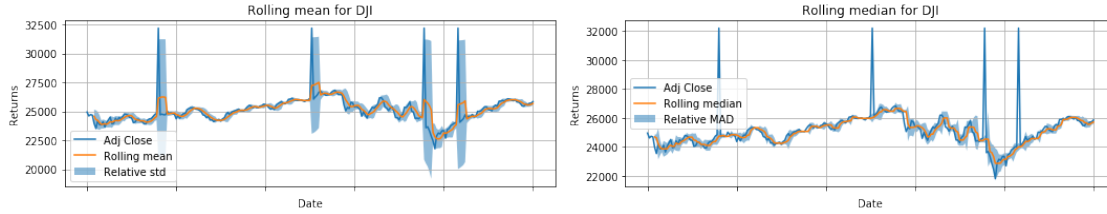
Discussion

By introducing oultiers, it can be seen that the median absolute deviation is less sensitive to such outliers than standard deviation is. This is shown by the width of the *blue* fill in the diagrams above. The relative standard deviation also sees delayed jumps where there are outliers placed. Considering the results it can be argued that while for an entire population the standard deviation works well, it does not give a good indication for smaller sample sizes (such as the window size of 5 used here). This concludes that Median Absolute Deviation is a more robust metric to use.
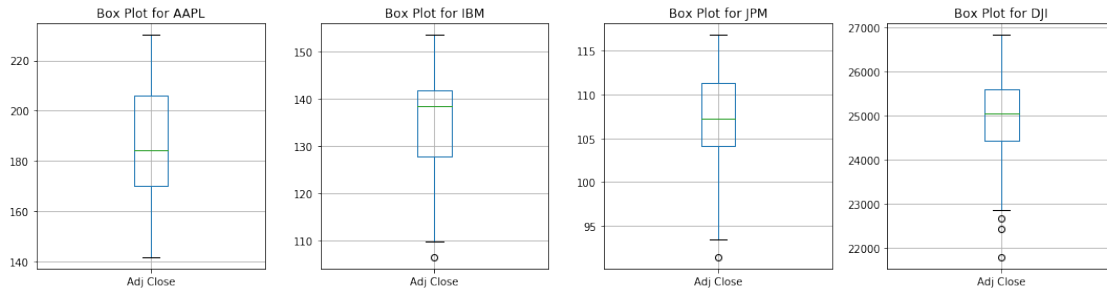
On the same axis (excluding DJI due to its larger values)
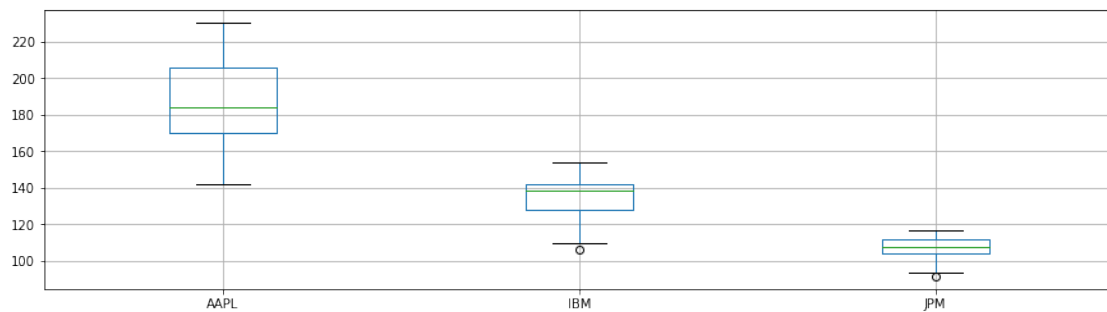
Discussion

The boxplot conveys the range, interquartile range and median. By plotting them on the same axis, it is possible to draw conclusions from these basic statistics with relative ease. For example one may expect lower returns from JPM (from the median and range), however the uncertainty of the value of return is also much smaller (from the IQR).

4.2 Robust Estimators

Custom implementations
   Library Implementations
   Computational Efficiency Comparison

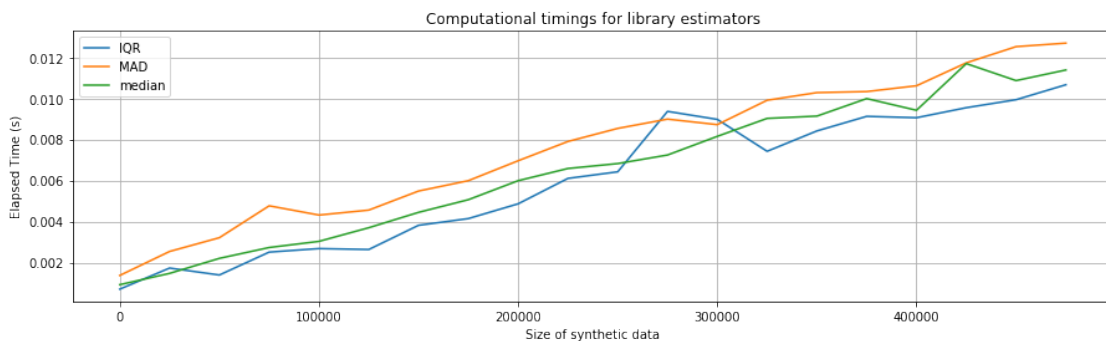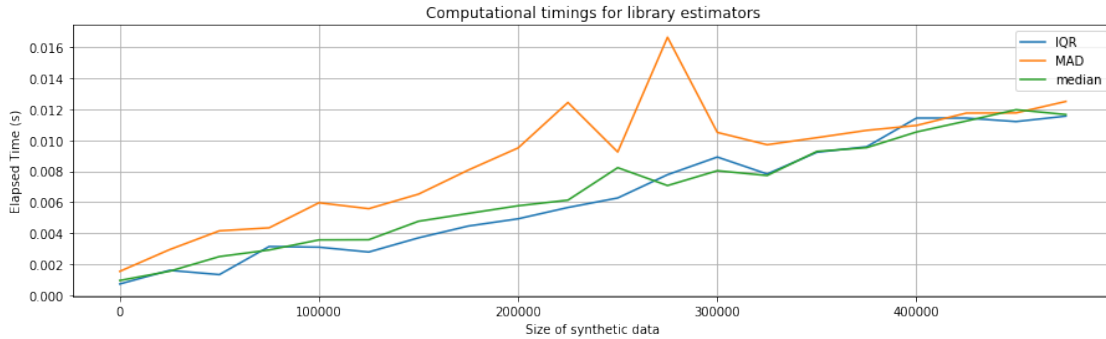Clearly MAD is the least efficient computationally, followed by IQR and median, in a custom as well as library implementation. This is due to the fact that MAD relies on performing the median twice, and IQR does the same but for half the dataset - thereby explaining the results obtained.

In terms of a theoretical computational complexity, the custom implementation differs to that used by *Pandas*. The defined implementaiton uses sorting and simply returns the median, giving $O(nlog(n))$ complexity. However a better implementation exists, which requires a linear $O(n)$ time - this is used by the library. As mentioned above, IQR and MAD will simply be linear combinations of this complexity, meaning an $O(nlog(n))$ for the custom definition and $O(n)$ for the library implementation.

Comparing the breakdown points for each estimator

As has already been seen, IQR is heavily affected by outliers. The break down point, as defined by (Hampel, 1971; Hampel, 1974), is the smallest percentage of outliers than can cause an estimator to take arbitrary large values.

The breakdown point of IQR is approximated to be around 25%. For the median absolute deviation, it is approximated at 50% - which is the same as median (robust locator estimator). Note that the median, therefore, requires the highest possible percentage of outliers to reach the breakdown point. Thereby making it more 'robust' than mean.

4.3 Robust and OLS regression

OLS regression

```
Fitting OLS to the stock: AAPL
Model score (R^2): 0.00013086731455969058
Model coefficients: [[-2.10330403e-06]]

Fitting OLS to the stock: IBM
Model score (R^2): 0.005918848314793057
Model coefficients: [[1.13916025e-05]]

Fitting OLS to the stock: JPM
Model score (R^2): 0.0005628951686514894
Model coefficients: [[2.95455942e-06]]

Fitting OLS to the stock: DJI
Model score (R^2): 0.0005628672758266617
Model coefficients: [[2.36497271e-06]]
```

Hubber regression

```
Fitting Hubber to the stock: AAPL
Model score (R^2): -0.001191390293084682
Model coefficients: [1.31714475e-09]

Fitting Hubber to the stock: IBM
Model score (R^2): -0.0020195690436943803
Model coefficients: [1.31722802e-09]

Fitting Hubber to the stock: JPM
Model score (R^2): -0.00013694029477084868
Model coefficients: [-3.88148858e-10]

Fitting Hubber to the stock: DJI
Model score (R^2): -0.0020304988248793787
Model coefficients: [9.08325237e-10]
```

From the R^2 score, it seems OLS performed better than Hubber regression. Hubber has a negative model score, meaning it gets arbitrarily worse. However the values are relatively close to 0, suggesting that it is a constant model.

Dataset with outliers

As in the earlier section, introduce outliers to dreturns

```
Fitting OLS to the stock: AAPL
Model score (R^2): 0.00011417728914719927
```

```
Model coefficients: [[2.22432054e-06]]

Fitting OLS to the stock: IBM
Model score (R^2): 0.006396847910867587
Model coefficients: [[1.53201307e-05]]

Fitting OLS to the stock: JPM
Model score (R^2): 0.0010108760335360012
Model coefficients: [[4.38339328e-06]]

Fitting OLS to the stock: DJI
Model score (R^2): 0.0018450602545707626
Model coefficients: [[5.23425218e-06]]

Fitting Hubber to the stock: AAPL
Model score (R^2): -0.0002944066850156801
Model coefficients: [2.28463051e-09]

Fitting Hubber to the stock: IBM
Model score (R^2): -0.0017193471868854893
Model coefficients: [1.47611153e-09]

Fitting Hubber to the stock: JPM
Model score (R^2): -0.002331485938392941
Model coefficients: [2.66418451e-10]

Fitting Hubber to the stock: DJI
Model score (R^2): -0.0006692098737812735
Model coefficients: [1.77103669e-09]
```

Discussion

When considering R^2 score, no difference was observed in this case. Theoretically it is expected that for the case of outliers, regression scores for (the most robust technique) Hubber should be better than OLS. One observes this when comparing the changes in the coeffecients to the datasets with and without outliers.

4.4   Robust Trading Strategies

Moving Average Cross Over for stock dataset
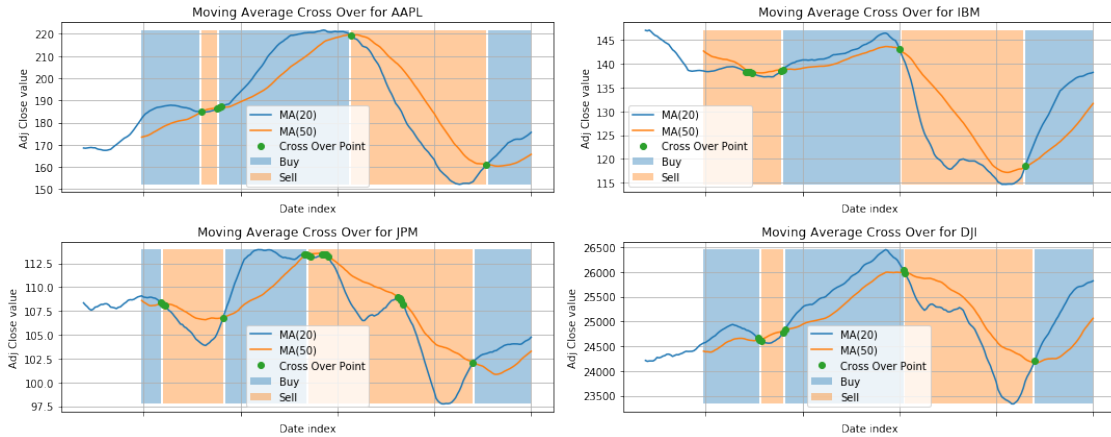
Moving Average Cross Over for outliers introduced dataset
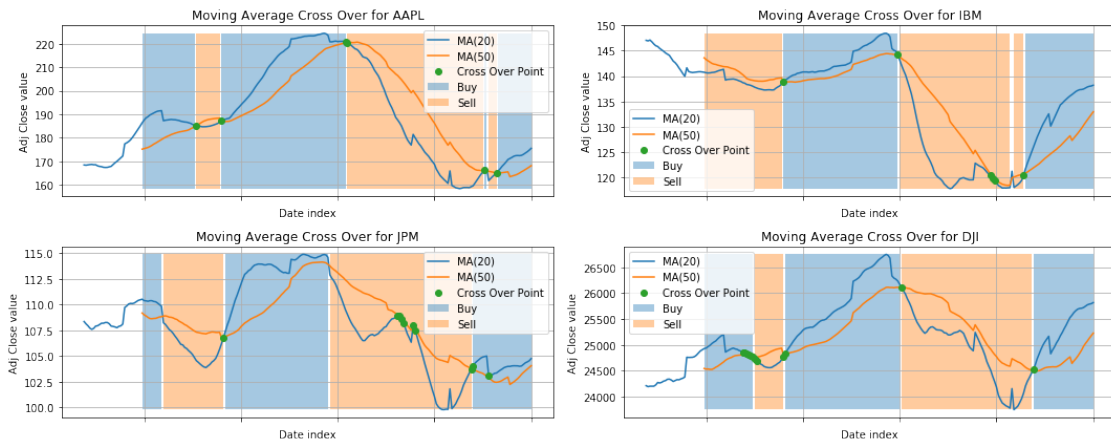
Discussion

In the case of IBM, it can clearly be seen that the cross over points are effected by the introduced outliers.

Moving Average Cross Over for stock dataset: Rolling median

Discussion

In general one observes that the rolling median is far less smooth that the rolling mean, which in this case can mean a higher frequency of bu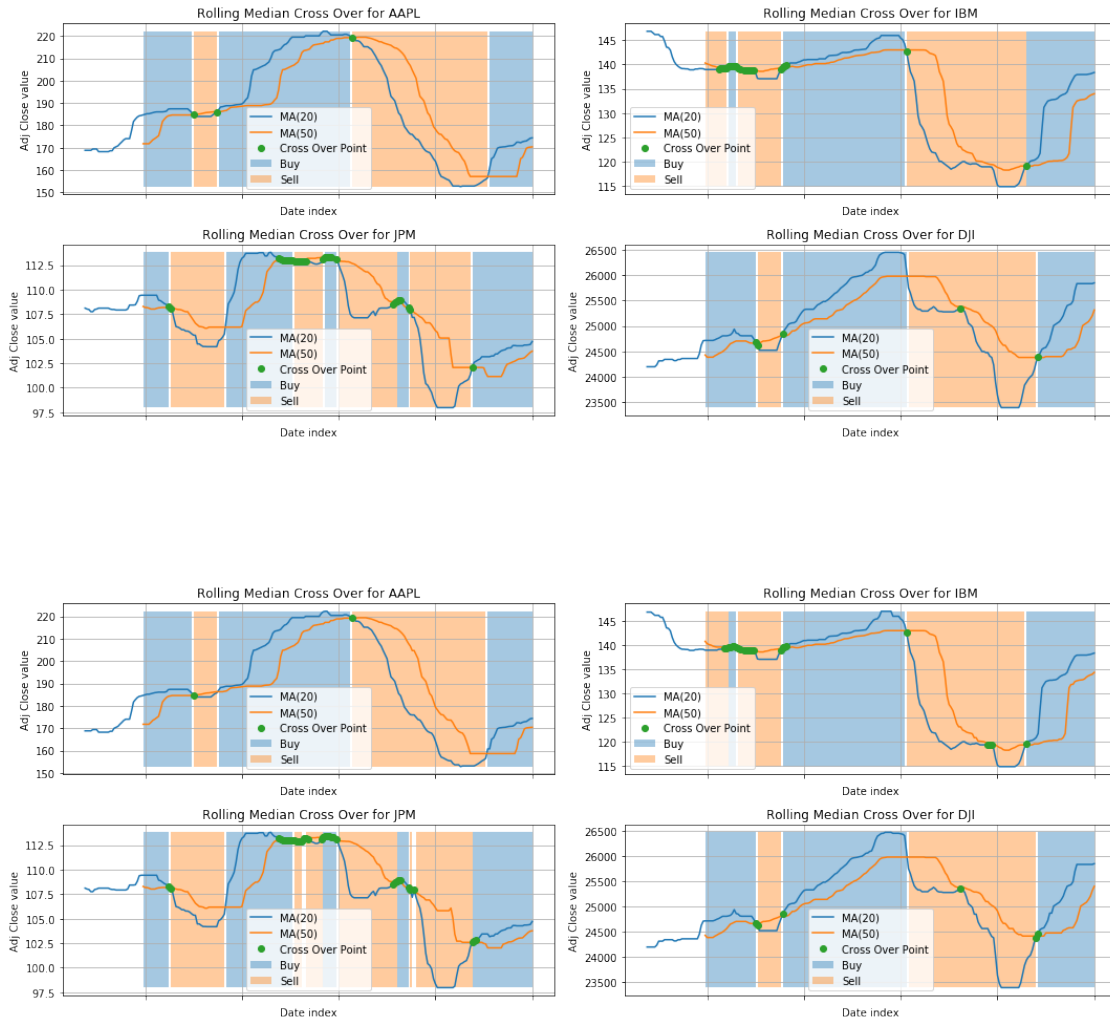ying and selling. In the case of IBM, it can be seen clearly that by using the rolling median the introduced outliers did not effect buying and selling decisions. However in the case of a dataset that has a more voltile decision outcome due to this strategy, the introduced outlier would have an effect on the median - as is observed for JPM.

4.4.1 Graphs in Finance

The task involves applying graphs to modelling and visualizing the relationship between stocks within the S&P 500 index. Please refer to Appendix 3 for a further investigation of the task that has been conducted.

1. Selecting 10 assets

```
There are 6064 missing values in the dataset
Dropping assets with nans...
```

For the purpose of this task, ten stocks with the highest market-cap of each sector were chosen. It is expected that a network graph for these will not demonstrate an evident relationship. The purpose here is to determine what sectors may be co-related to each other.

Appendix 3 demonstrates an example with the ten largest technology companies. Please refer to the analysis there to see a graph in which the nodes have a more clear relationship.

2. Constructing a network graph

An arbitrary threshold of 0.3 was used to filter out which links are to be plotted.
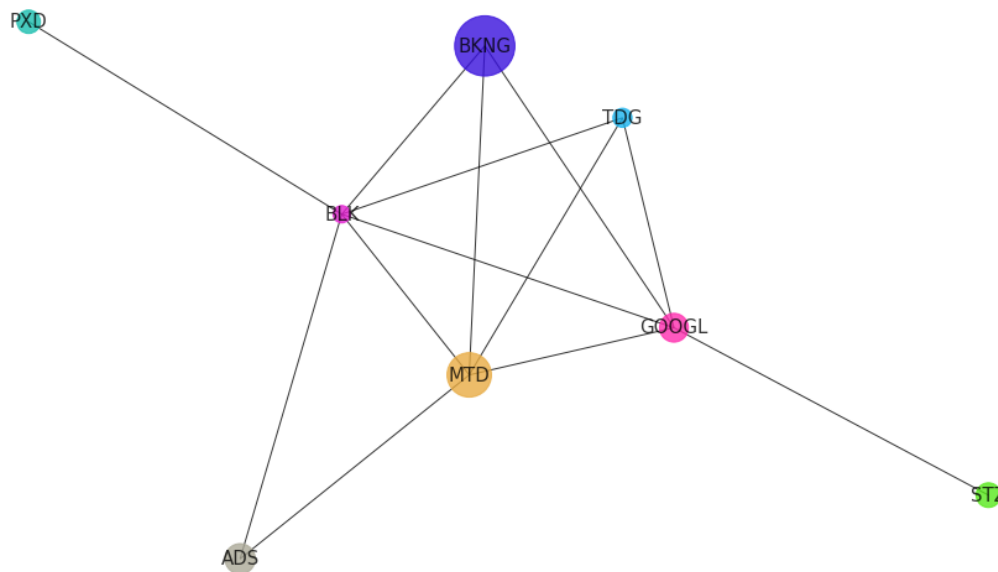
Although it is difficult to point out physical meaning behind the correlations observed, the graph does suggest that there might exist a relationship between sectors. Surpisingly it seems that (MTD) and (BLK) are more related to other sectors, perhaps suggesting more diversification than other companies - which would not be an entirely false conclusion.

Discussion: correlation matrix

The correlation matrix is used to create the network links. In an undirected graph as is drawn by the Networkx library here, each link to another has a certain probability. The probabilities, in the case of the stock datasets, are given by their correlations with respect to each other.

As expected, the correlation at the diagonal will be filled with 1's (as it is wrt the same stock). It goes without saying that the matrix will also be symmetric, as direction does not matter. A threshold (of 0.35) is applied in the first graph so that only correlations above that are shown as links.

3. Discussion of results

The topology of the graph will be affected by the nature, if a threshold is applied. That is to say if no threshold is used, then the same links will always be plotted (as in the second graph in the section above). However the topology is entirely dependent on the nature of the data in the case that a threshold is applied. This is because any change in the nature of the data could affect the measured correlation, and thereby changing the topology.

Continuing with this logic, reordering of the vertices should not have any effect on the topology as it does not affect the generated correlation matrix.

Reordering the time series data would drastically change the correlations as the underlying structure of the time series as a whole has changed. Thereby it would change the topology of the network graph.

4. Applying a different distance metric

Dynamic Time Warping (DTW) is a common distance metric for judging the correlations of time series. Unlike euclidian distances or correlation measures, DTW considers axis offsets - making it a more robust technique for measuring distance.

A matrix $d_{i,j}$ is constructed that best aligns the two sequences, where each element (i, j) is an alignment of point a_i, b_j. The mapping from one sequence: A to the other B is defined by a warping path, $W = w_1, w_2, ..., w_K$. Element w_k has to be adjacent to w_{k-1} in order to choose adjacent element in the defined path. The cost function therefore involves selecting a path that minimises the cost of warping:

$$DTW(X, Y) = min(\sqrt{(\sum_{k=1}^{K} w_k)}) \tag{29}$$

Another propery about DTW is that it is also symmetric, as correlation and euclidian distances, making it appropriate for an undirected graph.

```
/Users/Divyansh/anaconda3/envs/mne/lib/python3.6/site-packages/ipykernel_launcher.py:2: FutureW
```

```
/Users/Divyansh/anaconda3/envs/mne/lib/python3.6/site-packages/networkx/drawing/nx_pylab.py:61
  if cb.is_numlike(alpha):
```
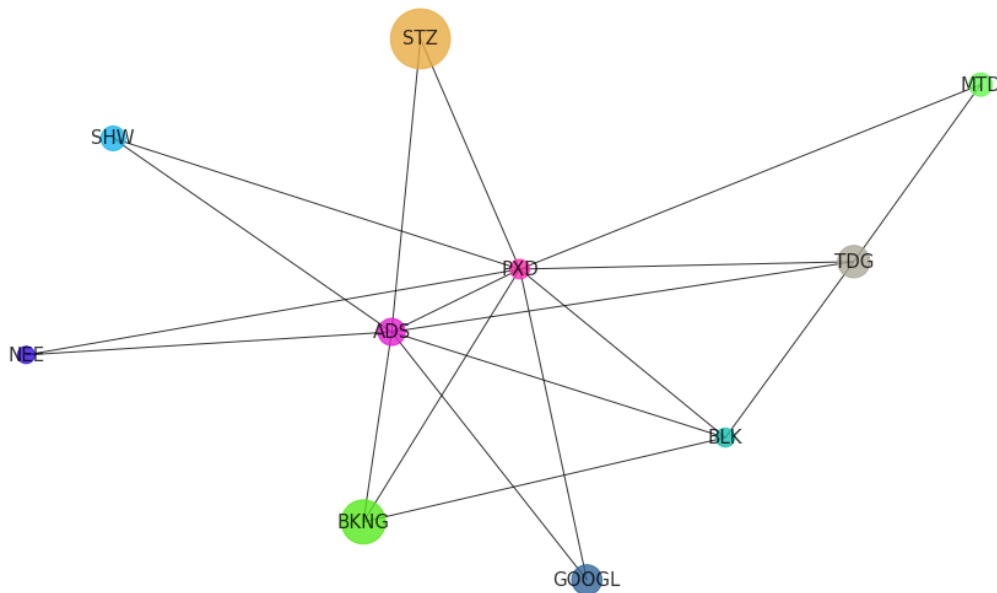
5. Discussion: Consider raw prices

In the case that raw prices are used instead of the logarithmic returns, one principal difference expected is the strength and relationships derived from the correlations. Significantly different correlations would be observed between the time series' as their values are not measured in percentages. The correlations will be higher, due to the pseudo random walk nature of stocks. For this reason it can be argued that the correlations of prices would be less meaningful than log returns. In terms of the topology of the graph, this would mean that everything seems to be better connected/related to each other.

To understand this from a mathematically sound perspective, let $P_t = \{P_0, P_1, P_2 ... P_T\}$

$$\implies R_t = P_t - P_{t-1}$$
$$\therefore P_t = \{P_0, P_0 + R_1, P_0 + R_1 + R_2, ..., P_0 + R_1 + \cdots + R_T\} \tag{30}$$

From this it can be seen that a correlation of raw prices is such that early changes (R_1) will have a greater impact that later changes (closer to R_T) in calculations. Therefore making it sesitive to the size of the x-axis (the time periods). However computing the correlation on returns will provide an equal weighting.
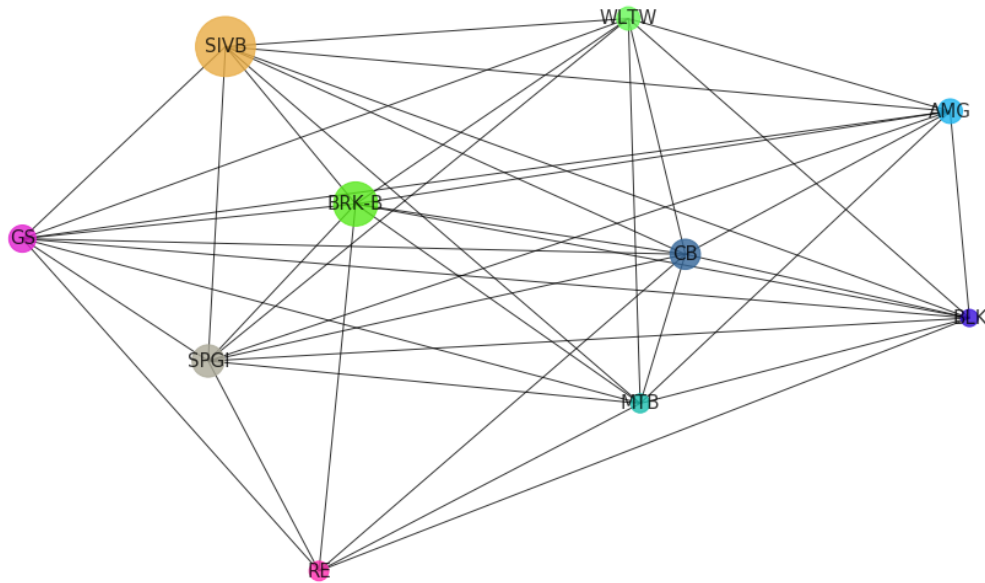
A   VAR model summary

B   VAR autocorrelation plots

```
<Figure size 432x288 with 0 Axes>
```

ACF plots for residuals with $2/\sqrt{T}$ bounds
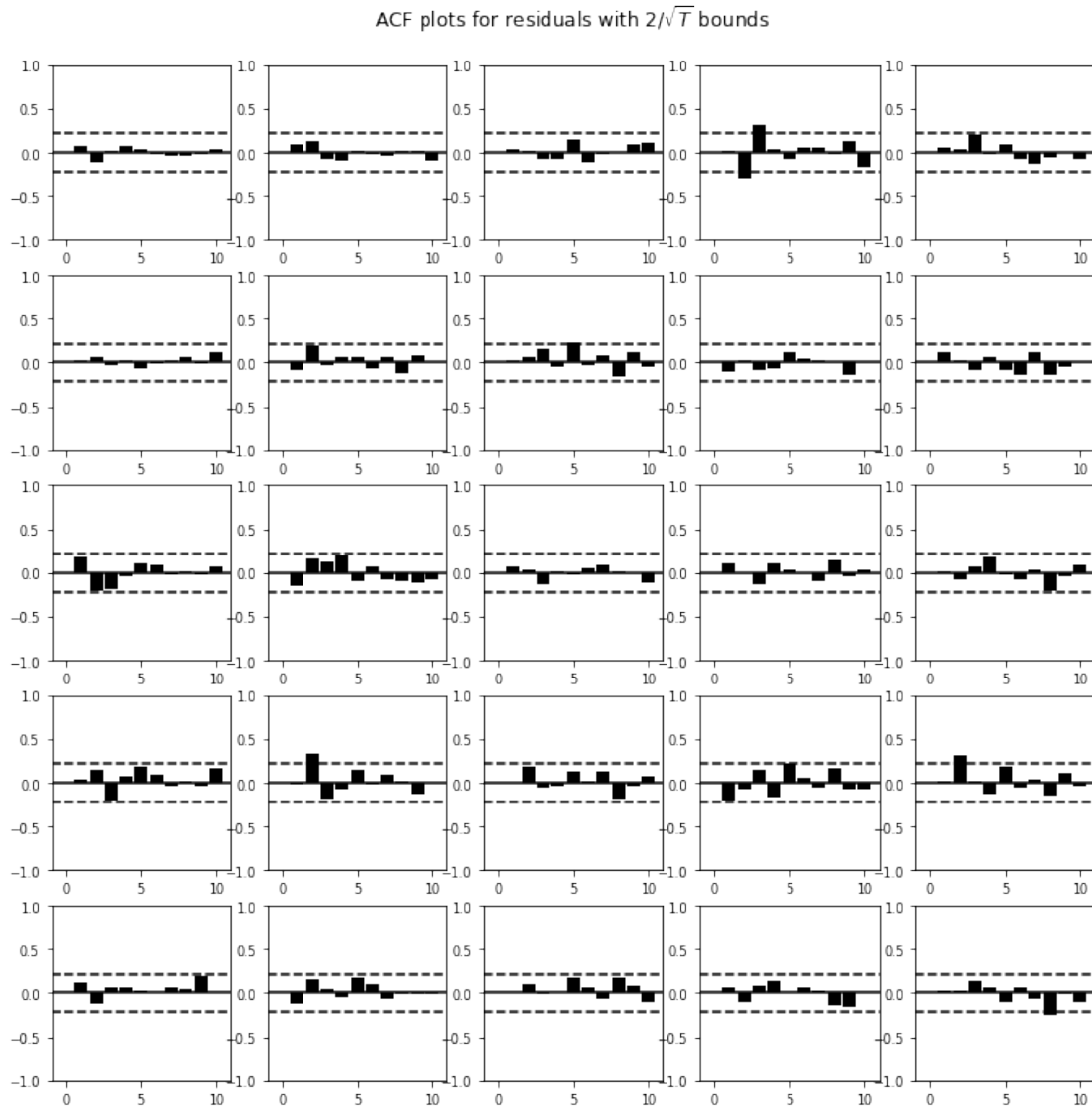
C   Graph relationships of the ten largest technology companies

```
/Users/Divyansh/anaconda3/envs/mne/lib/python3.6/site-packages/networkx/drawing/nx_pylab.py:61
  if cb.is_numlike(alpha):
```

Using the same threshold for building links, all companies seem to be much more correlated - as would be expected. Rather surprisingly, Adobe seems to be most correlated with the other technology giants - whilst Arista Networks has the least.

C.1   Bibliography

[1] Mikusheva, A. (2007). Stationarity, Lag Operator, ARMA, and Covariance Structure. [ebook] Available at: https://ocw.mit.edu/courses/economics/14-384-time-series-analysis-fall-2013/lecture-notes/MIT14_384F13_lec1.pdf [Accessed 26 Feb. 2019].

[2] Hudson, R. and Gregoriou, A. (2015). Calculating and comparing security returns is harder than you think: A comparison between logarithmic and simple returns. International Review of Financial Analysis, 38, pp.151-162.

[3] Meucci, Attilio. (2010). Quant Nugget 2: Linear vs. Compounded Returns – Common Pitfalls in Portfolio Management. GARP Risk Professional.

[4] (2006) Vector Autoregressive Models for Multivariate Time Series. In: Modeling Financial Time Series with S-PLUSő. Springer, New York, NY

[5] Plyakha, Yuliya & Uppal, Raman & Vilkov, Grigory. (2012). Why Does an Equal-Weighted Portfolio Outperform Value- and Price-Weighted Portfolios?. SSRN Electronic Journal. 10.2139/ssrn.1787045.

[6] Malladi, Rama & J. Fabozzi, Frank. (2017). Equal-weighted strategy: Why it outperforms value-weighted strategies? Theory and evidence. Journal of Asset Management. 18. 188-208. 10.1057/s41260-016-0033-4.

[7] Petersen, K. B. & Pedersen, M. S. (2008), 'The Matrix Cookbook' , Technical University of Denmark , Version 20081110 .

[8] People.duke.edu. (n.d.). Uses of the logarithm transformation in regression and forecasting. [online] Available at: https://people.duke.edu/~rnau/411log.htm [Accessed 17 Mar. 2019].