Time Series Analysis and Modelling of Financial Markets

Arman Tadjrishi

School of Physics and Astronomy

Cardiff University

A thesis submitted for the degree of

Master of Physics

May 2017

Acknowledgements

I am extremely grateful for the exceptional guidance of my supervisor, Dr Clarence Matthai. Without his help I would have never been able to meet such an ambitious challenge. Thank you for keeping me on the correct path as we moved forward with the project.

I am also grateful for the various assistance I was given from Cardiff University academics. Special thanks to Dr Paul Clark for the many useful explanations in data analysis, time and time again. Many thanks to Dr Andrey Pepelyshev from the Mathematics department for his help with time series analysis.

Abstract

The field of complex systems borrows expertise from a multitude of different disciplines. In physics our obsession is to detect patterns and build or identify models for our observation. In this project we looked at financial markets from the point of view of a complex system in hopes to build and identify models that describe it's many complex components using discrete time series analysis.

We showed that stock price series have small-form decaying memories that depends on the level of their complexity. We successfully identified a collection of characteristics in order to describe a series. We applied linear stochastic models to financial series and found that the description is adequate in some but not in others. We showed using regression and cross correlation that certain markets are very closely related, stocks such as Apple or Google were shown, statistically, to have causation relation with other stocks within the S&P500 information technology sector. We found that correlation between stocks also evolves with time, and certain periods of 'information delay' exist in were linear correlation takes time to move between markets.

We applied the BDS test of non-linearity to the industrial and information technology sector. We found that both series exhibit non-linearity in their trajectories and as a consequence we rejected the hypothesis of a random walk market. We calculated the Correlation Dimension, d, of both series in order to identify low dimensional chaos. We found that d did not converge in low dimensions and so chaos could not be identified, however the increase in d was not linear, suggesting that the series may be chaotic at higher dimensions.

Using inferences we built a mathematical model for the co-movement of stock price for the information technology and the industrial sector in the form of a recurrence relation. We tested a number of parameters on the deterministic model and found that we could not find a stability

in the trajectory. We made stochastic approximations to the model and found some promise. Simulations showed some similar characteristics as the original series, however these were not enough to reject independence with highest confidence. There is potential however, and in this work we identify some future avenues that could be taken in order to improve the model.

Finally we looked at influences from outside the system, we applied a simple, yet novel method of sentiment analysis to financial news data in order to quantify the effect of world events on the market. We found that significant linear correlation was present 5 business day (one week) after a news sources had been published. We found model parameters and our simulated series successfully rejected the hypothesis of independence with the original series.

Contents

Chapter 1

Introduction

1.1 Financial Markets

Financial markets are marketplaces where financial instruments are traded with prices that reflect the law of supply and demand. Often markets refer to particular exchanges that allow trade between parties for a small transaction cost. Exchanges can be physical institutions such as New York Stock Exchange (NYSE), or electronic ones such as the NASDAQ.

Definition 1.1.1. *Financial instruments* are monetary contract between parties. Instruments can be traded in many different forms, such as cash (foreign exchange), stocks or derivatives.

Financial markets are an important topic of study, often they are indicators of changes in a nation's economy or social outlook; hence why government bodies apply regulations and reforms constantly to control the dynamics of markets in order to better effect national prospects [1]. Furthermore, extensive research is done by institutions large and small in order gain a better understanding of the financial future and often, to further their vested interests.

Financial markets are an example of a highly complex system, often they are thought of as a closed system with a large number of interacting components and multiple dimensions that effect the dynamics of the market [2]. In the study of complex systems, we find that much of the same tools that can be used to model a system in a particular discipline may have important implication and uses in another, in this way we can define complex systems as a truly interdisciplinary field.

1.2 Time Series Analysis

A *time series* is a sequence of observations taken in time; a series is usually observed in regular time intervals that are equally spaced. In many systems and in fact in a financial market, the measurements observed are not equally spaced apart, and analysis on said data without careful consideration may have unwanted implications. Stock markets operate only on trading days, meaning that over weekends and public holidays the exchange is closed and no data is published. Daily stock market data therefore ignores these missed days and treats every iteration as a day apart. Thankfully, for stock market data, this has a negligible effect on the outcome of the analysis and so we can make the assumption that each point point in daily stock market time series is exactly one day apart.

An intrinsic feature of a time series is that, typically, series observations are *dependent* and *time series analysis* is defined as the techniques used to analyse the nature of said dependence. The requirement for time series analysis is the construction of stochastic and deterministic models that explain the dynamics of variables over time.

This project focuses on two sectors classified within the S&P500 index [3] listed companies, the information technology (IT) sector index and the industrial sector (IND) index. Complementary analysis is also carried out on commodity series'.

Definition 1.2.1. A stock market *index* a measure of value of a section of a stock market, typically a special weighted average of all its constituents. The S&P500 is an index of the 500 biggest companies in the US and bases its calculations on market capitalisation of each constituents.

For a list of constituents and tickers see [4] and [5].

1.3 Modelling Financial Markets

A mathematical model that expresses change in a system over time is known as a *dynamic* model; these models are often represented in the form of a differential or difference equation. A system can be described by either stochastic or deterministic models.

Definition 1.3.1. *Deterministic processes* are dynamic process where any two equation solutions that agree at one time agree every time [6]. In other words, knowing model parameters and variables exactly at one time, it is possible to know the exact position of the variables at another.

Definition 1.3.2. *Stochastic processes* have random variables sampled from probability space in their trajectory, therefore its impossible to know with absolute certainty the particular value of a measurement at a particular time.

Stochastic processes have different degrees of randomness and can often be seen as hybrids to deterministic processes when random factors are relatively small, however even in these processes predictability and forecasting becomes less accurate as the as the time variable increases, as even tiny random fluctuations would lead to exponentially large uncertainty over long periods. This ties in with a concept in dynamics known as ergodicity

Definition 1.3.3. *Ergodicity* defines a dynamic that, independent of starting position, would have the same statistical properties over time. Ergodicity in time series can be deduced from a single, sufficiently long, random sample of the process.

The efficient market hypothesis is a successful economic theory that explains that markets evolve as a direct result of new information instantaneously, i.e it would not be possible to determine the outcome of a market based on past values. The hypothesis has resulted in a number of debates, with studies showing strong evidence both for and against it [7]. Stochastic models have had the most successful in explaining financial observations. Econometric models such as those developed in 1970s by Box and Jenkins, known as ARIMA models have had huge success and have found themselves established in every time series analysis text book ever since[8]. Another very famous model is the Black-Scholes model [9], a differential equation that models option price based on an underlying volatility and interestingly, to solve the equation under certain boundary condition, its required that it be converted into the heat equation [10]. This isn't the only instance where physics has been utilised in financial modelling. Over the last 20 years, several different ideas from physics have been used in modelling, such as phase transitions [11] and agent based simulations influenced by the Ising magnetisation model, leading to varying levels of success in explaining market dynamics[12][13][14].

All this is not to say that deterministic models have no place in financial economics, non-linear models such as ARCH have been found to successfully predict short term volatility in certain time series[15], of course where these models fall off is when stochastic approximations are used to fill in gaps[16].

In the last few decades however, thanks for early pioneers such as Lorenz[17] and Mandelbrot [18], Hsieh[19], Labaron[20], there has been a certain excitement for the

development of non-linear deterministic models in hopes to be able to explain the movement of stock markets using chaos theory. The concept of chaos is often hard to gauge as there is no universally accepted definition however chaotic systems all have the following in common:

Definition 1.3.4. *Chaotic systems* are systems with extreme sensitivity to initial conditions, such that two dynamics with initial values $y(t)$ and $y(t) + \Delta y(t)$ will diverge at a rate proportional to $\exp \lambda t$ where λ is known as the Lyapunov exponent.

Chaos theory looks to explain seemingly random phenomena, and in the recent past there has been some success, albeit little, in modelling markets via chaos. Several works such as works by David Hsieh [19] have successfully shown evidence for non-linearity and some chaotic 'footprints', however this is still fundamentally a new field and more research needs to be done in order to develop a convincing argument for chaotic markets.

1.3.1 Model Building and Complexity

We discussed that financial markets can be characterised as highly complex systems, these systems can often be analysed in multiple layers. In its most microscopic model, a financial market consists of agents (traders) that are able to interact with other agents and move assets based on varying amounts of capital and interaction rules; these interactions are highly non-linear and are determined by one goal, to maximise profit for the agent. Other models take more of a macroscopic approach, looking at changes caused by large events and their time-scale. Often the most intuitive modelling option has been a statistical analysis of the copious amount of financial data gathered over the last 100 years and interestingly, there remains a number of generalisations that can be applied to all markets around the world no matter how diverse their markets are. Examples of these generalisations are fat-tailed distributions of squared price returns or scale invariance.

There's strong evidence for complexity when we consider the following:

- Linear statistical properties of a single financial time series (a single stock, commodity etc) show autocorrelation that decay quickly with time depending on how often trades occur. Non-linear properties have a time scale that is much larger, complexity is evident when there is both long and short range movements. [2].

- There is a high degree of correlation between different financial time series, this co-movement is closer for time series belonging to the same sector. This dynamical property between time series is a key aspect of complexity.
- The collective behaviour observation after particular market events whether typical or extreme. Real life events are often the cause of spikes in prices or short-term trends seen in the time series.

1.4 Project Aims and Thesis Layout

The project was set out in order to accomplish the following:

- Demonstrate and apply common time series analysis methods to a number of financial markets in order to make inferences about market dynamics.
- Test markets for non-linearity and assess whether markets are chaotic.
- Build a mathematical model based on reasonable assumptions and inferences from previous findings for a closed system of financial markets.
- Modify model to more closely align with real observations.
- Look at how forces outside of the market may effect the complex system, namely the effect of news as a reaction to world events.

In chapter 2 of this thesis, statistical methods is discussed, we looked at the statistics of financial time series by applying tools such as Fourier transforms, regression analysis, autocorrelation and ARIMA modelling. In chapter 3 we present a set of equations constructed that looks to model two co-moving time series and apply it to index data of different sectors within the same market. In Chapter 4 we discuss the implications of news on financial markets by employing text mining and language model techniques to construct a time series of news, this is compared to financial time series and ultimately put into the built model. Chapter 5 discusses our attempts to find chaotic characteristics within the stock market and the output of our model.

Chapter 2

Time Series Methods and Analysis

In this project we looked at many different time series from many different markets in multiple countries; most of the financial data was gathered using data available on Quandl[1][21]. Throughout the project, we utilise programming languages Python and R in order carry out analysis, Quandl has a built in API that works with both languages.

2.1 Financial times series

Data from the stock market is recorded as opening price (price at market open time), closing price, highest/lowest prices of the day and volume of trades. Below is the time series for the shares of Apple (ticker: AAPL) dating from 1981 until 2017. Immediately obvious is the sudden trend upwards in price that starts off around 2005, this is evidence that markets are not completely random, and a particular event or set of events triggered the motion. We can also create windows of time series such as the bottom of Figure 2.1, inspecting time series closer shows that movement is a result of both small and large fluctuations in price.

Other financial series have different overall trends, Figure 2.2 shows a number of different stocks traded under NYSE, 2 of which (Bank of America, goldman Sachs) are investment banks, as its visible, overall trends throughout time aren't always similar, even for companies that may operate closely. Equally there are similarities, take for example the large dip in price at around 2007, as many know this is the result of the worldwide financial crisis of 2007/08.

[1]Some data is freely available, others require a premium.

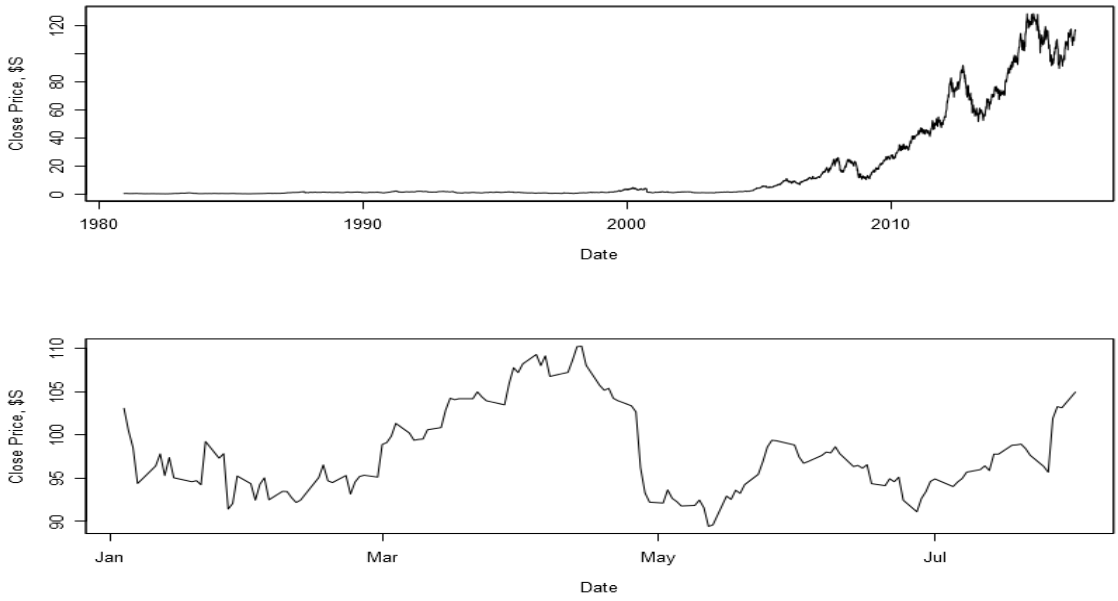

Figure 2.1: Time series plot Apple (AAPL) closing price. Top: movement of price since 1982. Bottom: time window between Jan 2016 to Aug 2017.

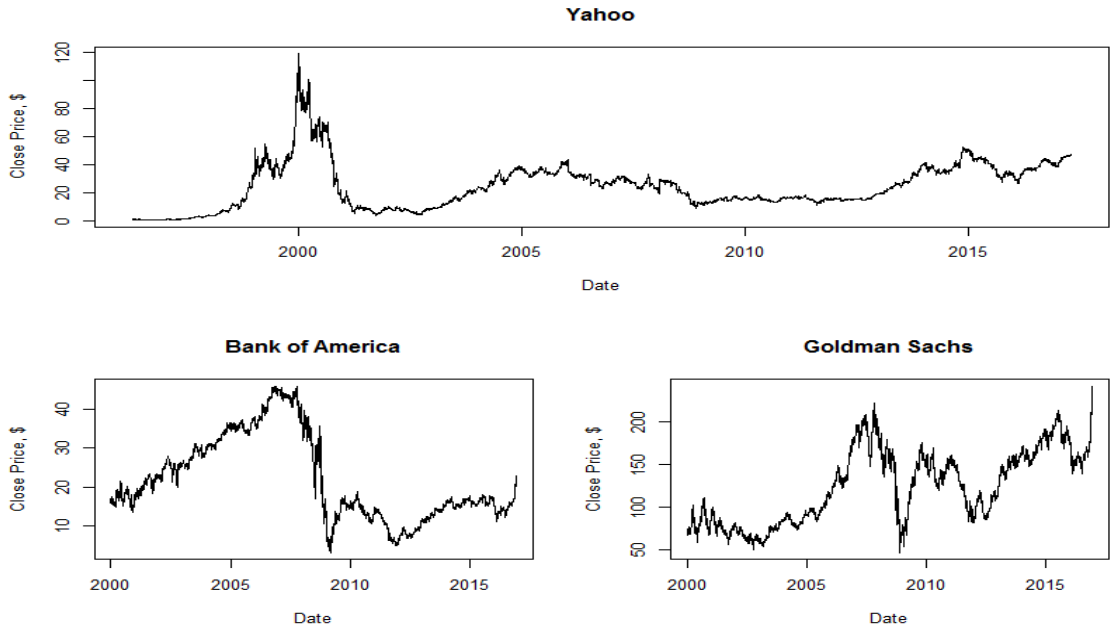

Figure 2.2: Examples of stock market time series traded under NYSE.

2.2 Statistical Methods and Analysis

In statistical analysis, it is useful to regard an observed series, $(x_1, x_2, x_3, x_4, \cdots, x_T)$ as a particular *realisation* of a stochastic process. A realisation is denoted as $\{x_t\}_1^T$ and the stochastic process is usually denoted as $\{X_t\}_{-\infty}^{\infty}$, we restrict the index on the stochastic process to match the realisation (1,T). Stochastic process can be described by a T dimensional probability distribution resulting in an analogy to population and sample statistics. These processes can be characterised by T expected values $E(x_1), E(x_2), \cdots, E(x_T)$ and T variances $V(x_1), V(x_2), \cdots, V(x_T)$ with the standard definition of the expected value

$$E[X] = \sum_{i=1}^{\infty} x_i p_i \tag{2.1}$$

with p_i being the probability of the random variable. Variance is given by

$$Var(X) = [E(x_i - E(X))^2] \tag{2.2}$$

its important to note that using a single realisation to infer the properties of a stochastic process is only valid if the process is ergodic. Testing for ergodicity is a difficult task using a single time series, so we make the assumption that for any *stationary* time series this requirement is met [22].

2.3 Stationarity

Stationarity is a very important in both linear and non-linear time series analysis. Trends can be deterministic (linear or non-linear) or stochastic via stochastic shocks that cause permanent movement in series position. In general we require stationarity as we generally look to describe system as a set of recurrence or regression relation, both of whom need to be independent of time in order to have constant parameters. Stationarity is therefore a requirement in both mean and variance.

Definition 2.3.1. *Strictly stationary* processes are those where all characteristics of the time series are independent of time, i.e. the joint probability distribution for one set of times $(t_1, t_2, \cdots, t_n)$ remains the same as the joint probability distribution at times $(t_{1+k}, t_{2+k}, \cdots, t_{n+k})$. where k is an arbitrary shift in time.

Definition 2.3.2. *Weakly stationary* processes are defined as those with a mean, variance and autocovariance that remain constant with time.

Strictly stationary processes imply weak stationarity, though this is not always the case the other way round. For Gaussian processes, weak stationary would also imply strict stationarity as distributions are defined by the stationary moments. The *autocovariances* are defined as

$$\gamma_k = Cov(x_t, x_{t-k}) = E[(x_t - E(X))(x_{t-k} - E(X))] \tag{2.3}$$

and *autocorrelations*

$$\rho_k = \frac{Cov(x_t, x_{t-k})}{[V(x_t)V(x_{t-k})]^{\frac{1}{2}}} \tag{2.4}$$

for a stationary process

$$\rho_k = \frac{\gamma_k}{\gamma_0}. \tag{2.5}$$

2.3.1 Types of Stationarity

There are a number of methods to test for stationarity and we will discuss some of the ones that are utilised in the project here.

Definition 2.3.3. *Trend stationarity* is a process with a (linear or non-linear) deterministic trend term in in the model. If we consider a model

$$Y_t = \theta Y_{t-1} + \mu + \alpha t + \epsilon_t \tag{2.6}$$

Where μ is a constant, ϵ_t are random variables and $|\theta| < 1$, solving for Y_t we find that $E[Y_t]$ also contains a linear trend, i.e non-stationary. Any stochastic shocks to the process would be temporary and so the process is *mean reverting*.

Definition 2.3.4. *Unit root* processes non-stationary processes where stochastic trend is present and drifts are non mean reverting. A simple model for a unit root process is

$$Y_t = \theta Y_{t-1} + \mu + \epsilon_t \tag{2.7}$$

A unit root is present if $\theta = 1$. Unit root processes are very sensitive to initial conditions (Y_0) and processes with different starting positions would have completely different trajectories.

2.3.2 Stationarity Tests

Stationarity in a deterministic trend is often very easy to identify and remove, if linear, simple linear regression with with the time variable can be utilised and have trend

subtracted. For non-linear trends usually a known model is fit or one is estimated with an n order polynomial and subtracted in a similar way. Unit roots are more difficult to identify and so there are a number of tests that can be used to identify them.

2.3.2.1 Augmented Dicky-Fuller Test

The Dicky-Fuller[**?**] tests the null hypothesis that in equation 2.7 $\theta = 1$ and the alternate hypothesis $\theta < 1$. The Dicky-Fuller is the t-test

$$\hat{\tau} = \frac{\hat{\theta} - 1}{\text{Standard Error}(\hat{\theta})} \tag{2.8}$$

In most circumstances, a 5% critical value is is given by a statistic of -1.95. The Augmented Dicky-Fuller test takes this further by modelling a time series as

$$y_t = \theta_1 y_{t-1} + \theta_2 y_{t-2} + \theta_3 yt - 3 + \cdots + \theta_n y_{t-n} + \epsilon_t \tag{2.9}$$

Defining

$$\pi = \theta_1 + \theta_2 + \theta_3 + \cdots + \theta_n - 1 \tag{2.10}$$

The hypotheses therefore are

$$H_0 : \pi = 0 \text{ and } H_A : \pi < 0 \tag{2.11}$$

and the t-test is

$$\hat{\tau} = \frac{\hat{\pi}}{\text{Standard Error}(\pi)} \tag{2.12}$$

A known deterministic trend can be accounted for to test for unit roots by the addition of a trend term in equation 2.10.

2.3.2.2 KPSS

KPSS (Kwiatkowski, Phillips, Schmidt and Shin) [23] test reverses the hypothesis test and has stationarity as the null hypothesis. The null hypothesis assumes a stationary process

$$Y_t = \hat{\mu} + \epsilon_t \tag{2.13}$$

i.e.

$$H_0 : Var(\epsilon_t) = 0 \text{ and } H_A : Var(\epsilon_t) > 0 \tag{2.14}$$

The test statistic is given by:

$$KPSS = \frac{1}{T^2} \frac{\sum_{t=1}^{T} S_t^2}{Var(\epsilon)^2} \tag{2.15}$$

where $S_t = \sum_{s=1}^{t} \epsilon_s$. The two tests, KPSS and ADF can be used in conjuncture for confirmatory analysis.

2.3.3 Finding Stationarity

There are four main methods of stationarising a time series; *smoothing and decomposition*, *trend subtraction* and *transformation*. Non-stationary time series can be made stationary by applying a smoothing function or a *filter* to the original time series, and removing the smoothed time series from the original one. Moving averages are one method of filtering, whereby an average of a subset of the series is calculated for each point leading up to point of interest. This results in a smooth function with a trend similar to the original time series. Seasonal *decomposition* is a related idea, time series with expected seasonal peaks can have them decomposed by calculating an average value at each particular season and subtracting those averages from the time series. In Figure 2.6 we show both of these methods. For monthly data from the S&P500 Financial Sector from the years 2002-2014, we find the average values for every month and subtract the time series from the original data. Then we fit a moving average time series of the decomposed series and subtract. The residuals are in the form of a stationary time series. We can reach stationarity by fitting a particular trend to the time series, for example we fit a model in the form of $Y = mt + c$ to data from the S&P500 GICS Information Technology Index, in Figure 2.4 fit is shown, removing the fit we are left with a time series that looks stationary, however this is not the case after running an ADF test. We find $d_f = -2.96$, $p(d_f) = 0.286$, for 5% significance, we cannot reject the null hypothesis that there remains a unit root in the residuals. In fact the main issue with using models to stationarise data is the requirement of prior knowledge of the model itself and for stock market data linear models (or simple polynomials) are almost never fully satisfactory for detrending. Non-linear transformations can often be used to help approximate a linear trend, for example taking the natural logarithmic transform of S&P IT sector time series we can apply a log transformation so that we are left with the time series

$$Y_t = log(y_t) \tag{2.16}$$

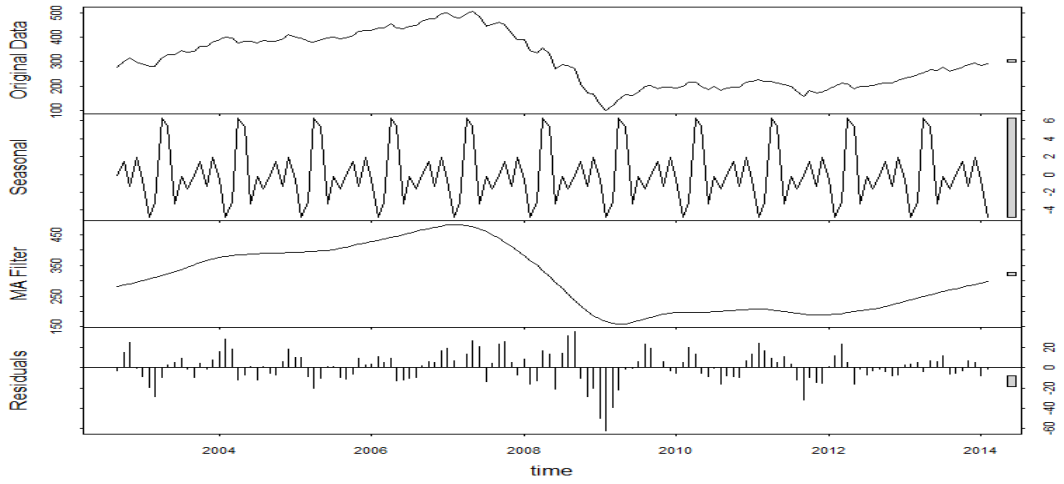

Figure 2.3: Seasonal and moving average decomposition of monthly S&P500 financial sector data

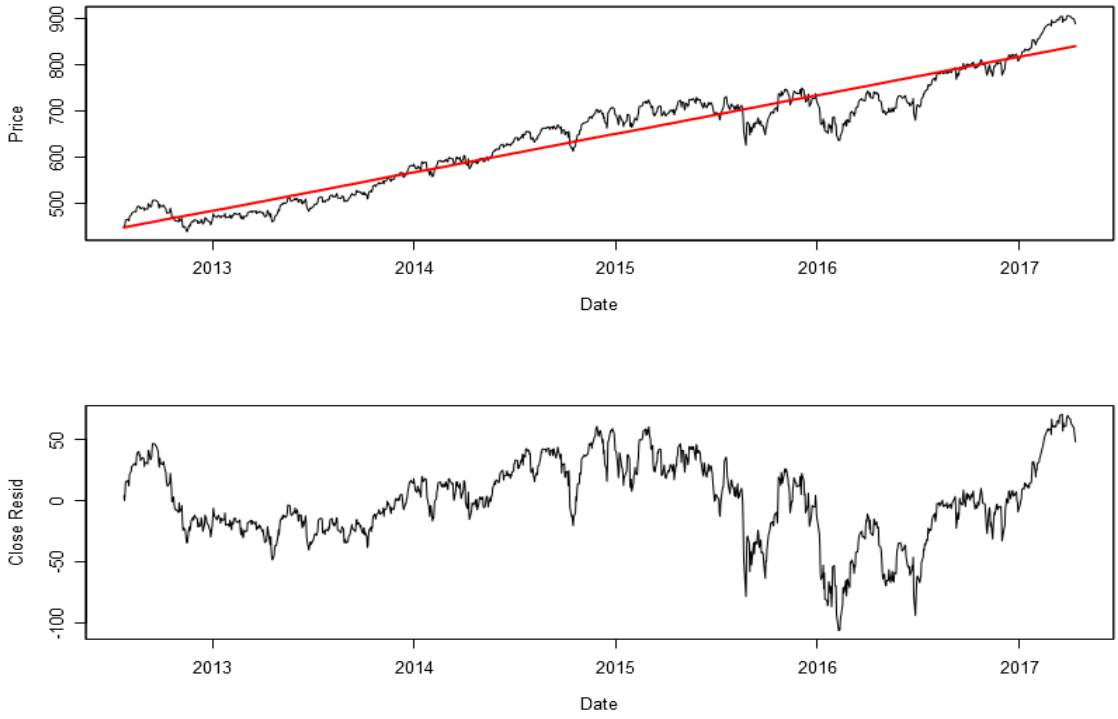

Figure 2.4: Data from S&P500 IT sector. Top: Linear fit of time series with time as the dependent variable. Bottom: Residuals after fit. subtraction

We can then apply the linear trend to our data, this method is known as Log-Linear Detrending (LLD) and is a frequently used method in stock market analysis.We find

$d_f = -3.2$,$p(d_f) = 0.146$ The ADF tests from the residuals from LLD show that the null hypothesis can't be rejected, however the time series is now more stationary than before. Figure 2.5 shows this. Box and Jenkins [8] introduced a method make a time

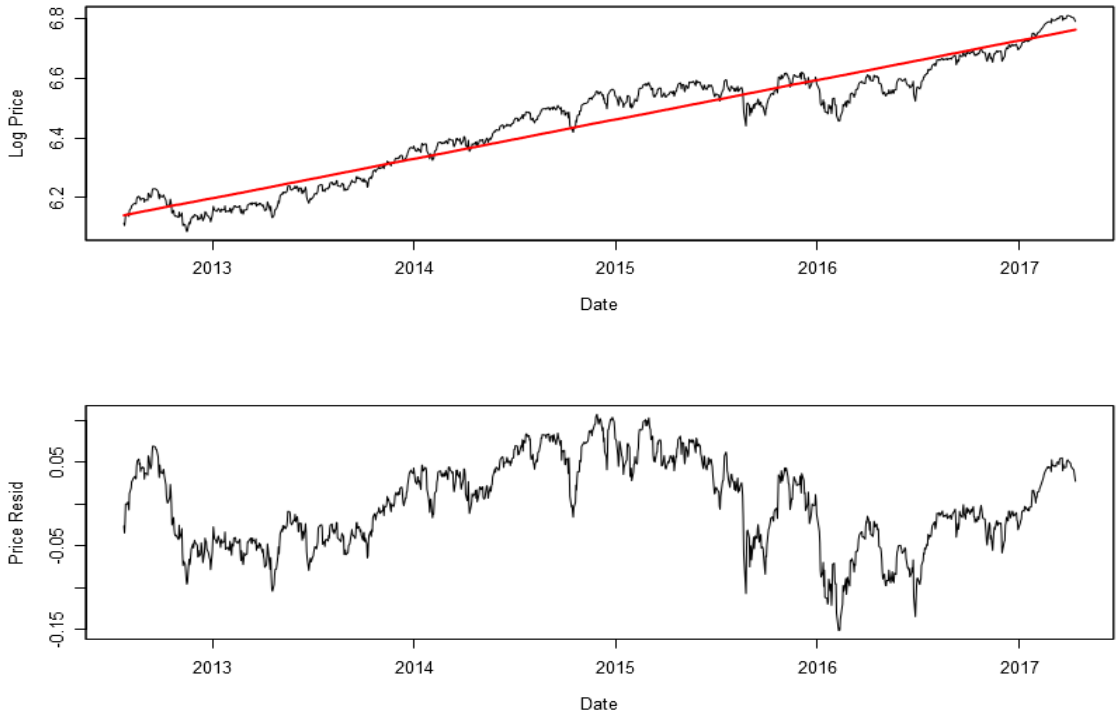

Figure 2.5: Data from S&P500 IT sector. Top: Log Linear fit of time series with time as the dependent variable. Bottom: Residuals after fit. subtraction

series stationary, known as *differencing*. The first difference of a time series can be shown as

$$\Delta x_t = x_t - x_{t-1} \tag{2.17}$$

Introducing the backshift operator

$$Bx_t \equiv x_{t-1} \tag{2.18}$$

and, in general,

$$B^m x_t = x_{t_m} \tag{2.19}$$

we can define the difference operator

$$\Delta^m = (1 - B^m) \tag{2.20}$$

where m is the order of difference. We can show that this transformation leads to a stationary series from Figure 2.6 in the S&P500 IT sector index, the ADF test of the first differences show that the null hypothesis of a unit root is rejected. We find $d_f = -11.031$,$p(d_f) = 0.01$ Transformations such as logarithmic ones can be used in

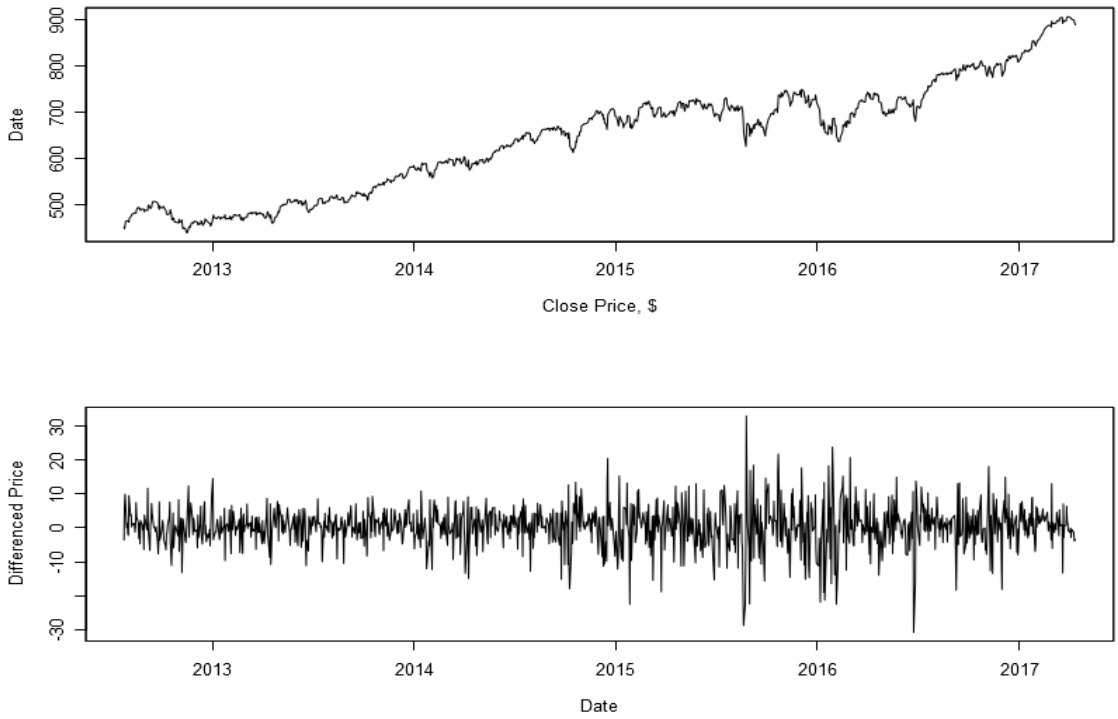

Figure 2.6: Data from S&P500 IT sector. Top: Index data with no transformation Bottom: First difference of index. subtraction

conjunction with differencing. In fact, in economics, geometric returns of a stock is just the difference of the logged series, i.e.

$$r_t = log(x_t) - log(x_{t-1}) \tag{2.21}$$

the returns, r approximately equal to the percentage change in price through time, and time series' are often modelled using returns as the variable of interest.

2.4 Fourier Analysis

One of the main distinguishing features of time series analysis to ordinary regression is the presence of repetitive or regular behaviour over time. The concept of regularity

can be expressed in terms of periodic variations, such that the series is produced by a combination of sins and cosines expressed as Fourier frequencies. Most may be familiar with the concept of the Fourier series, where any truly periodic function can be reorganised as a finite sum of sins and cosine functions. Generally, Fourier transforms can be used to describe any function with an infinite sum over continuous space. The Fourier transform is highlighted by the following equations

$$f(t) = \int_{-\infty}^{\infty} F(v) \exp(2i\pi vt) dv \tag{2.22}$$

Where $f(t)$ is a function or signal, v are the frequencies and $F(v)$ are the fourier coefficients given by

$$F(v) = \int_{-\infty}^{\infty} f(t) \exp(-2\pi ivt) dt. \tag{2.23}$$

In the fields of time series analysis we deal with discrete space and a finite series. Given a trajectory, the discrete Fourier transform (DFT) can break the equally spaced series into equivalent length cycles. Each cycle has an amplitude, a phase delay and frequency. The equation for the DFT is

$$X_f = \sum_{t=0}^{N-1} x_t \exp(-i2\pi ft/N) \tag{2.24}$$

where X_f is the Fourier coefficient, translating to the amplitude of frequency f[2]. Each f is a complex number, including both an amplitude and phase shift. N is the number of samples in the series x_t. Fourier coefficients can be converted to the time domain via the inverse Fourier transform (IFT)

$$x_t = \frac{1}{N} \sum_{f=1}^{N-1} X_f \exp(i2\pi ft/N). \tag{2.25}$$

DFTs are the most important transforms in spectral analysis [24], however computing the functional form above for large series becomes exceedingly taxing on larger series, thankfully we can utilise modern computational algorithms such as the Fast Fourier Transform (FFT) in order to greatly increase calculation speeds[3]. The plot of frequencies against their amplitude is known as the *spectrum* and sometimes as the

[2]Here we have used f for frequency, rather than v in equation 2.22

[3]The FFT has built-in libraries in Python and R, for details on the algorithm see [25]

periodogram[4]. For example we can show the result of a sinusoidal of the form

$$x_t = \frac{3}{4}\sin(3(2\pi)t) + \frac{1}{4}sin(7(2\pi)t) + \frac{1}{2}\sin(10(2\pi)t) \tag{2.26}$$

where t is from $0 : 6s$. Figure 2.7 shows this, on the left is the time domain plot and the right is the spectrum, where each peak shows the amplitude of the particular harmonic frequencies[5].

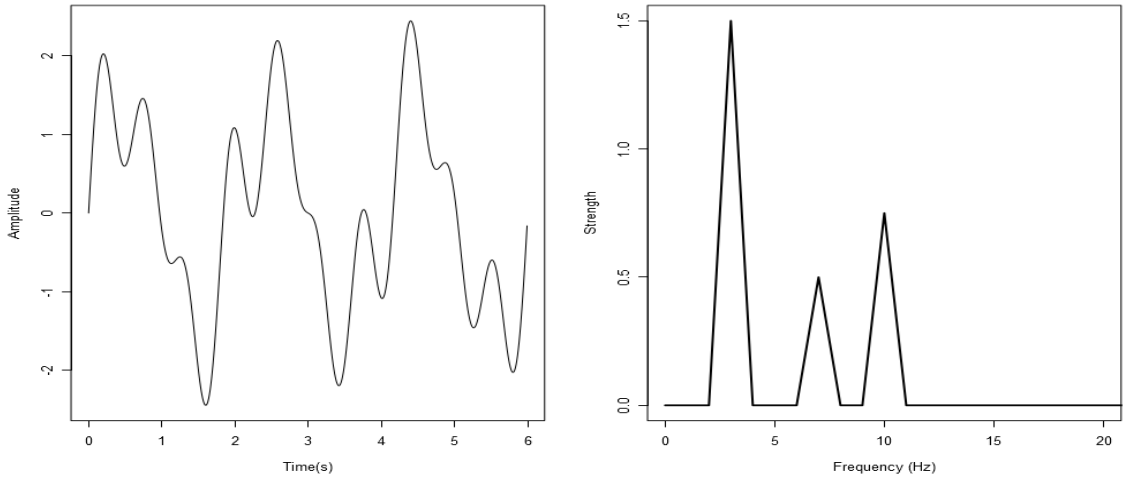

Figure 2.7: Left: Cyclic trajectory in the time domain, Right: Spectra

Another example is a wave with a more complex cycle, in order resemble a financial series we add a linear trend and Gaussian random variables at each t:($\epsilon = i.i.d.\ N(0,1)$) i.e

$$x_t = 20t + 3\sin(3(2\pi)t) + \epsilon_t. \tag{2.27}$$

In Figure 2.8's spectrum, we see the largest amplitude at first harmonic frequency ($1Hz$), this corresponds to the linear trend component of the wave, the second frequency also has a high amplitude, which is not part of the sin function, in fact the trend in the series has distorted the spectrum and the DFT hasn't captured the signal. The harmonic frequencies are shown in the bottom left, the presence of both noise and trend has masked the true periodic frequency of $3Hz$. We can also see that the presence of random variables scatters the spectrum and leads to a broadband spectra. This result can be seen when we find the Fourier transform of real stock data, Figure 2.9 shows the transform of the S&P500 IT sector data. Like the sinusoidal example, the first frequency is by far the largest, corresponding to the time trend. This forms

[4]In signal processing, the periodogram is actually a scaled version of the DFT

[5]The Fourier coefficients have been normalised, so what you're seeing is in fact X_f/N

a problem if we want to identify periodicity in the market, as non-stationarity clearly has a large effect on the Fourier transform; in fact we show this in the bottom right of Figure 2.9 where the first 10 harmonics are summed to reconstruct the series, the overall trend is approximately represented by the first number of frequencies. We note that at the beginning and end points of the series, the reconstructed series greatly diverges from the original, these 'warparound' effects are a result of the requirement for periodicity in the series; what we see is an attempt for the series to match its endpoints for periodic repetition. .

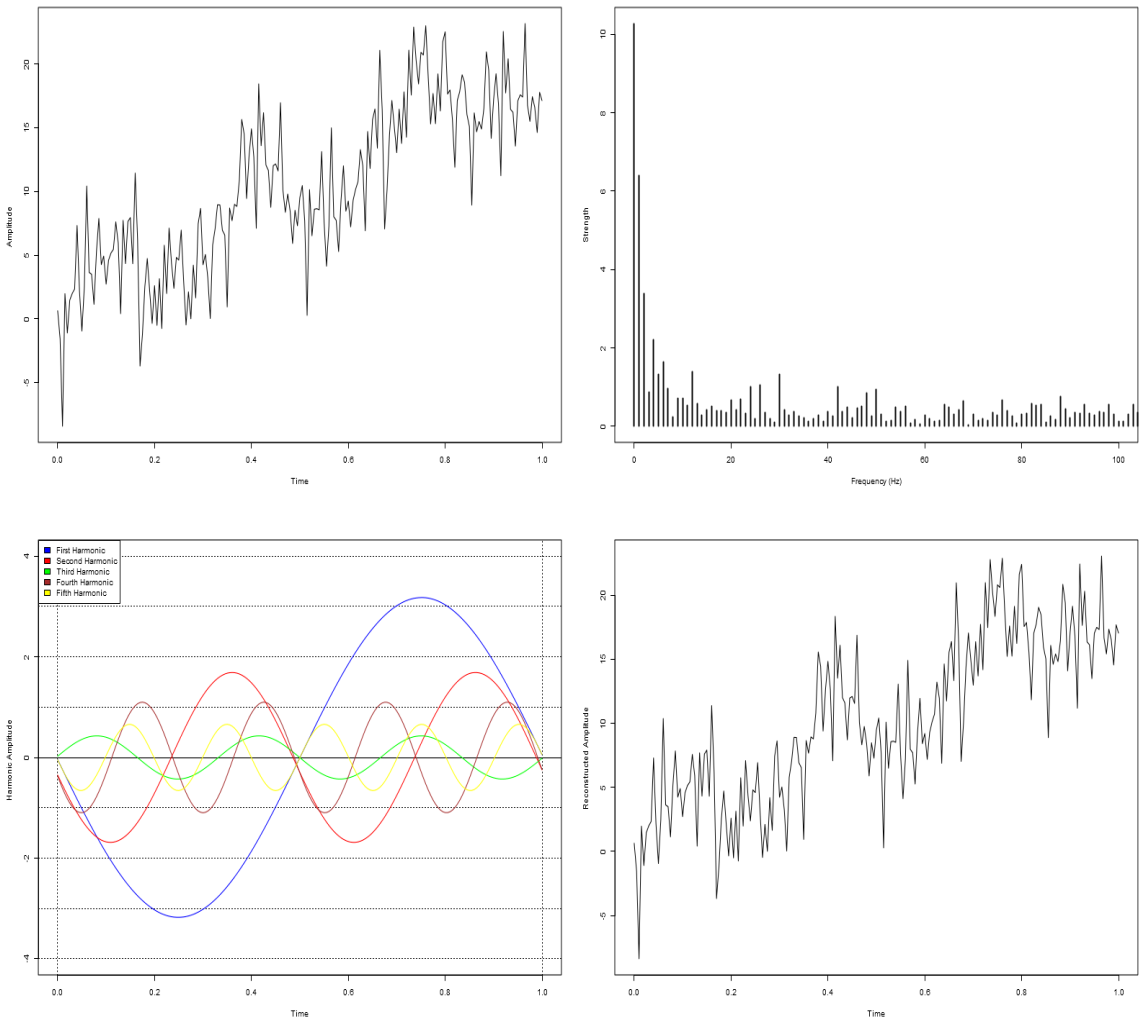

Figure 2.8: Top left: Cyclic trajectory in the time domain. Top right: Frequency Spectra. Bottom Left: First five harmonics. Bottom Right: Reconstructed time series using all harmonics.

Stationarity is also useful for Fourier analysis. We employ the use of the differencing to achieve stationarity in the IT index, and apply the Fourier transform. We

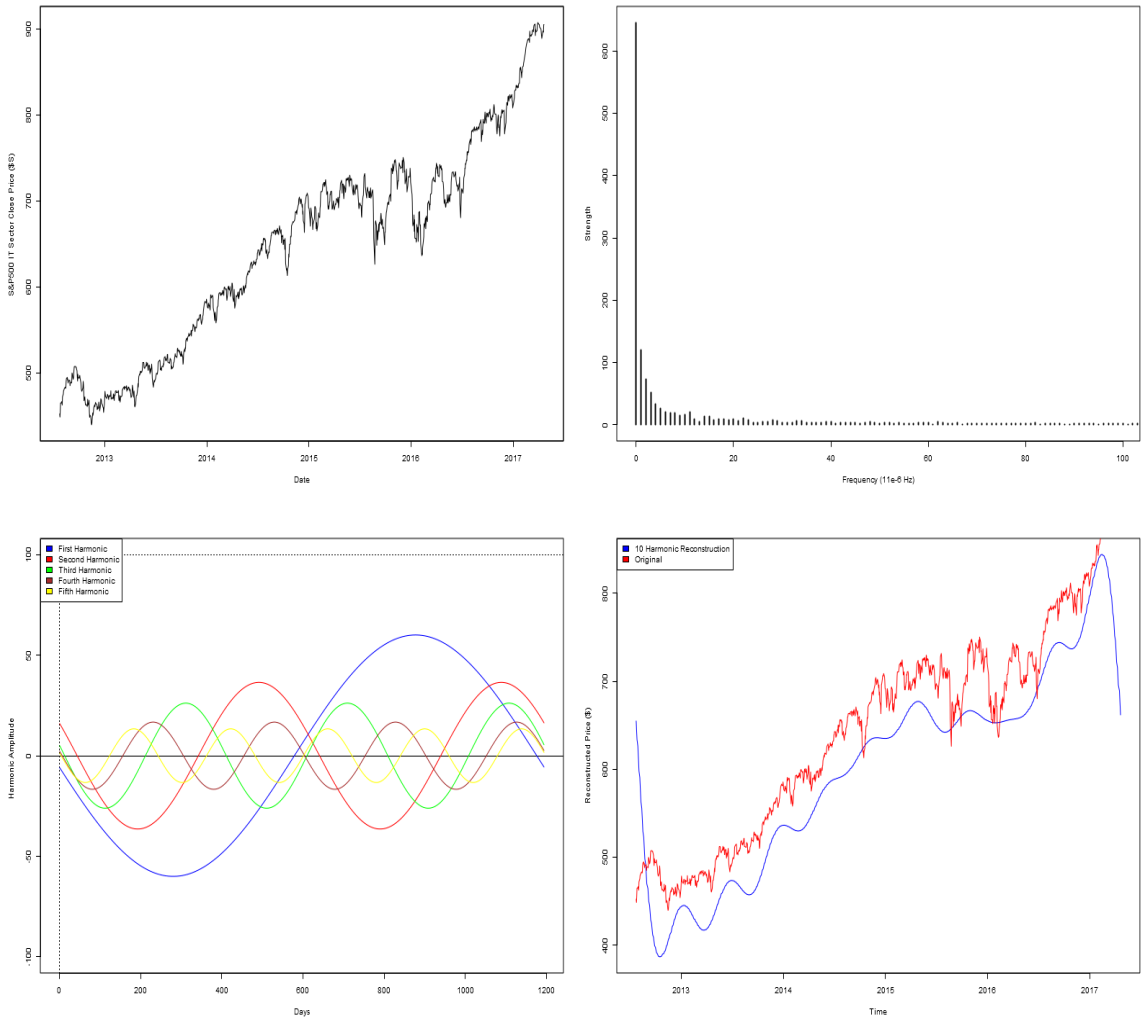

Figure 2.9: Top left: S&P500 IT sector trajectory in the time domain. Top right: Frequency Spectra. Bottom Left: First five harmonics. Bottom Right: Reconstructed time series using first 10 harmonics.

can see the spectrum on Figure 2.10 we see that all frequencies over the maximum resolvable frequency range of 596 day^{-1} are somewhat homogeneous in amplitude. There are certain frequencies that are more pronounced than others, however its not possible to identify them with any certainty. this is evidence for aperiodicity in the time series. Due to the Nyquist frequency limit, reconstructing the time series leads to some discrepancy, with a error much larger than the non-stationary reconstruction, this is because every harmonic frequency now adds a much larger proportion to the series; this problem may be resolved by using data sampled at every $12h$ intervals.

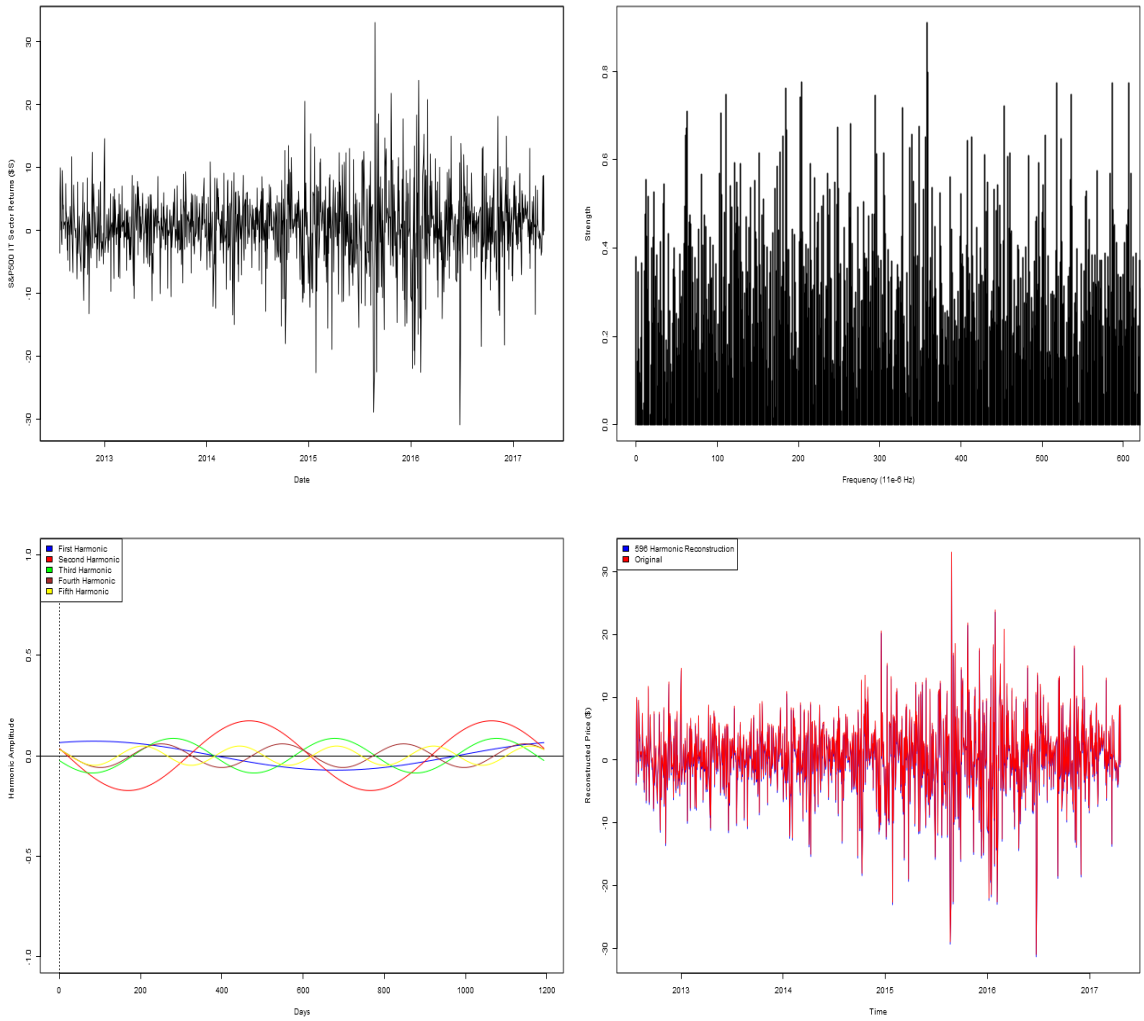

Figure 2.10: Top left: S&P500 IT sector differenced trajectory in the time domain. Top right: Frequency Spectra. Bottom Left: First five harmonics. Bottom Right: Reconstructed time series using 596 harmonics.

2.5 Autocorrelation

Earlier we introduced the autocorrelation statistic in equation 2.28, autocorrelation in time series is a useful measure of how well correlated a sample is with previous subsets of the series. Knowing if a series is serially correlated would help us determine models for the time series, as we have more insight on how the system evolves. Autocorrelation is based on Pearson's correlation coefficient (R), which determines if two sets of variables show linear dependence, for a sample, R is given by

$$\rho_{x,y} = \frac{E[(X - \mu_X][Y - \mu_Y)]}{\sigma_Y \sigma_Y} \tag{2.28}$$

Where μ is sample mean, and σ is the sample standard deviation. The autocorrelation function (ACF) represents the sample autocorrelation with respect to order of lag k, given by equation 2.28. The plot of ACF against lag is known as the *correlogram*. Inspecting the correlogram can also tell us if there are repeating or periodic patterns in the data. By definition, a white noise series ω_t with $E(\omega_t) = 0$ and variance σ_ω^2 will have autocorrelation

$$\rho_\omega(s,t) = \frac{E[\omega_s, \omega_t]}{\sigma_\omega^2} = \begin{cases} 1, & s = t \\ 0, & s \neq t \end{cases} \tag{2.29}$$

and we can see this to be true if we simulate the series, in Figure 2.11 this is shown, the ACF is at 1 for 0^{th} lag and drops to insignificant values thereafter. The ACF distribution follows a normal one and has $\mu = -1/N$ and $\sigma^2 = 1/N$ where N is sample length. The significance is of critical value of 95%, this is calculated from this, i.e error bars of $\pm 2\sigma = 2/\sqrt{(}N)$. ACF can be used to detect non-stationarity in time series, as long term trend would show very steady but gradual decay with lag. This is only the case when the series is non-stationary in μ, if a series is non-stationary in variance alone, the ACF will not identify non-stationarity.

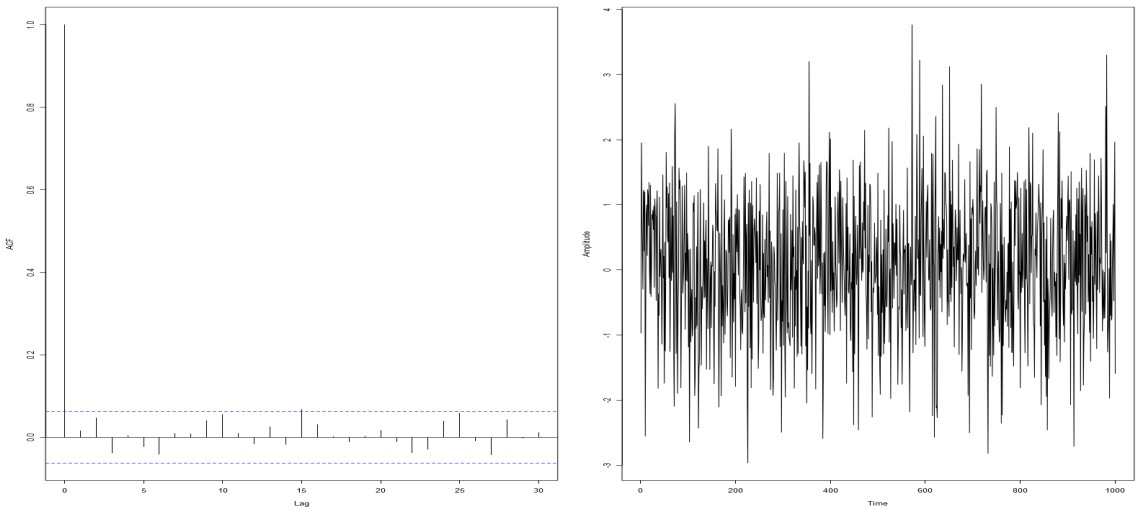

Figure 2.11: Left: White noise time series ($NID \sim (0, 1)$, Right: Autocorrelation Function

A similar measure is the *partial autocorrelation* function. In general, the correlation between two variables can be due to the variables being correlated with a third. For example the correlation between x_t and x_{t-3} may be due to the correlation this pair has with the intervening lags x_{t-1} and x_{t-2}, the PACF adjusts for this.

The k^{th} partial autocorrelation is the coefficient ϕ_{kk} in the serial process

$$\hat{x}_t = \phi_k 1 \hat{x}_{t-1} + \phi_{k2}\hat{x}_{t-2} + \cdots + \phi_{kk}\hat{x}_{t-k} + \omega_t \tag{2.30}$$

Where $\hat{x}$ is the mean subtracted series. We can find ϕ_{kk} via multiple regression, or alternatively through the relationship

$$\begin{bmatrix} \rho(0) & \rho(1) & \cdots & \rho(k-1) \\ \rho(1) & \rho(0) & \cdots & \rho(k-2) \\ \vdots & \vdots & \vdots & \vdots \\ \rho(0) & \rho(1) & \cdots & \rho(0) \end{bmatrix} \begin{bmatrix} \phi_{k1} \\ \phi_{k2} \\ \vdots \\ \phi_{kk} \end{bmatrix} = \begin{bmatrix} \rho(1) \\ \rho(1) \\ \vdots \\ \rho(k) \end{bmatrix} \tag{2.31}$$

The white noise series' PACF is plotting in Figure 2.12, as expected, this follows the same form as the ACF for a white noise series with no significant correlations.

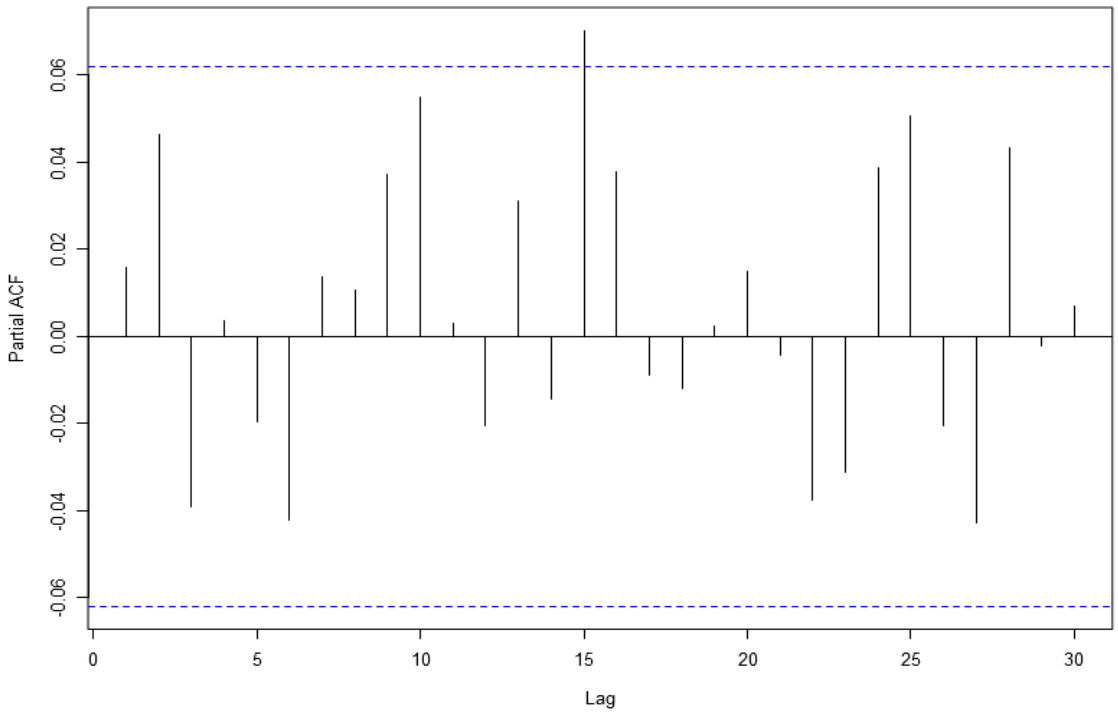

Figure 2.12: PACF of white noise series.

We applied the ACF and PACF to some financial time series to make inferences on their dynamics. Figures 2.16 and 2.13 show the ACF and PACF for the differenced S&P500 IT index, respectively. The left side of the ACF plot shows correlation for the first 60 ks, correlation looks very small, suggesting that the series is progresses

independently from previous values. Extending the plot to 800 k we see that this may not be the case, the right hand plot reveals that

- Significant correlations exist on a scale larger than 60 days, for example at lags of $k \approx 80$ we see significant correlation.
- The ACF gradually decreases over long lags, k.

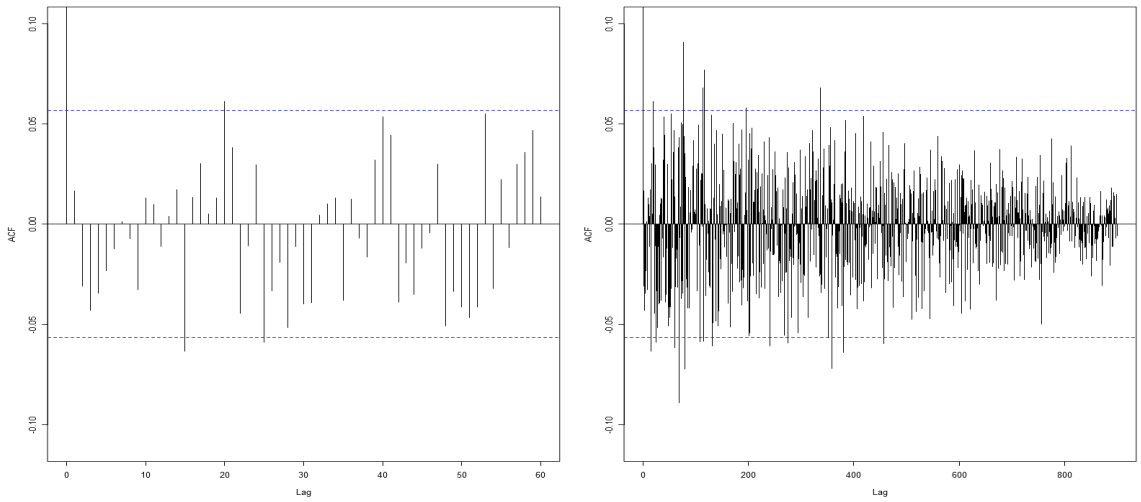

Figure 2.13: Left: partial autocorrelation with max lag $k = 80$. Right: autocorrelation with max lag $k = 900$

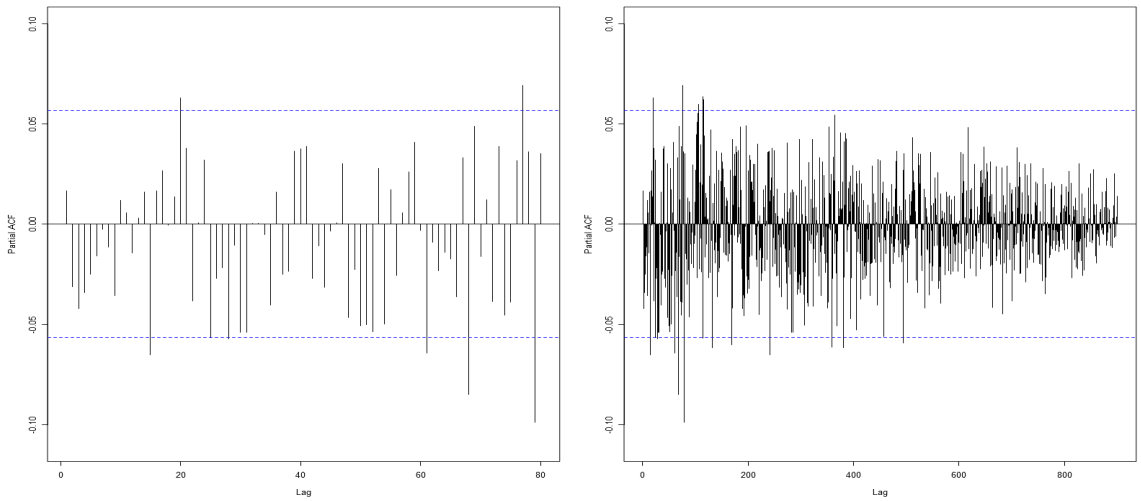

Figure 2.14: Left: partial autocorrelation with max lag $k = 80$. Right: autocorrelation with max lag $k = 900$

What we can infer is there is both a seasonal component in the system that arises every four months and that there may be some long term memory in the system. Short term correlation is still not evident from the ACF. From the PACF plots we can see that there exists significant correlation at lag $k = 20$, this corresponds to a months difference. This result makes sense, a month's end is often a time when companies/ news sources release progress reports, information would persuade potential investors to buy or sell, leading to significant correlation months apart (and little in between). As stated earlier, evidence for complexity comes from the short term memory of stocks when we analyse the $\{x_t^2\}$ series. In Figure 2.15 this is shown, the autocorrelation for the squared differences show a decay of significant correlations over time, suggesting that the system is in fact auto correlated non-linearly. It was stated that for memory

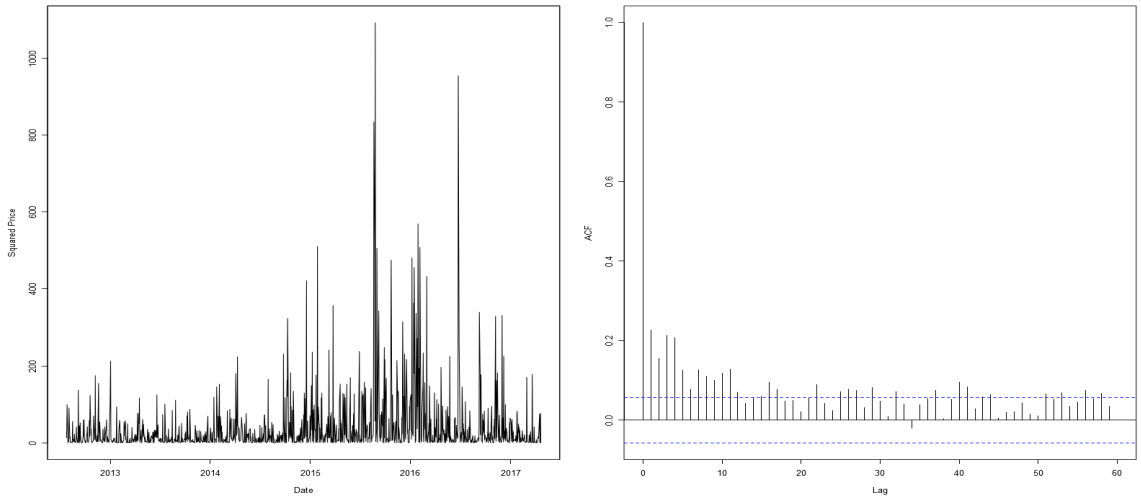

Figure 2.15: Left: plot of difference squared IT index. Right: autocorrelation, showing slow decay of correlation.

in the system decays at a rate proportional to how often a the instrument is traded, we can demonstrate this by inspecting correlation of the prices of three commodities that have different trade volumes. In Figure 2.16 we calculated the ACF and PACF of the commodities copper, cotton and gold. Gold is a heavily traded commodity and traded a number of magnitudes higher than both copper and cotton [26][27]. What is clearly evident that there is significant autocorrelation in the $k = 1$ of both cotton and copper but this is not the case for gold. We can infer that the volume of trades has a direct effect on the rate of short term correlation. its likely that there exists autocorrelation in gold however on a scale smaller than $k = 1$ to find this we would need to study intraday market data. We can see that some time series are correlated

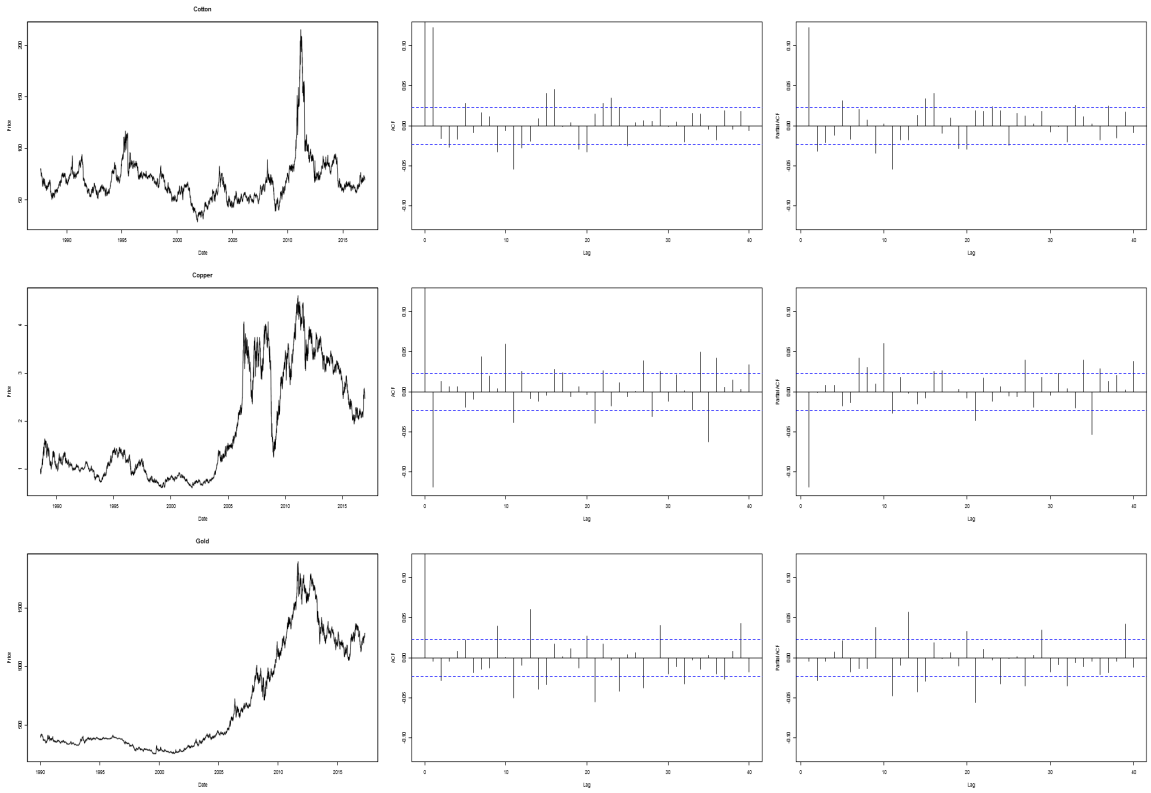

Figure 2.16: Time series of commodities copper, cotton and gold, with the respective ACF and PACF

with their own past values at specific lags, using this knowledge it is possible to build mathematical models to express the evolution of a system, in fact Box and Jenkins [8] showed that the power of ACF and PACF can be used to explain a stochastic process using ARIMA modelling.

2.6 Linear Stochastic Models

There are a set of stochastic difference equations, that aim to model any weakly stationary process as a linear combination of a sequence of random variables. These are known as ARIMA models and we introduce them in this section.

Definition 2.6.1. An *autoregressive model* is a generalised stochastic difference equation of the form

$$x_t = \alpha + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdot + \phi_p x_{t-p} + \epsilon_t \tag{2.32}$$

where α is the mean subtracting term of the form $\alpha = \mu(1 - \phi_1 - ... - \phi_p)$, for a series

with 0 mean this disappears. By convention, the model is abbreviated as **AR(p)**. ϵ_t is a series of independent Gaussian random variables denoted by $\epsilon_t \sim NID(0, \sigma_\epsilon^2)$. The $AR(1)$ model can be generalised by considering

$$\begin{aligned} x_t &= \phi_{t-1} + \epsilon_t \\ &= \phi(\phi x_{t-2} + \epsilon_{t-1}) + \epsilon_t \\ &= \phi^k x_{t-k} + \sum_{j=0}^{k-1} \phi^j \epsilon_{t-j} \end{aligned} \tag{2.33}$$

We used Monte-Carlo methods to simulate $AR(1)$ and $AR(2)$ and calculate their ACF and PACF in order to make inferences about their structure; Figure 2.17 shows this. What we see is that for the $AR(1)$ series with $\phi = 0.9$, observations close to

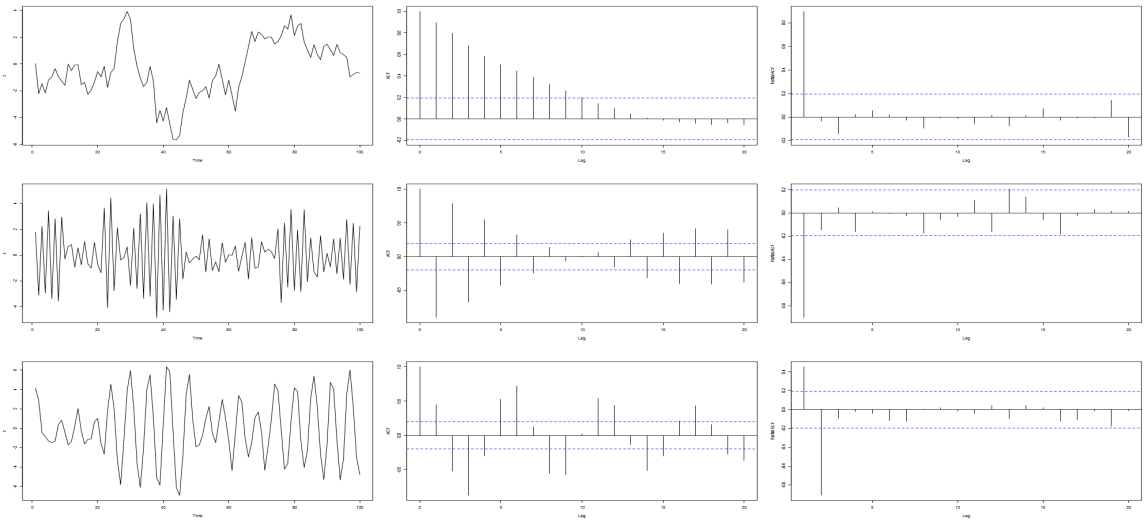

Figure 2.17: Top: simulated $AR(1)$ with $\phi_1 = 0.9$ with ACF AND PACF adjacent. Middle: simulated $AR(1)$ with $\phi_1 = -0.9$ with ACF AND PACF adjacent. Bottom: simulated $AR(2)$ with $\phi_1 = 0.9$ and $\phi_2 = -0.9$ with ACF AND PACF adjacent. $\sigma_\epsilon^2 = 1$

each other will be positively correlated, stationarity requires $|\phi| <$ in all models, and so what we see is that the decimal coefficient acts as a force driving the series closer to the mean $\mu = 0$. The ACF of shows a smooth decay to zero, whereas the PACF ($= \phi_1 = 0.9$) cuts off to zero at lags after k. In this way, the $AR(1)$ the series is a Markovian process i.e

$$P(x_t|x_{t-1}, x_{t-2}, \cdots x_0) = P(x_t|x_{t-1}) \tag{2.34}$$

The $AR(1)$ process with $\phi_1 = -0.9$ shows that nearby observations are negatively correlated with each other and so we observe a fluctuating pattern. From the ACF we see an oscillating decay towards zero and the observed cutoff i.e. PACF $= \phi_k$. The $AR(2)$ process with $\phi_1 = 0.9, \phi_2 = -0.9$ shows that the process (now non-Markovian) has a damped oscillating pattern, and the PACF cuts off at lag p with correlations $\phi_p = \phi_p$.

Definition 2.6.2. *Moving average* models are alternatives to the autoregressive representation, which x_t are no longer the linear combination, but rather the white noise series ϵ_t are linearaly combined, i.e

$$x_t = \epsilon_t + \theta_1 + \epsilon_{t-1} + \theta_2\epsilon_2 + \cdots + \theta_q\epsilon_q \tag{2.35}$$

with q lags in the moving average and so by convention the model is described as $MA(q)$. We created Monte-Carlo simulations of two $MA(1)$ processes in Figure 2.18, comparing with the $AR(1)$ processes in Figure 2.17, the realisations from the $MA(1)$ are often quite similar, suggesting that it may, on occasions, be difficult to distinguish between the two. Furthermore, we see the opposite pattern with regard to the ACF and PACF, in the $MA(1)$ model, it is the PACF that decays smoothly, whereas the ACF cuts off.

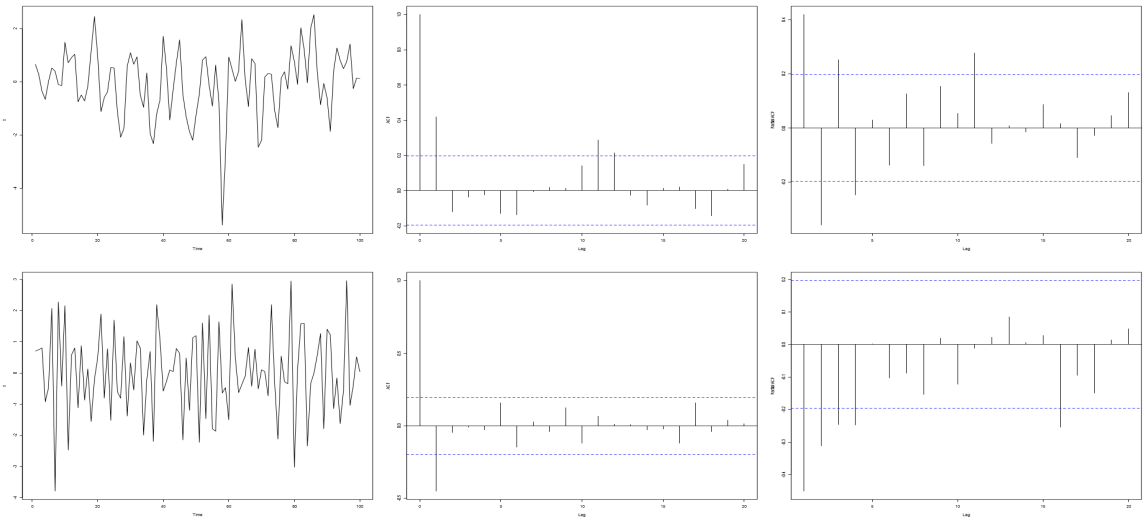

Figure 2.18: Top: simulated $MA(1)$ with $\theta_1 = 0.9$ with ACF AND PACF adjacent. Bottom: simulated $MA(1)$ with $\theta_1 = -0.9$ with ACF AND PACF adjacent. $\sigma_\epsilon^2 = 1$

Definition 2.6.3. A stationary time series can be generalised as an **ARMA(p,q)**

or *autoregressive moving average* process and follows

$$x_t = \alpha + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + \epsilon_t + \theta_1 + \epsilon_{t-1} + \cdots + \theta_q \epsilon_q \tag{2.36}$$

with $\theta_q \neq 0, \phi_p \neq 0$ and $\sigma_\epsilon^2 > 0$. Once again, we simulated the model, as shown in Figure 2.19, the patterns in the series are now an obvious combination of both models. With regards to financial data, we have already seen that time series are

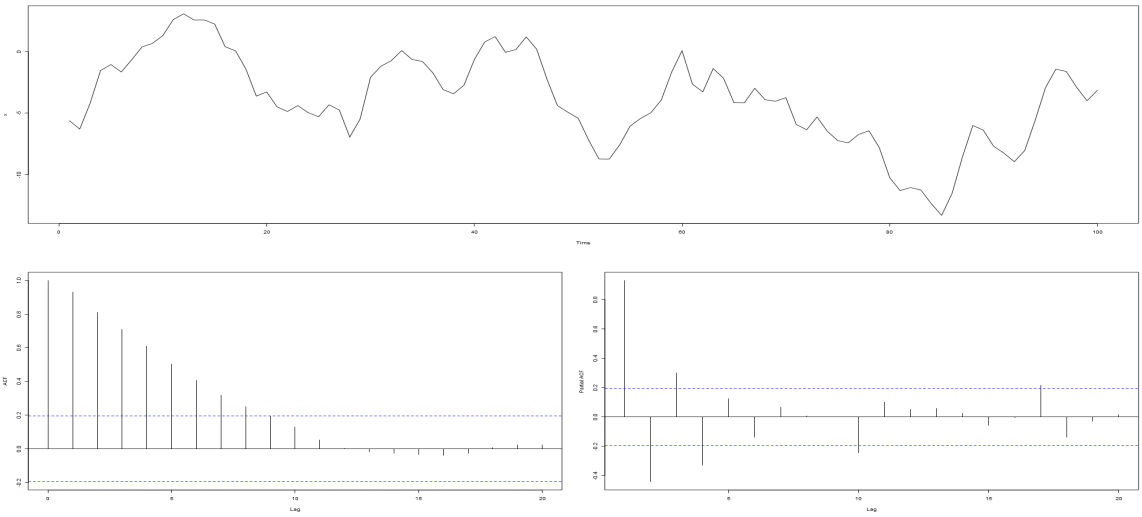

Figure 2.19: Top: simulated $ARMA(1,1)$ with $\theta_1 = 0.9, \phi_1 = 0.9$ with ACF AND PACF shown below. $\sigma_\epsilon^2 = 1$

often non-stationary and so in order to fit a model, any unit roots or trends must be removed. a non-stationary series that is made stationary through the difference operator is known as an integrated series or $I(d)$ where d is the number of differences required to stationarise the series, and so a further generalisation can be made to $ARMA(p,q)$in that a non-stationary process can be described as an *autoregressive* ***integrated*** *moving average* process or $ARIMA(p,d,q)$ where d is the order of difference in the series.

There is a final model that we must consider, and that is the discrete *random walk* model of a time series

$$x_t = x_{t-1} + \epsilon_t \tag{2.37}$$

which is essentially an **ARIMA(0,1,0)** model or an **AR(1)** with $\phi_1 = 1$, a random walk is strictly non-stationary.

ARIMA processes are useful in both fitting models to dynamic systems and creating forecasts, but as we'll see in chapter 3, they can help explain non-independent errors after fitting a particular model.

2.6.1 Parameter Estimation

For very low order ARIMA models, its generally enough to study the ACF and PACF of the data to estimate the exact order and parameters and magnitude of the coefficients; however its immediately obvious that when multiple orders of positive and negative coefficients are expected, it would be too difficult to fit a model with certainty. A routine method of model fitting is to fit a number of models and use parameter estimation techniques such as maximum likelihood estimation (MLE) to compare them and decide which fit is correct.

The likelihood function of a set of parameters for a time series $\{x_t\}$ is defined as

$$\begin{aligned} L(\theta) &= P(\{x_t\}|\theta) = P(x_0|\theta)P(x_1|\theta)...P(x_n|\theta) \\ &= \prod_{t}^{n} P(x_t|\theta) \end{aligned} \tag{2.38}$$

If we consider **ARMA(pq)** with $\epsilon_t \sim NID(0, \sigma^2_{epsilon})$ and a parameter vector β= (μ, ϕ_p, θ_p) the likelihood function can be written as

$$L(\boldsymbol{\beta}, \sigma^2_\epsilon) = \prod_{t=1}^{n} P(x_t|x_{t-1}, ..., x_1) \tag{2.39}$$

The conditional distribution P is a Gaussian with mean x_t^{t-1} and variance $v_t^{t-1} = \sigma^2_\epsilon r_t^{t-1}$ where r is a function of the model parameters. Without explicit derivation the likelihood can be written as

$$L(\boldsymbol{\beta}, \sigma^2_\epsilon) = (2\pi\sigma^2_\epsilon)^{-n/2}[r_0^1(\boldsymbol{\beta})r_2^1(\boldsymbol{\beta})...r^n - 1_n(\boldsymbol{\beta})]^{-1/2} \exp\left(- \frac{S(\boldsymbol{\beta}}{2\sigma^2_\epsilon)} \right) \tag{2.40}$$

where $S(\boldsymbol{\beta})$ is the conditional sum of squares

$$S(\boldsymbol{\beta}) = \sum_{t=1}^{n} \left\{ \frac{(x_t - x_t^{t-1}(\boldsymbol{\beta}))^2}{r_t^{t-1}(\boldsymbol{\beta})} \right\}. \tag{2.41}$$

If we take the partial derivative of the natural log of equation 2.40 with respect to $\boldsymbol{\beta}, \sigma_\epsilon$ and setting the result to zero we can find the parameters that fit the model. This is done numerically, and with the use of certain algorithms like the Kalman filter[28]. The main problem with using MLE to fit models is that we're comparing models with different numbers of parameters, this can be solved by using criteria measures such

Table 2.1: Model parameters estimation for copper commodity

(p,d,q)	ϕ_1	θ_1	θ_2	Log(L)	AIC	BIC
(0,1,0)	0	0	0	29500	-58997	-58989
(1,1,0)	-0.190±0.01	0	0	29728	-59453	-59437
(0,1,1)	0	0	-0.196 ±0.01	29742	-59480	59464
(1,1,1)	0.070±0.05	-0.26 ±0.04	0	29743	-59480	-59457
(2,1,1)	-0.10 ±0.16	-0.04±0.035	-0.09±0.16	29744	-59481	-59449
(2,1,0)	-0.19±0.009	-0.06±0.01	0	29744	-59482	-59459

as the *Akaike information criterion.*

$$AIC = 2k - 2\ln(\hat{(}L)) \tag{2.42}$$

or the *Baysian information criterion*

$$BIC = \ln(n)k = 2\ln(\hat{L}) \tag{2.43}$$

where k is the number of parameters, n is the number of data points in x and $\hat{L}$ is the MLE of the model. Given a set of models, the most likely fit is the one that has the maximum AIC or BIC. Both are criteria for goodness of fit, but also penalise for overfitting (i.e a higher order estimate of parameters), the BIC penalises overfitting more than the AIC.

As a preliminary, we used this method to fit a model on the copper commodity price we previously demonstrated, we fit a number of potential models and show the results in Table 2.1. The best fit model chosen from our MLE based criteria was an $ARIMA(0,1,1)$ with $\theta_1 = -0.105 \pm 0.001$, $\sigma_\epsilon^2 = 0.00053$ and $\mu = 0.04$. We then compared the result with the original series by looking at the ACF, PACF and the Fourier transform, as shown in Figure 2.20. We can see that the model results agrees fairly well especially when we look at the ACF and PACF, however with the FT, we see that there is a strong tailing to the higher frequencies of the series. its not fully understood as to why this occurred, the MA model wouldn't suggest stronger periodicity at high frequencies compared to lower ones.

To get a better picture of our result, we created 5 more simulations, this time taking intial conditions x_t at the 1st August 2016, and simulated the data until 1st February 2017 in order to see how accurate a forecast using this model would be, the results are shown in Figure 2.21. We can see that there is a tendency for the simulations to

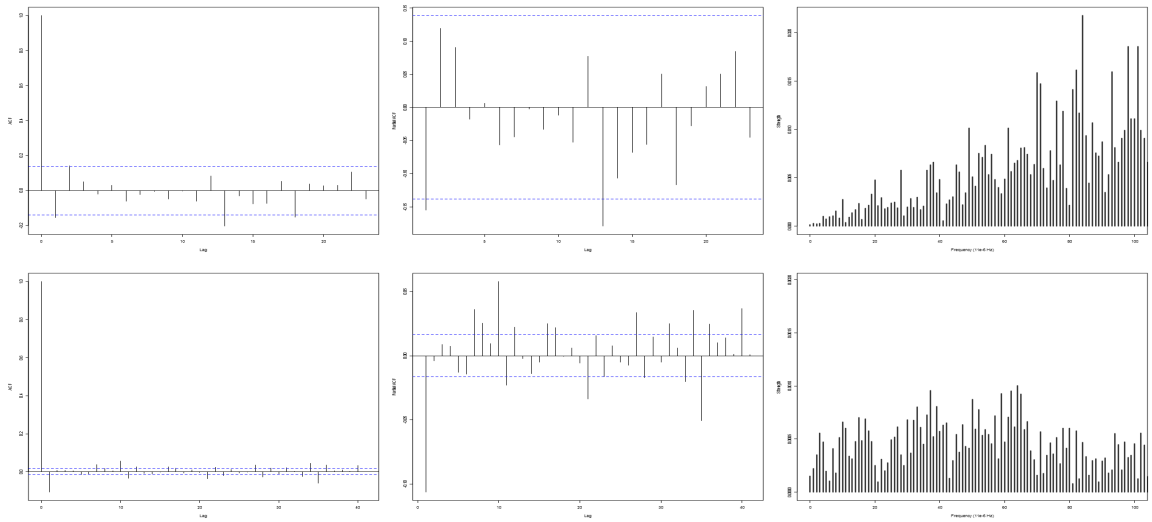

Figure 2.20: Top: ACF, PACF and DFT of simulated $ARIMAMA(0,1,1)$ fit of copper price. Bottom: Original ACF, PACF and DFT of copper.

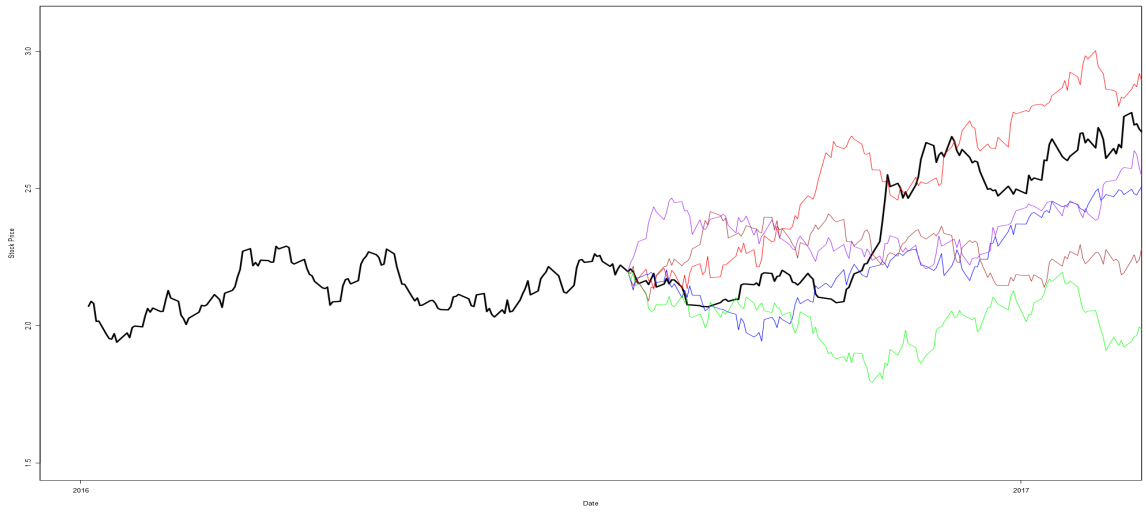

Figure 2.21: Price of copper Commodity from 1st January 2016 until 1st Febuary 2017, coloured lines show simulated trajectories beginning 1st August 2016.

follow the original trajectory, however there is also a large error with prediction. We can see this error more clearly if we create many more simulations and and use the ensemble to create an average trajectory, this is seen in Figure 2.22. The shaded area represents the 80% and 95% confidence intervals (assuming normality). What we see is that forecast accuracy greatly deteriorates with increasing time. This is a general weakness of stochastic time series models; these models may serve to explain more simple systems more accurately, but in order to describe more complex systems, such as a stock markets, we may need to branch out.

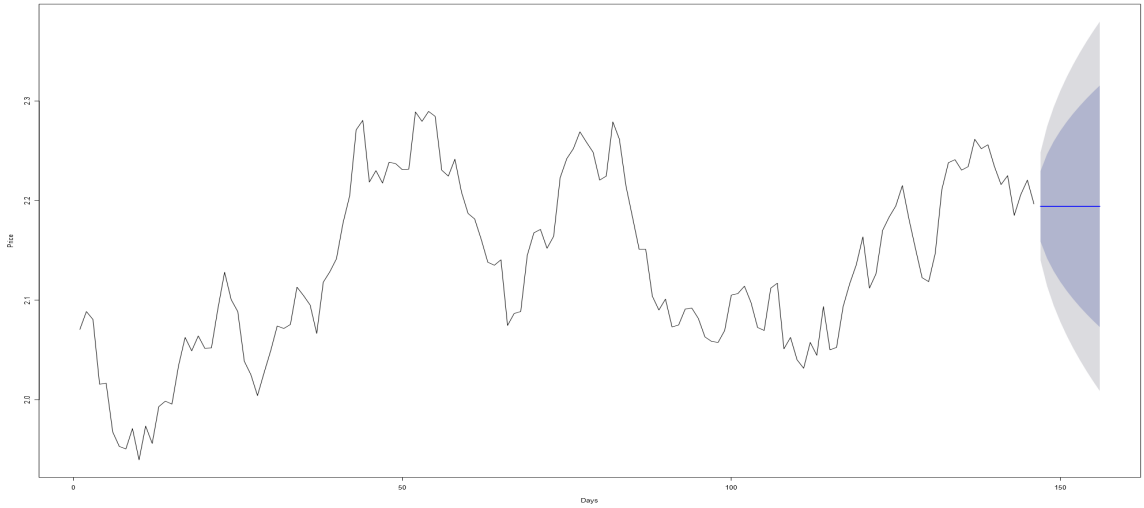

Figure 2.22: Price of copper Commodity from 1st January 2016 until 1st August 2016 with the addition of the forecast line. The shaded areas represent 80% and 95% confidence intervals.

When we applied the ARIMA model to a more complex market, the S&P500 IT sector, we found that the only plausible fit was an $ARIMA(0, 1, 0)$ with $\mu = 0$. This is not an unexpected result, in fact we showed earlier that the IT index showed very little autocorrelation in its daily price and so the closest linear model would be the random walk. With many stocks, this turns out to be the case, yet it is still useful to make inferences using these models, in order to make inferences that aid with our own model.

2.7 Multivariate Time Series Analysis

So far we have only discussed univariate models and analysis of time series. Multivariate analysis extends much of the same concepts we have seen and applies them to multiple time series. In this section we look to see the extent of correlation between markets, whether they belong to the same groups or not.

We began our analysis by applying ordinary least squares method of regression to three stocks that are present in the S&P500 IT sector, these are the companies (and their tickers) Google inc. (GOOGL), Apple inc. (AAPL) and IBM inc. (IBM) starting from 1st January 2000 until present day.

Before we can apply analysis techniques, we need to determine stationarity in the time series; we applied the stationarity tests described in the previous section. We

found that a difference of order one was suitable to achieve stationarity as ADF null hypothesis was rejected every time whilst the KPSS null was not, The results of the tests are shown in Table 2.2, and in Figure 2.24 we plot the resulting time series after differencing.

Table 2.2: Stationarity Test Results

	IBM	ΔIBM	AAPL	ΔAAPL	GOOGL	ΔGOOGL
KPSS ∼5	15.185	0.11312	20.195	0.1401	18.63	0.26493
$p(K)$	0.011	0.19	0.010	0.11	0.011	0.18
ADF, ∼5	-1.6691	-13.496	-2.2431	-12.784	-1.7026	-14.341
$p(A)$	0.732	0.01	0.4754	0.009	0.7042	0.087

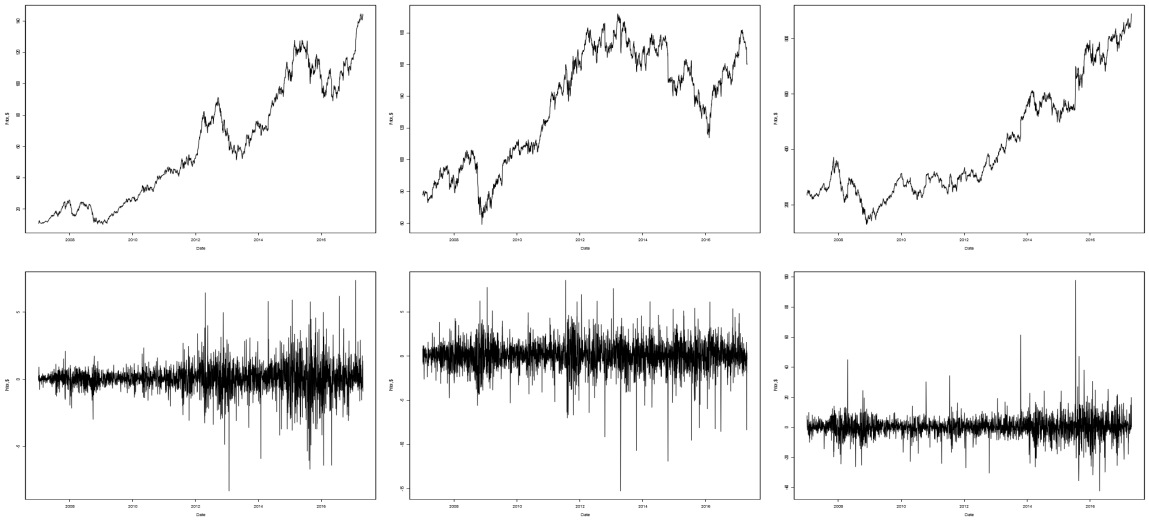

Figure 2.23: Three stocks series AAPL,GOOGL and IBM and their differenced transformation

The first, and most obvious analysis, is to fit a linear regression model, of the form

$$x_{(1,t)} = \alpha + mx_{(2,t)} + \epsilon_t \tag{2.44}$$

where $x_{(1,t)}$ and $x_{(2,t)}$ are two separate series and ϵ_t are independent and normally distributed errors or the *residuals* of the fit. A linear regression model is often sufficient if the errors show no autocorrelation, the presence of correlated errors would suggest that the fit may be non-linear in nature. Using ordinary least squares method (Minimising S from Equation 2.41), we plotted fits of the three stocks together, the results and the errors of the fit are shown in Table 5.1. The ρ is the (Pearson's)

correlation coefficient

$$\rho = \frac{Cov(x_{1,t}, x_{2,t})}{[V(x1,t)V(x_{2,t})]^{\frac{1}{2}}}. \tag{2.45}$$

The ρ can be used as a statistic in order to test the significance of correlation with the hypothesis

$$H_0 : \text{True Correlation} = 0; H_A : \text{True Correlation} \neq 0.$$

p values can be deduced by solving the Pearson's distribution integral numerically or by looking up values in available Tables [29]. We can see that for every fit, the null hypothesis is rejected with over 99% confidence.

Table 2.3: Regression Results

	$m \pm \delta m$	$\alpha \pm \delta\alpha$	ρ	$p(\rho)$
AAPL vs IBM	0.22046 ± 0.01169	0.04416± 0.02016	0.3472757	2.2e-16
AAPL vs GOOGL	0.06221± 0.00289	0.03547 ± 0.01982	0.3892971	2.2e-16
IBM vs GOOGL	0.095585 ± 0.004573	0.007615 ± 0.031351	0.3462729	2.2e-16

We plotted the fits along with the ACF of the residuals in Figure 2.24. What we observed is that the see that there appears to be some, albiet small, forms of autocorrelation in the fit residuals of all three models. This can be interpreted as either a goodness of fit measure, where there is still room to improve our regression fits or that there is some non-linear component in the correlation between the stocks.

These observations take into account the ensemble as a whole, we can understand the correlations more clearly by plotting correlations based on a rolling window of the series. We applied rolling windows of 1 month (20 data points) to each observation, and calculated both the correlation coefficient and the regression slope m. The results are shown in Figures 2.25 and 2.26.

Clearly, neither regression or the correlation coefficient are uniform over time. Correlation and regression both change with one another over time; this means that when two stocks become highly correlated i.e. co-moving, they also exert movements to one another to a higher degree, this leads us to believe there may be a force, outside the system that causes this effect. Even more interesting is the comparison between the measures, for instance during February of 2012 both ρ and m of all three stocks switched signs into negative regions, this occurrence is seen throughout the series, we can infer that the change in correlation between two related stocks X, Y stocks

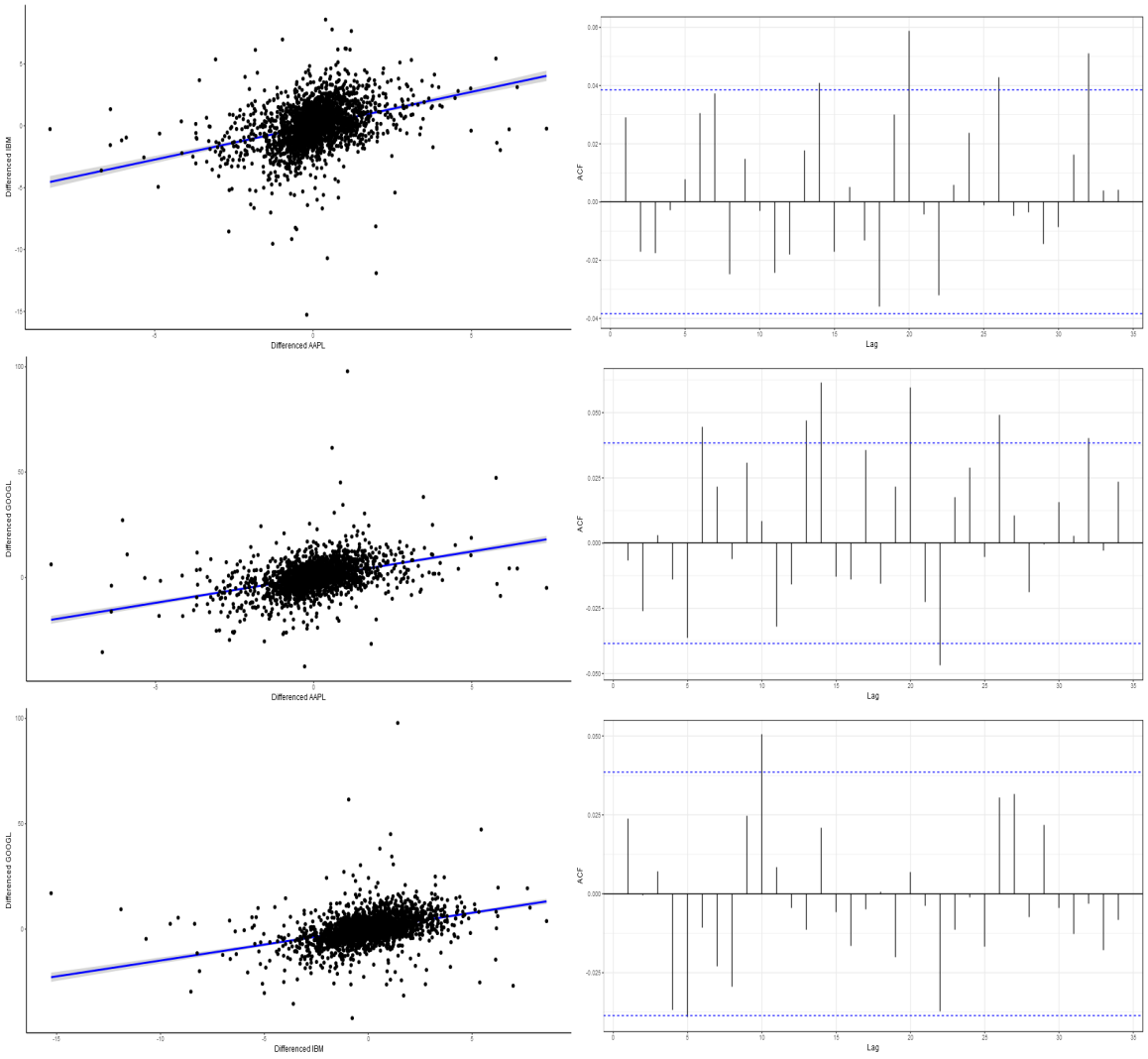

Figure 2.24: Linear Regression plots (with errors shaded around trendline) with ACF of residuals adjecent. Top: AAPL vs IBM. Middle: AAPL vs GOOGL. Bottom: IBM vs GOOGL

is also related to the change in correlation between one of the stocks and a third stock, seemingly related stock Z. This suggests that these related stocks belong to the same dynamic system. It is not immediately clear as to what causes the signs of both parameters to switch at certain periods, but we can infer that this could be more evidence for a complex, non-linear system. We have not yet considered if two series are correlated over time periods; in fact its more useful when building a dynamic model, to see what the effect of two series are over time. One method is to study the CCF or *cross correlation function*; CCF is essence just an extension of the ACF but correlated with a different series, and negative lag values are allowed in the

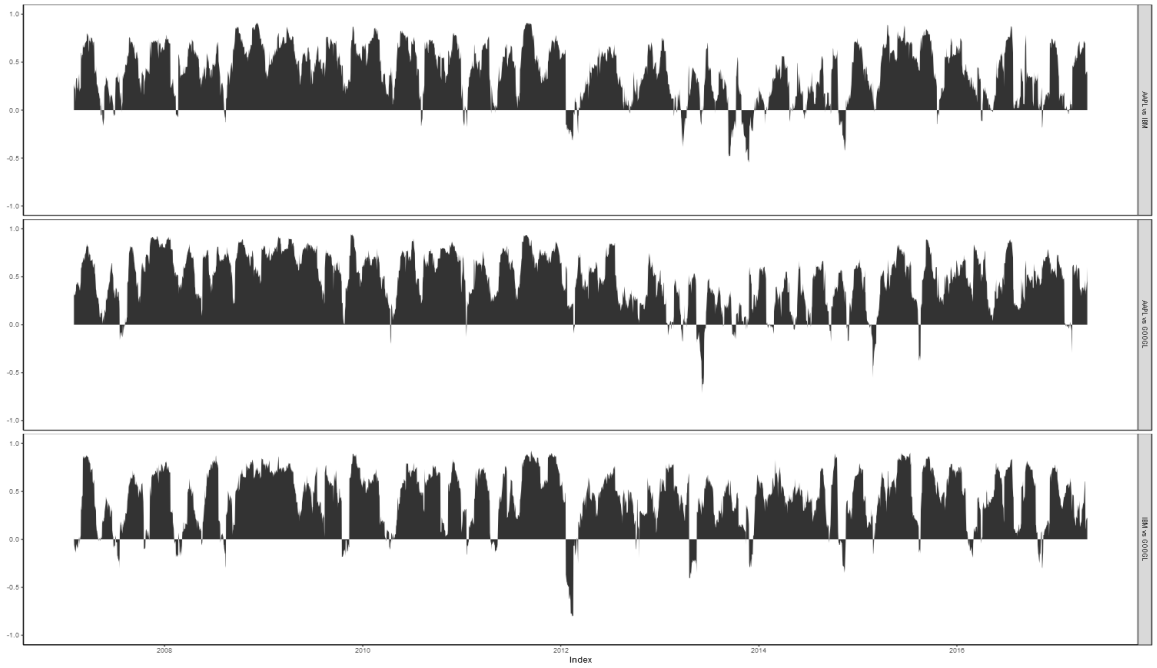

Figure 2.25: Rolling Correlation ρ of AAPL, IBM and GOOGL via a one month window.

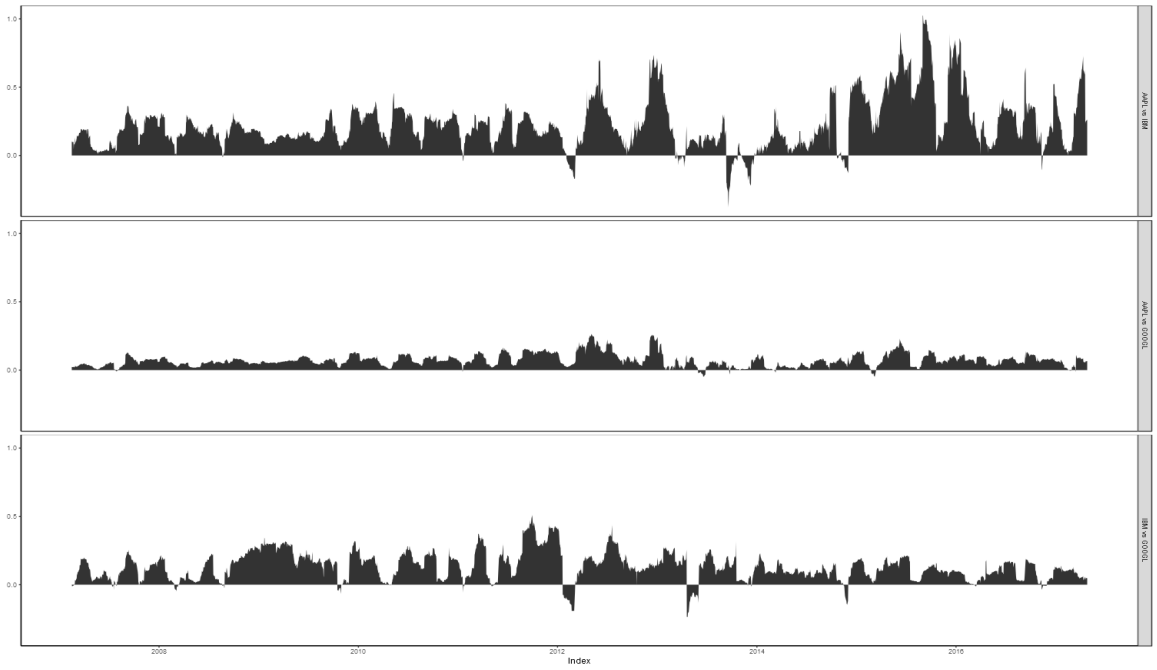

Figure 2.26: Rolling Regression of AAPL, IBM and GOOGL via a one month window.

correlogram. The CCF is defined

$$\rho_{XY}(\tau) = \frac{E[(X_t - \mu_X)(Y_{t+\tau} - \mu_Y)]}{\sigma_X \sigma_Y} \tag{2.46}$$

. where X and Y are two jointly stationary processes and τ is the number of lag shifts in Y. If the correlogram shows significant correlation with positive τ, we can see that there is in essence a τ 'time delay' in how X effects Y; for negative τ the time delay is on X. If correlation in one region is higher than another, it can be inferred that one variable leads the other. If positive τ has more correlation than negative τ, X is said to be leading the lagging Y. If the both negative and positive τ are similarly distributed, causation cannot be inferred and so a different variable must be analysed. Insignificant correlation would suggest that two variables change instantaneously to each other (or within one day). In Figure 2.29 we plot the CCF for three stocks. One the left are CCFs from a one year window of 2016. On the right is the entire series from 2007-2017, with max $\tau = 1500$. What is clear is that the highest correlation is at $\tau = 0$ for each stock, this tells us that most of the 'information' travels between stocks within at least one day, however significant correlation (and anti-correlation) is visible for small $|\tau|$ around 0, suggesting that a small fraction of how stocks effect each other is delayed in time. For small data sets and time windows (Figure 2.29 left) these delayed correlations are hardly significant however when studying long term movements between stocks over a long lag window, we notice these lagged correlations of small τ hold significance; we also observe that there is a steady decay over τ. From this we can infer that market information can take time to fully develop. In the microscopic picture this is attributed to the agent interactions when trading, when one stock return suddenly increases an agent may take this to be a good omen and invest into the market and similar stocks, driving those prices higher, leading to the instantaneous correlation; other agents may take their time to gather more information (sacrificing higher potential rewards) before making the decision to invest into the market. Furthermore, we see that in the case of AAPL vs IBM and IBM vs GOOGL there is a visible asymmetry in the CCF, in both cases the stock IBM is being lead by two series, allowing us to conclude that the change in stock price from AAPL and GOOGL *causes* the price to also change in IBM. For the case of AAPL vs GOOGL, the CCF is approximately symmetric, no causation relationship can therefore be inferred. Next, we plotted the rolling CCF based on a rolling window of one year and one month with a maximum τ of ± 1, the results are shown in Figure 2.28.Lagged correlations are not fully uniform over time but are similar. Interestingly, there are times when the $\tau = 0$ correlations are small, the $\tau \pm 1$ correlations become larger, this is seen predominantly at the start 2014 in AAPL vs IBM and AAPL vs GOOGL. This suggests that there are periods of time where information travels slower between two stocks; looking at the 20 day window

we see that these in fact these periods appear frequently. We next looked at the

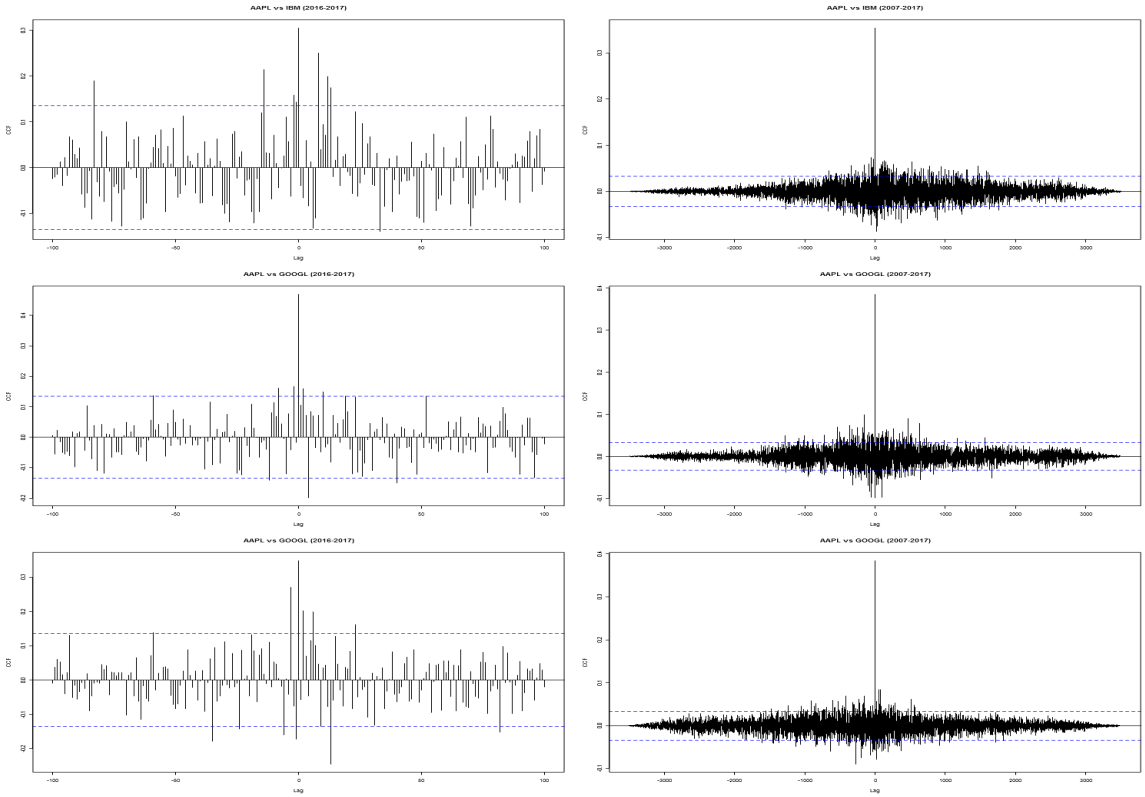

Figure 2.27: Cross correlation functions of AAPL, IBM and GOOGL.

CCF between less related markets, we wished to look at how stocks from two different sectors correlate over time and so we used the S&P500 IT and Industrial sector indices as the time series to analyse. Figure 2.28 shows this, one the left of the Figure is the CCF from 2016-2017 with a max $\tau = 100$. We see a highly positive instantaneous correlation between the two stocks, furthermore we see an asymmetry in the CCF favoring the IT index being the leading series (There is also a very significant anti correlation at $\tau = -4, \rho_{XY} = 0.24$). This asymmetry is seen in the right of Figure 2.28, however to a slightly smaller degree. No causation relation can therefore be inferred. The CCF is strictly a measure of linear correlation, we need to utilise other tools to assess whether a non-linear correlation exists between series. Several methods exist, such as Spearman's Rank [30], copula[31] or mutual information. We utilise the latter, mutual information in a similar fashion as the CCF. Given two series X_t and Y_t, the mutual information I is a measure of the redundancy of information contained between the two series, similar to the CCF, we can vary one series as $Y_{t+\tau}$ and measure the redundancy of information as a function of τ and so the average

mutual information is given by

$$I(\tau) = \sum_{t=1}^{T} P(X_t, Y_{t+\tau}) \log_2 \frac{P(X_t, Y_{t+\tau})}{P(X_t)P(Y_{t+\tau})} \tag{2.47}$$

Where $\mathrm{P}(X_t, Y_t)$ is the joint probability of X and Y and $I(\tau) \geq 1$. The difference between the CCF and MI method are that with MI, we are assessing what non-independence does to the joint probability of two series. MI is not concerned whether the series are correlated linearly, but it cannot tell us if two datasets are anti-correlated. To calculate the MI, an estimate for both probability distributions and joint distributions needs to be determined. In Figure **??** the probability density histogram for both series are shown.We fit a Gaussian function to the histogram;. We find that the Gaussian fit does not correspond to the normal distribution, i.e for a Gaussian Function

$$f(x) = a \exp\left(- \frac{(x-b)^2}{2c^2} \right) \tag{2.48}$$

The relationship $\frac{1}{\sqrt{2\pi c^2}} = a$ is not satisfied, i.e Gaussian is non-normal. This is shown in the Figure with an additional Normal fit using the parameters μ and σ. The fitting parameters are shown in Table 2.3. We can show the difference between the distributions more explicitly by plotting the quantiles of a theoretical normal distribution against the same data in a *Q-Q plot*. in Figure 2.31 we show the resulting plot; it is visible that around there distribution tails, a significant deviation from the linear normal line is present in both distributions. This result agrees with previous analysis in industry [32].

To calculate the joint probability distribution, we approximate with a 2D histogram at each τ, the joint distribution for $\tau = 0$ is plotted in Figure 2.34.

Table 2.4: Gaussian Fit

	$\mu \pm \delta\mu$	$\sigma \pm \delta\sigma$	$b + \delta b$	$c + \delta c$	$a + \delta a$
IT index	0.39 ± 0.18	6.13 ± 0.10	0.56 ± 0.01	4.31 ± 0.01	0.042 ± 0.001
IND index	0.22 ± 0.10	3.87 ± 0.08	0.32 ± 0.01	3.02 ± 0.01	0.062 ± 0.001

Figure 2.32 demonstrates the MI correlogram, we see the characteristic peak at $\tau = 0$ demonstrating that the highest correlation between the two indicies are indeed instantaneous, the pattern surrounding the peak now differs only a small degree. Its not possible to infer whether this is evidence for a non-linear relationship to any certainty.

Figure 2.28: Rolling Cross correlations of AAPL, IBM and GOOGL with both yearly and monthly windows shown; yearly windows correspond to roughly 260 trading days while monthly are 20 trading days.

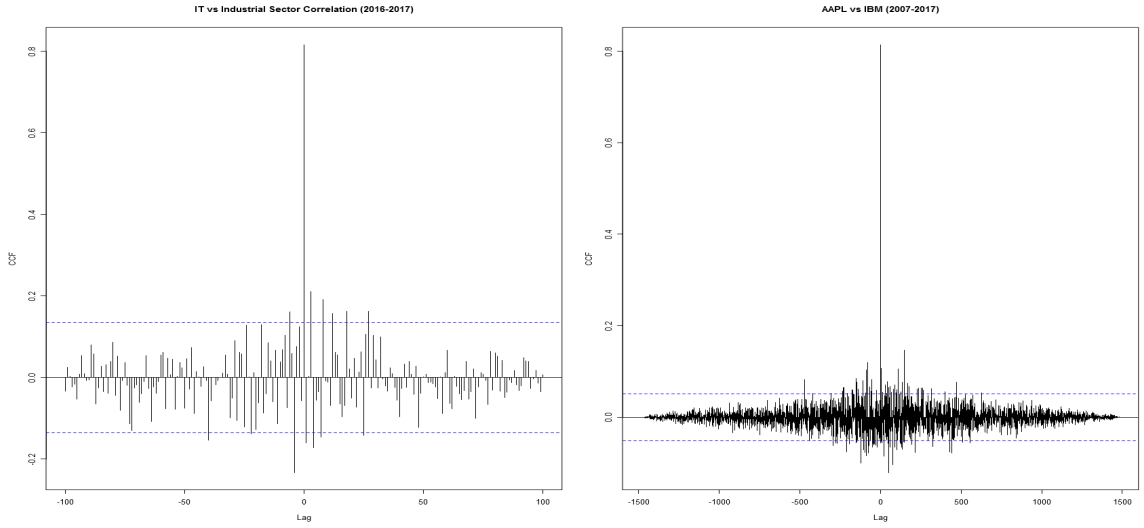

Figure 2.29: Cross correlation functions of S&P500 IT vs IND index.

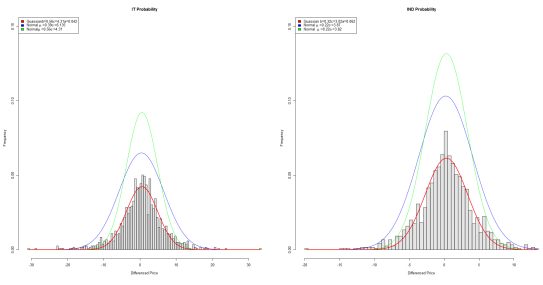

Figure 2.30: Estimated Probability density functions for S&P500 IT index (left) and IND index (right)

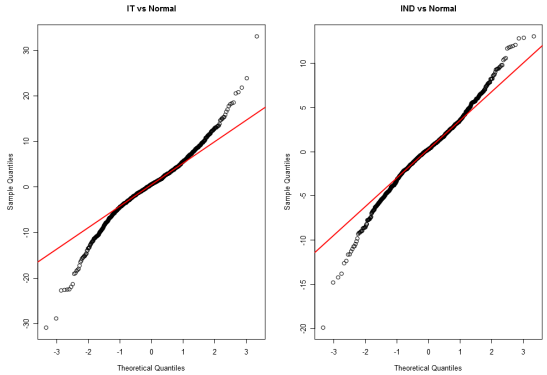

Figure 2.31: Q-Q plots of IT(left) and IND(right). Straight line is the perfect normal distribution

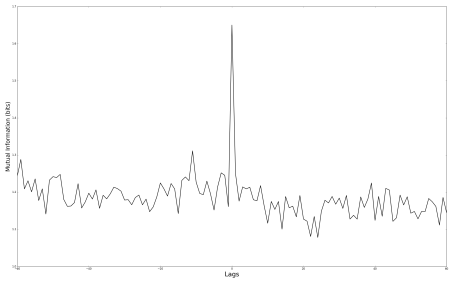

Figure 2.32: Mutual Information Correllogram of IT vs IND.

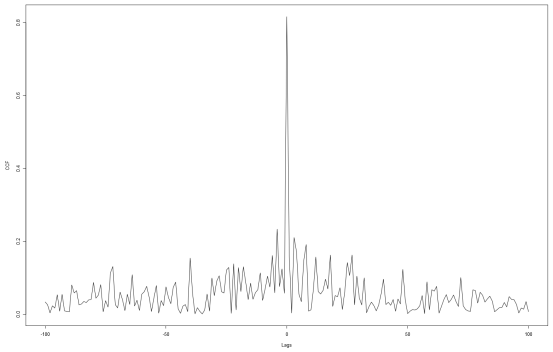

Figure 2.33: CCF Correllogram of IT vs IND.

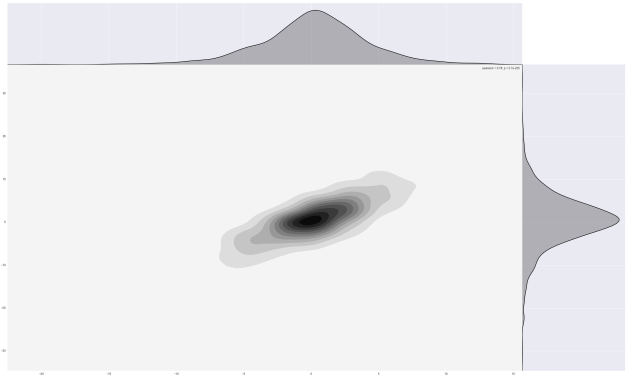

Figure 2.34: Estimated Joint Probability density functions for S&P500 IT index and IND index.

Chapter 3

Non-linearity and Chaos

In the analysis of the S&P500 IT sector index, we observed that the lack of linear autocorrelation lead to the the classification of a random walk (integrated white noise) series in the $ARIMA$ landscape, closely agreeing with the efficient market hypothesis. Many view this theory as brushing aside questions that are deemed too difficult to answer and look to use non-linearity and chaos to explain the seemingly random.

A requirement of chaotic dynamics is non-linearity, linear models can only generate dynamics of a combination of [oscillatory,non-oscillatory] $\times$ [stable, explosive]. Non linear can generate much richer types of trajectories such as sudden bursts of volatility or crashes. Another requirement is that the system is very sensitive to initial conditions.

Non-linearity is not exclusive to deterministic chaos, however, stochastic systems can also be non-linear. The distinction is defined

$$x_t = f(x_t, x_{t-1}...) + \epsilon_t \text{ Stochastic non-linear} \tag{3.1}$$

$$x_t = f(x_t, x_{t-1}...) \text{ Deterministic non-linear} \tag{3.2}$$

where f is a non-linear function of x and its past values. We can give a brief intuition to chaos here from the simplest model, the *Tent map*. The tent map is a combination of piece-wise linear functions that make a non-linear system, in the form of

$$x_t = \begin{cases} \mu x_{t-1}, & ,x < 0.5 \\ \mu(1 - x_{t-1}), & x \geq 0.5 \end{cases} \tag{3.3}$$

The tent map is chaotic is exactly in the phase $\mu = 2$. The tent map takes the

interval of (0,1), stretches it to twice its length, and folds it in half, this stretching folding is repeated, pulling apart points that are close together and leading to a system that is unintuitive to predict, but not in anyway random. We can demonstrate the sensitivity of initial conditions by providing some simulations, in Figure 3.1 we plotted four simulations, the first three have initial conditions that are different by only a small percentage and the final differs much more, $x_0 = [0.91587624823323, 0.917525262363, 0.913000000000, 0.01114423323]$. As $t \to \infty$, x_t fills the unit interval [0,1] uniformly, i.e the proportion of x_t falling into an interval [a,b] is (b-a) for any $0 < a < b < 1$.The smallest difference in initial condition, as observed, will diverge trajectories exponentially fast. The plotted ACF shows no significant serial correlations, this is in fact a characteristic of chaotic non-linearity. A

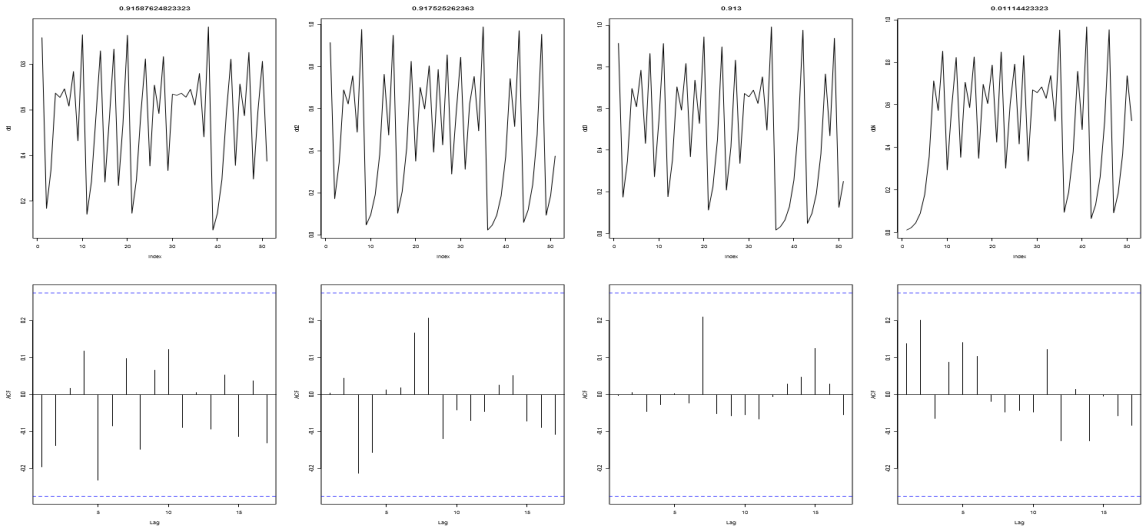

Figure 3.1: Simulated tent map with intial conditions shown, the autocorrelation is plotted below.

systems can be redefined in terms of its phase space (or state space), the equilibrium state to which a system evolves is called the *attractor*. For a damped system, such as a real pendulum the attractor tends to a point. Pendulums with replenished energies have periodic attractors. With a chaotic attractor, equilibrium applies to a region, rather than a point or orbit. Indeed chaotic equilibrium is a very dynamic set of states. If a trajectory starts with intial conditions in the attractor's basin, it will eventually fall within its periodic phase space. Chaotic attractors often display special symmetries with self-similarties on multiple scales called *fractals*. For a number of simulations we plotted the attractor for the tent map (3.2) , and one can see where the name originates.

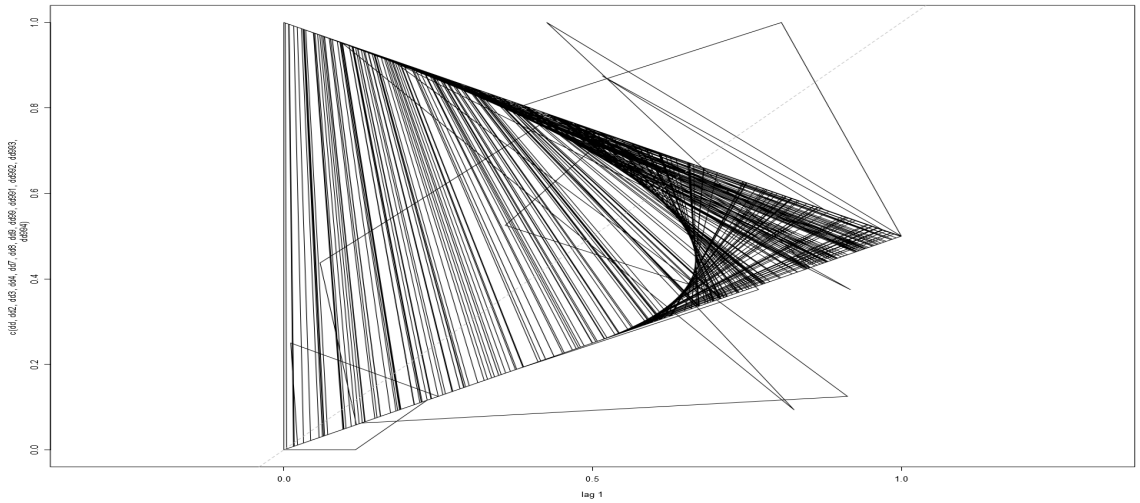

Figure 3.2: Tent map attractor, with $x_t vs x_{t-1}$, the lines show movement between points on the lines of the "tent" (no points are inside). The outliers are a result of computational shortcomings with regard to high decimal places.

3.1 Tests for non-linearity

3.1.1 The BDS test

The BDS test, named after its developers Brock, Dechert and Scheinkman is the most popular test for non-linearity, first published by the group in 1996 [33]. Originally developed to distinguish data that is independent and identical distributed from non-linear chaos; however since then it has deemed suitable as a test for non-linearity in general [34]. It can be used as goodness of fit measure for fit residuals (even when autocorrelations are insignificant). Indeed the strength of the BDS test is that it is a statistical test, with a null hypothesis that data is IID (white noise). For a financial time series that has either been fitted with a linear model so that autocorrelations are 0, or one that was found to be white noise from a linear prospective - non-linearity can be determined after rejecting the null hypothesis.

The BDS test is based on the *correlation integral*. This is a measure of the average 'closeness' of two states within the attractor. For a time series x_{t0}^{T} with a defined embedding dimension (A phase space vector) or its $m-$history $\boldsymbol{(x_t^m)} = (x_t, x_{t-1}, .., x_{t+(-m+1)})$, the correlation integral as a function of m is defined as

$$C_{m,\epsilon} = \frac{2}{T_m(T_m - 1)} \sum_{m \leq s} \sum_{< t \leq T} I(\boldsymbol{x_t^m}, \boldsymbol{x_s^m}; \epsilon) \tag{3.4}$$

where $T_m = T - m + 1$ and $I(x_m^t, x_s^m; \epsilon)$ is equal to one if $|x_{t-i} - x_{s-i}| < \epsilon$ for $i = 0, 1, ..., m-1$ and zero otherwise. We are estimating the probability that any two m dimensional points are in a distance of ϵ. i.e

$$p(|x_t - x_s < \epsilon, |x_{t-1} - x_{s-1} < \epsilon, ..., |x_{t-m+1} < x_{s-m+1}| < \epsilon) \tag{3.5}$$

if x_t are white noise (iid), this probability should be equal to

$$C_{1,\epsilon}^m = P(|x_t - x_s| < \epsilon)^m \tag{3.6}$$

The BDS statistic for a time series is defined as

$$V_{m,\epsilon} = \sqrt{(}T)\frac{C_{m,\epsilon} - C_{1,\epsilon}^m}{\sigma_{m,\epsilon}} \tag{3.7}$$

where σ is the standard deviation of the numerator. The BDS statistic is distributed normally: $N(0, 1)$ and so the null hypothesis is rejected at 5% whenever $|V| > 1.96$. The embedding dimension is an intrinsic part of the system, in the tent map attractor, $m = 2$ as we could describe the entire system with two dimensions. We perform the BDS test on the S&P500 IT and IND index, a simulated white noise series and the data from the tend map as a comparison, all 4 series showed know linear autocorrelation. In table 3.1 the results are shown for the BDS test for a number of m, we find that the null of iid is rejected for both the IT and IND index, suggesting very high likelihood of non-linearity in the system. The null is rejected for the tent map data and non rejected for white noise. ϵ is chosen as a fraction of the standard deviation of the sample series. We chose a time difference $(t - s = 1)$ as there was no significant autocorrelation in any series.

Table 3.1: BDS test results, for embedding dimension up to 5, ϵ is chosen as a fraction of standard deviation

m		ΔIT				ΔIND				Tent Map				ω_t			
		0.5σ	σ	1.5σ	2σ	0.5σ	σ	1.5σ	2σ	0.5σ	σ	1.5σ	2σ	0.5σ	σ	1.5σ	2σ
2	v	5.80	6.149	6.83	7.123	3.37	4.20	5.30	5.985	427.503	38.08	-2.52	-6.38	0.34	0.318	0.689	0.561
	$\rho(v)$	0	0	0	0	8e-04	0	0	0	0	0	0.011	0	0.72	0.749	0.490	0.574
3	v	7.65	7.74	8.00	7.99	3.92	5.28	6.71	7.61	590.63	41.78	-4.12	-4.76	0.39	-0.34	0.15	0.12
	$\rho(v)$	0	0	0	0	1e-4	0	0	0	0	0	0	0	0.69	0.72	0.87	0.89
4	v	9.043	9.20	9.24	9.04	5.47	6.79	8.00	8.58	845.54	43.1696	-3.27	-1.68	-0.41	-0.78	-0.37	-0.37
	$\rho(v)$	0	0	0	0	0	0	0	0	0	0	0	0.09	0.68	0.43	0.70	0.70
5	v	10.35	10.51	10.36	9.98	7.11	7.90	8.94	9.3213	1246.99	42.8000	-3.3776	0.02	-0.45	-0.83	-0.66	-0.65
	$\rho(v)$	0	0	0	0	0	0	0	0	0	0	0	0.98	0.65	0.40	0.50	0.51

3.1.2 Correlation Dimension

The correlation dimension, a related concept, is a method developed by Grassberger and Procaccia (1986) [35] in order to detect chaos within a series. In the paper, they showed that

$$C(\epsilon) = Constant \times \epsilon^d \tag{3.8}$$

where d is called the correlation dimension and it is a measure of the fractal dimension of the system. For a particular m, the correlation dimension can be obtained via

$$d_m = \lim_{\epsilon \to \infty} \frac{\log(C_m(\epsilon))}{\log(\epsilon)} \tag{3.9}$$

This can be found by studying the slope of equation 3.8. We find that if d does not increase with m, the data is consistent with chaotic behaviour. the Grassberger-Procaccia dimension is more accurately defined as

$$d = \lim_{n \to \infty} d_n \tag{3.10}$$

Intuitively, if we consider the tent map; since the series is uniformly distributed over interval [0,1], $C_1(\epsilon$ doubles if ϵ does too. And so for small ϵ

$$d_1 = \frac{\log(C_1(\epsilon))}{\log \epsilon} = 1 \tag{3.11}$$

When the embedding dimension is increased to 2, phase vectors now fill up a triangle in phase space (as opposed to a square) and so

$$d_2 = \frac{\log(C_2(\epsilon))}{\log \epsilon} = 1$$

and the pattern continues as

$$d_n = \frac{\log(C_n(\epsilon))}{\log \epsilon} = 1$$

and so $d = 1$ for the tent map. In the case of a white noise series, we find that the $m = 2$ will fill a [0,1]×[0,1] unit square with the given distribution, this increases. this leads to the conclusion that $d = \infty$ for a random process. d essentially represents how much phase space is "filled up" by a series, and so they need not be integers. The importance of d is that the minimum number of variables required to model an attractor, is the smallest integer greater than d We calculated the correlation dimension for the IT, IND, tent map and the Gaussian white noise for m up to 10.

The results are plotted in Figures 4.4 and tabulated in table 3.2. The results from the white noise series showed the expected proportionality i.e $d = \infty$. The same result is found for the tent map, where the correlation dimension is estimated to be 1.024 ± 0.1; which agrees with the theoretical value. The tent map data does become distorted after $m = 4$. This is because embedding reduces the number of data points to analyse by a factor $1/m$. Our results showed that for up to $m = 10$, the correlation dimension did not converge and chaos could not be detected in the system using this method. From the graphs we can make the distinction that the increase in d is not linear with m, meaning that we are likely to find that chaos exists in the system but at a very high dimension. The implications of our result that the two systems are non-linear and non-chaotic at low dimensions are as follows

- A system can have non-linear components without necessarily being chaotic, these systems can be stochastic or deterministic (without exponential sensitivity to initial conditions).
- Chaos could potentially be detected at much higher dimension, however this result would be less useful as with very high dimensions, the usefulness of chaos diminishes greatly, and no great distinction is made with randomness.

There is still a likely hood that chaos exists in this system and has gone undetected, In the discussion section, explain some of the pitfalls with the correlation dimension method and potential for future work.

Table 3.2: Results from the correlation dimension, "NaN" values represent insufficient data for embedding

m	d_{IT}	d_{IND}	d_{ω}	\$d$_{tent}$
1	1.175	1.113	0.998	0.911
2	1.945	2.126	1.993	1.200
3	3.371	3.207	2.977	0.984
4	3.353	3.282	3.858	1.001
5	4.598	4.390	4.784	2.998
6	5.583	5.1318	5.601	3.278
7	6.619	6.498	0.396	NaN
8	7.290	7.226	7.188	NaN
9	8.559	6.624	8.318	NaN
10	8.702	7.43	1.252	NaN

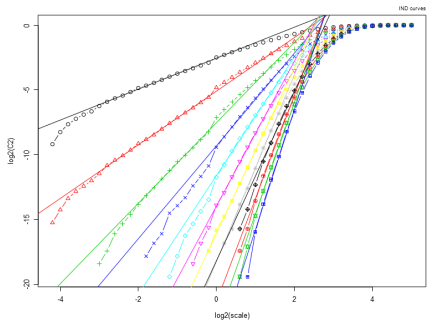

Figure 3.3: Correlation Dimension of IND index

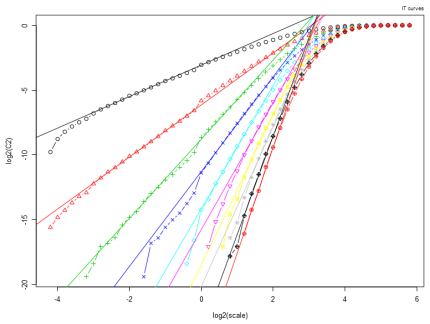

Figure 3.4: Correlation Dimension of IT Index

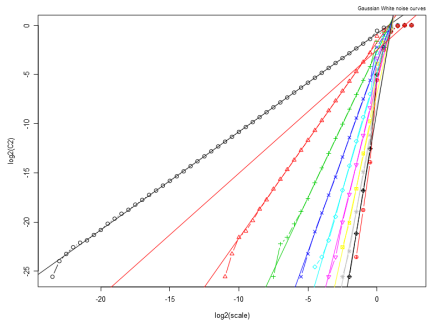

Figure 3.5: Correlation Dimension of white noise series.

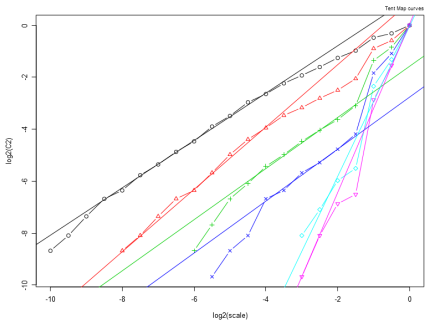

Figure 3.6: Correlation Dimension of tent map series.

Chapter 4

Building a Dynamic Model

The previous chapter built the foundations to make assumptions for a mathematical model, these foundations were particular inferences from stock market time series analysis. We found that wide-sense stationary stock market time series

- Show no prominent peaks in their frequency spectrum and can be estimated to be aperiodic.
- Trajectories have fast decaying memories, although partial autocorrelation is shown at certain lags. Auto-correlation for squared prices are significant and decay slowly.
- Linear stochastic models do not fully capture more complex stocks, random walk series does not account for partial autocorrelation nor the correlation between different stocks.
- Stocks belonging to the same system show high instantaneous correlation and smaller time delayed correlation that decays with increasing time difference.
- Linear regression shows non-independence in residuals, suggesting a non-linear relationship.
- Stocks from different sectors show correlation relationship and have variables sampled from a small region in probability space. Their probability distribution functions are asymptotically Gaussian but non-normal.

Our proposal was a set of simultaneous difference equations that simultaneously map the transformations of two discrete series $\{X\}_0^T$ and $\{Y\}_0^T$ from one state to

another, i.e.

$$Y_t \to Y_{t+1} \text{ and } X_t \to X_{t+1}$$

Analysis suggests that a time dependent, non-linear model that describes the dynamics of two related stocks may be appropriate, with this knowledge, we built a model under the following assumptions

- A series evolves over time with some contributions from its previous values, this contribution becomes significantly smaller with increasing time.
- Stock cross correlation varies with time, we hypothesize that this may be related to the magnitude of difference between one lag and the next.
- A stable dynamic recurrence relation can describe the system given a set of initial conditions and parameters.

4.1 Model 1

The initial proposal for this recurrence relationship is given by

$$Y_{t+1} = Y_t + \chi_t \Delta Y_t; \tag{4.1a}$$

$$\chi_t = \alpha \Delta X_t; \tag{4.1b}$$

$$X_{t+1} = X_t + \gamma_t \Delta X_t; \tag{4.1c}$$

$$\gamma_t = \beta \Delta Y_t \tag{4.1d}$$

Where Δ is the first order difference operator i.e. $\Delta Y_t = (Y_t - Y_{t-1})$,The product of time dependent coefficients χ_t and γ_t with the differenced Y_t and X_t, respectively, form a sum with the opposite variable to map determine the transformation. α and β are linear coefficients of related differenced stock at t. As stated in Chapter 1, we our goal was to model the movement of sectors and so we solved the above equations numerically for the S&P500 IT and Industrial sectors. We chose solve the above using data from the period of 01-01-2016 to 01-08-2016. The χ_t and γ_T were solved via the relationship

$$\chi_t = \frac{\Delta Y_{t+1}}{\Delta Y_t} \tag{4.2a}$$

$$\gamma_t = \frac{\Delta X_{t+1}}{\Delta X_t}. \tag{4.2b}$$

Figure 4.1 shows the form of these coefficients over time, we see that they fluctuate around zero, however in some instances spikes of very high magnitude are observed. These spikes occur as prices 'accelerate' in an instance, i.e ΔY_t is very small compared to ΔY_{t+1}. We used OLS regression to determine α and β the results are shown in

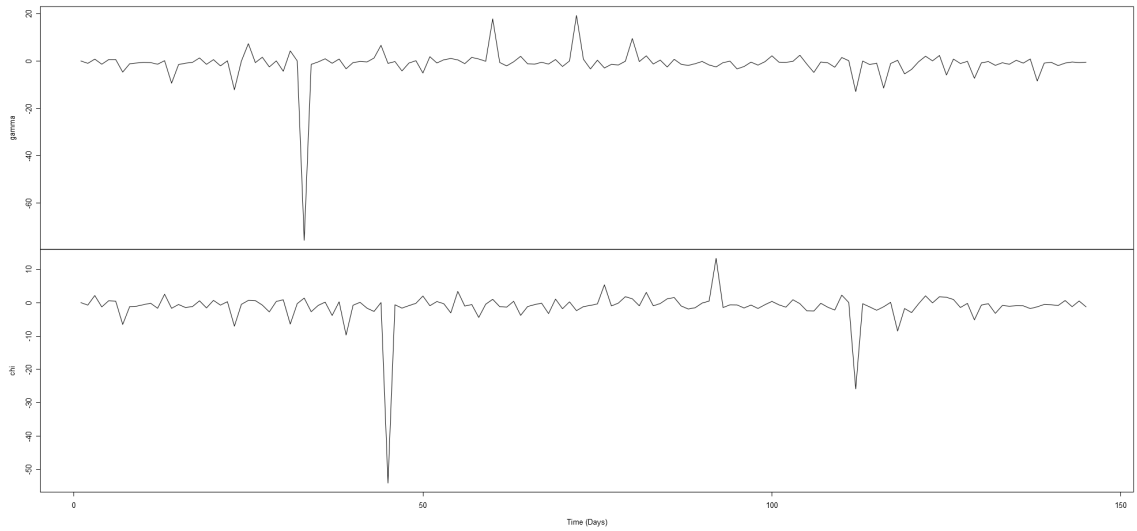

Figure 4.1: Top: $\gamma(t)$. Bottom: $\chi(t)$

Table 4.1. The Pearson's null hypothesis is rejected for γ_t regression relationship with 80% confidence, it is not rejected for the 95% confidence interval. For χ_t correlation is not rejected for either confidence intervals and therefore we cannot infer if there is a non-zero correlation. The results of the fits are shown in Figure 4.2. The relatively large error seen in the regression results is attributed to low correlation, especially for χ_t. Failing to reject the null hypothesis is not necessarily its acceptance [36], correlation could be masked by a non-linear or non-independent relationship or the lack of enough data points could to a lower $p(\rho)$. We move forward with the analysis using these results but we bare in mind the large bias that may be associated with this. Using the original time series as 'previous states' we simulate new values of X_t

Table 4.1: Regression Results

	$m \pm \delta m$	$\alpha \pm \delta\alpha$	ρ	$p(\rho)$
α	-0.0115 ± 0.100	0.0360 ± 0.101	0.003	0.72
β	0.16 ± 0.10	0.30 ± 0.35	0.15	0.17

and Y_t. We find that for these coefficients, the model converges to a fixed constant, so

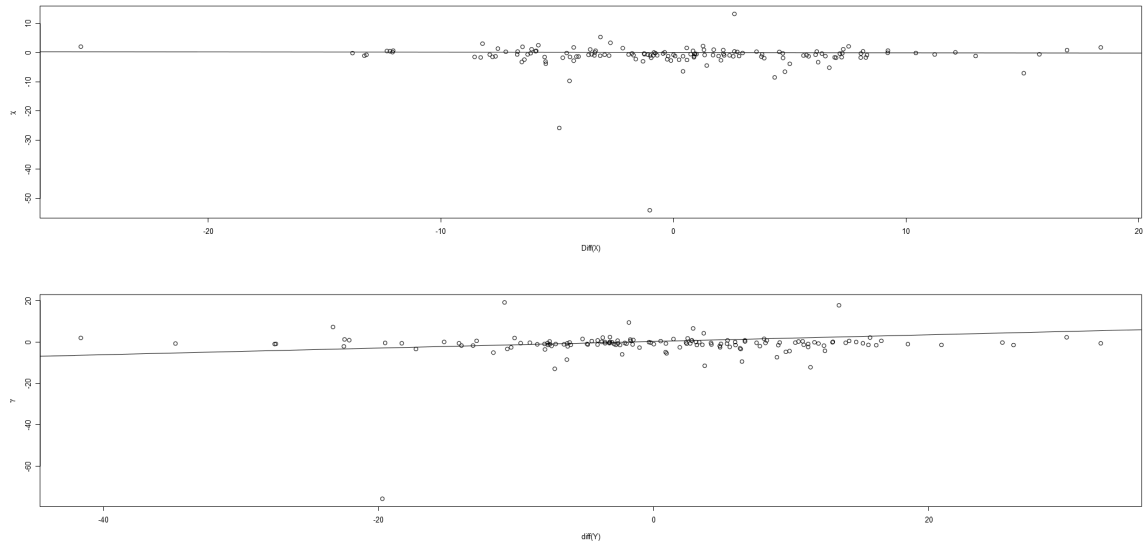

Figure 4.2: Top: $\gamma(t)$. Bottom: $\chi(t)$

therefore we can accept that either the parameters are unsuitable or the model itself is incorrect for our fixed initial conditions (the stock market). We ran numerical tests using combination of α and β in the range of $(-1:1)$ for multiple significant Figures and found that the parameters either lead to an unstable trajectory or a convergence to a fixed point. All recorded results are uploaded online see: [37][38]. So we can say that given our initial conditions, Equation 1.1 is not a recurrence relation for the two stocks and some modifications must be made.

We looked at the residuals of the regression, a requirement is that any error be negligible and completely independent of time, in Figure 4.3 we plotted the residuals and their ACF and PACF. Immediately it is visible that the residuals are serially correlated, the form of the residuals resembles an MA process we introduced in Chapter 2. Application of ARIMA parameter estimation, we identified both as an $ARIMA(0,0,1)$ with coefficients given by $\theta_{\alpha,1} = -0.8250 \pm 0.043$ and $\theta_{\beta,1} = -0.9335 \pm 0.0300$. With the addition of this as an error term, we must therefore add the regression constant of χ_t and γ_t to the model.

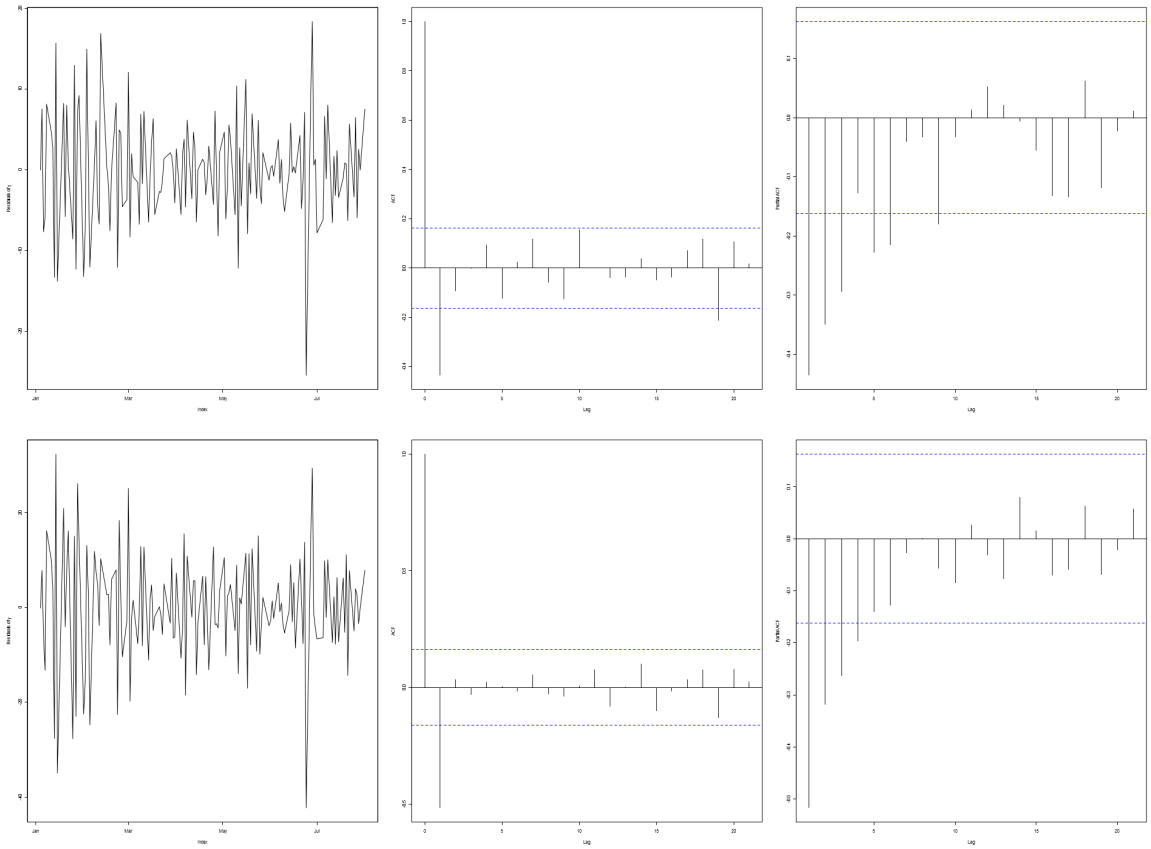

Figure 4.3: Residuals of the numerical solution for χ_t (top) and γ_t(bottom) with ACF and PACF shown.

4.2 Model 2

The modified model takes the form

$$Y_{t+1} = Y_t + \chi_t \Delta Y_t; \tag{4.3a}$$

$$\chi_t = C_\chi + \alpha \Delta X_t + \epsilon_t; \tag{4.3b}$$

$$X_{t+1} = X_t + \gamma_t \Delta X_t; \tag{4.3c}$$

$$\gamma_t = C_+ \gamma \beta \Delta Y_t + \omega_t \tag{4.3d}$$

Where C_χ and C_γ are the regression constants and ϵ_t and ω_t are the ARIMA models of the residuals. We made simulations using the modified model, in Figure 4.4 the simulation results for both time series are plotted. For IT index simulation: $\mu_{sim} = 0.18, \mu_{IT} = 0.39, \sigma_{sim} = 4.56 \sigma_{IT} = 6.14,$, for IND: $\mu_{sim} = 0.19, \mu_I ND = 0.22, \sigma_{sim} = 2.80, \sigma_{IND} = 3.86$. The simulation shows some of the same patterns as the real data,

however the variance is lower than expected. We determined stationarity through the ADF and KPSS tests, with results shown in Table 4.2. In Figures 4.5 and 4.6 we show the characteristic ACF, PACF, DFT and estimated probability distribution of the simulation, for IT and IND, respectively. In both series, the discrepancy in variance leads to a thinner density function. The ACF and PACF also show a significant correlation at $k = 1$, this is expected as map of new states is a linear combination of the previous state, however it is not understood why this is an anti-correlation. The frequency spectra shows no dominant peaks, and has a similar form as the original series. We run 4 more simulations and plot the integrated series to forecast the price over the next 8 month period. The results shown in Figure **??** show that the trajectories are unable to correctly predict the upward trend seen in the original series'. A statistical test for independence of two different series is the χ^2 [1] hypothesis test

$$\chi^2 = \sum_{i,j} \frac{(Observed - Expected)^2}{Expected} \tag{4.4}$$

with hypotheses

H_0 : Two data sets are independentH_A : Two data sets are non-independent

P values for the χ^2 test can be obtained from the χ^2 distribution. Table 4.4 shows the results the test, we find that for our simulations, we are not able to reject the null of independence, meaning that the model likely requires to be modified again.

Table 4.2: Stationarity Test Results

	IT Simulation	IND Simulation
KPSS ~ 3	0.22875	0.17292
$p(K)$	0.30	0.19
ADF, ~ 5	-4.7418	-5.7609
$p(A)$	0.01	0.01

Table 4.3: χ^2 test results

	IT Simulation	IND Simulation
χ^2	28392	28223
$p(\chi^2)$	0.2393	0.2402

[1] χ here is not related to the model coefficient

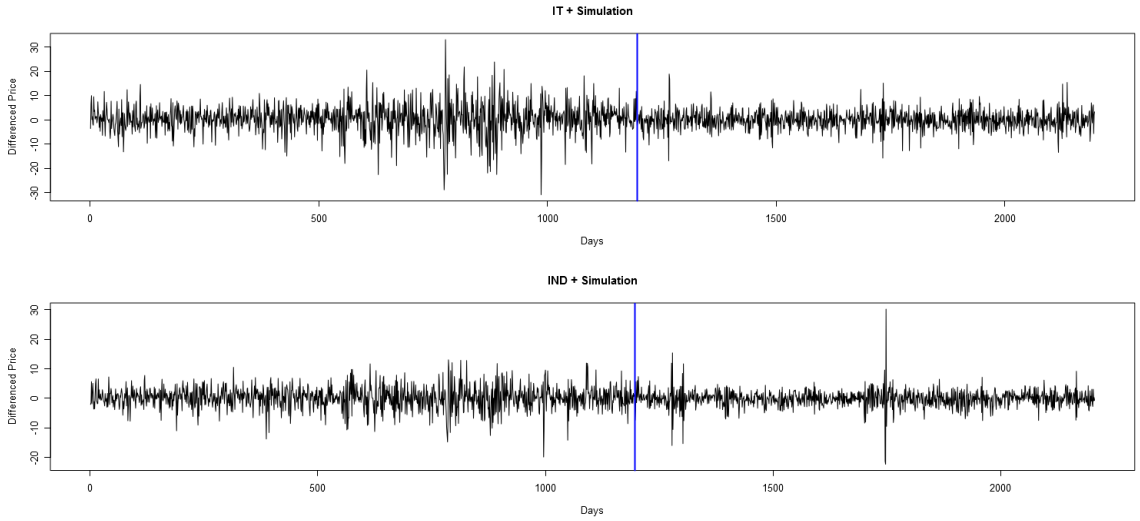

Figure 4.4: Original and simulation of differenced IT (top) and IND(Bottom), the blue verticle line shows where the simulation begins.

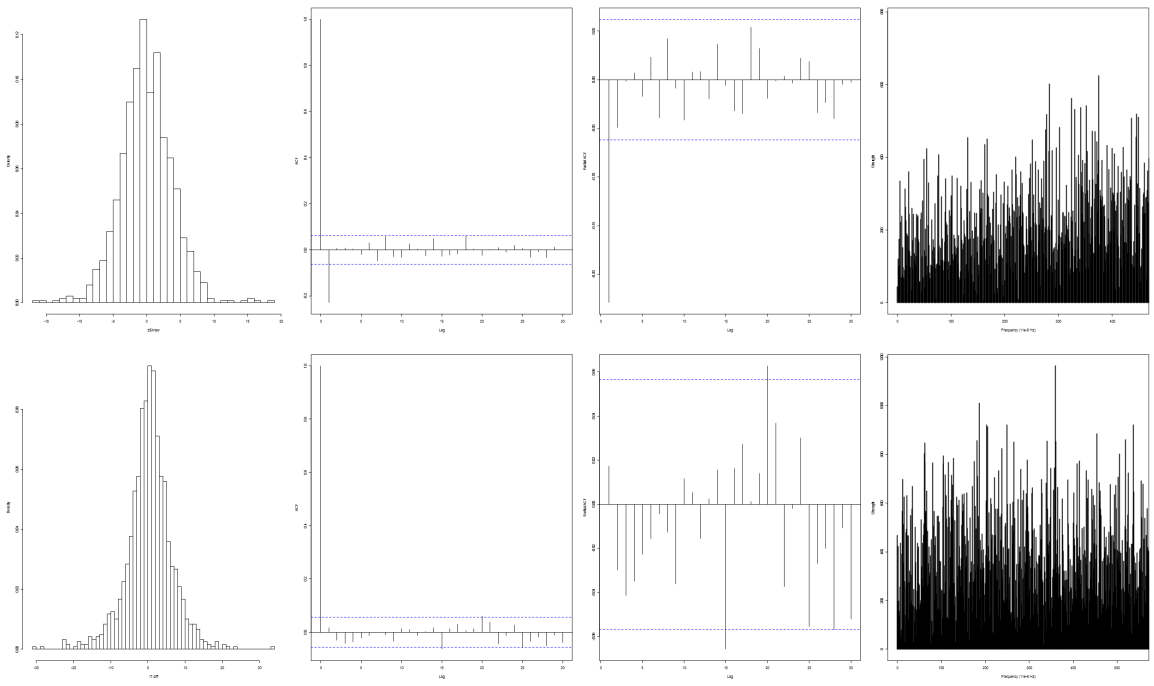

Figure 4.5: Characteristics of simulation (Top) vs IT (Bottom). In order: Probability desnity histogram, ACF, PACF and DFT

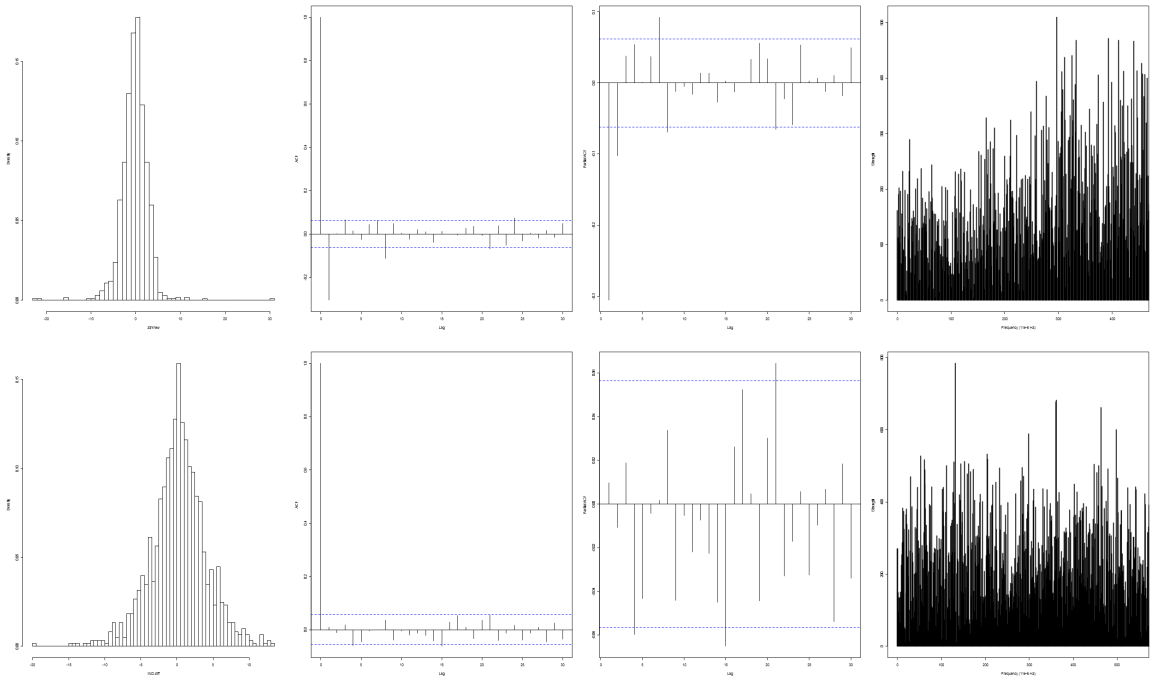

Figure 4.6: Characteristics of simulation (Top) vs IND (Bottom). In order: Probability desnity histogram, ACF, PACF and DFT

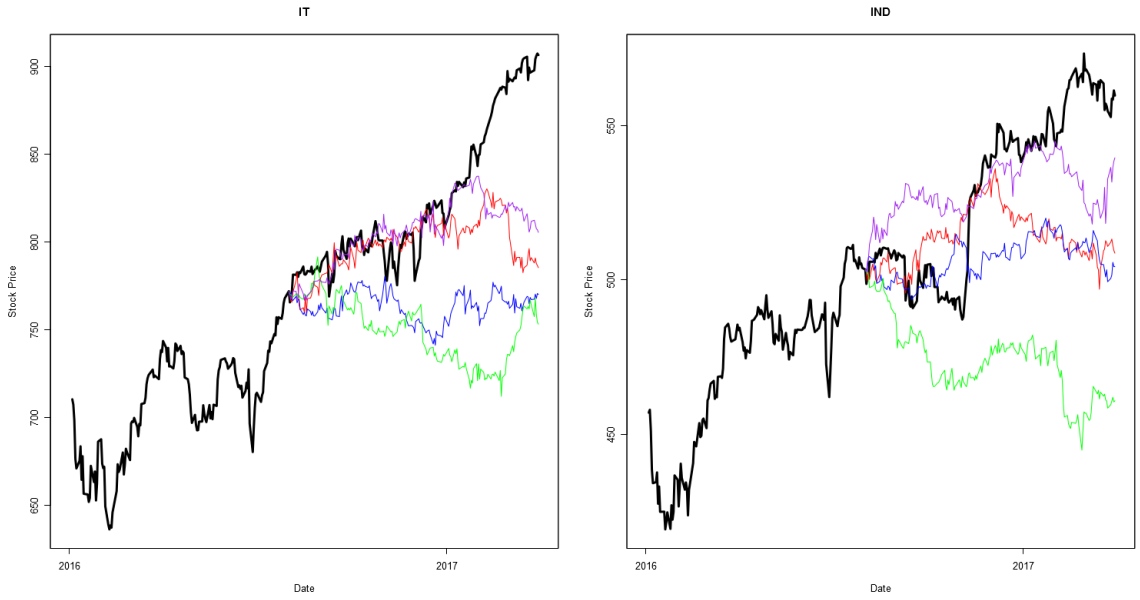

Figure 4.7: Several simulation forecasts of IT (left) and IND (Right).

4.3 Model 3

Next, we modified the model so that previous contributions of X_t and Y_t would have an effect on the outcome of χ_t and γ_t, the model proposed was

$$Y_{t+1} = Y_t + \chi_t \Delta Y_t; \tag{4.5a}$$

$$\chi_t = C_\chi + \alpha_1 \Delta X_t + \alpha_2 \Delta X_{t-1} + ... + \alpha_n \Delta X_{t-n} + \epsilon_t; \tag{4.5b}$$

$$X_{t+1} = X_t + \gamma_t \Delta X_t; \tag{4.5c}$$

$$\gamma_t = C_\gamma + \beta \Delta Y_t + \beta_1 \Delta Y_{t-1} + ... + \beta_m \Delta Y_{t-m} + \omega_t \tag{4.5d}$$

The values multiple regression coefficients α_n and β_m can be found through maximising the likelihood function, for χ_t the likelihood[2] function is

$$L = \prod_{t=1}^{T} N(\chi_t : \Delta X_t, \boldsymbol{\alpha_n}, \sigma^2) = (2\pi\sigma^2)^{-T/2} \exp\left(\frac{-1}{2\sigma^2}(\chi_t - \boldsymbol{\alpha_n}\Delta X_t)^2\right) \tag{4.6}$$

It is generally easier to maximise the log likelihood

$$l = -\frac{T}{2}\ln 2\pi - \frac{T}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(\chi_t - \boldsymbol{\alpha_n}\Delta X_t)^2 \tag{4.7}$$

The choice of m and n was chosen somewhat arbitrarily, we knew that these numbers would be relatively low from the CCF, we tried a combination of different values and chose the numbers based on a t test of $t = \frac{\alpha_n}{\sigma_n}$ and trajectory stability. Results are shown in Table 4.4. We create a simulation and plot the characteristics as before in Figure 4.9 and 4.10. Similar to the previous model we see that the probability distribution is less wide than expected, and several peaks in the ACF and PACF appear when they are not expected. The results of the stationarity and χ^2 are shown in Table 4.5. Again we find that we cannot reject the χ^2 null hypothesis of non-Independence. Running several simulations we plotted potential trajectories as forecasts, we see that one particular trajectory leads to instability.

Although we saw certain agreements, We were unable to produce a model that describes the two stocks better than the previously discussed models; however it should be noted that there is a lot of room for modifications and we have only scratched the surface so far. In the discussion section we go over some of the promising potential for future work regarding the model.

[2]Assuming normality

Table 4.4: Regression Paramters and Residuals

	Parameters	t statistic	$P(t)$
C_χ	-1.26670 ± 0.44768	-2.829	0.00536
α_1	-0.01738 ± 0.09467	-0.184	0.85464
α_2	0.01764 ± 0.08482	0.208	0.83554
α_3	0.02287 ± 0.10013	0.228	0.81969
α_4	0.06514 ± 0.10011	0.651	0.51630
α_5	-0.14596 ± 0.08481	-1.721	0.08750
ϵ_t	ARIMA(0,0,0)	σ^2=0.57	
C_γ	-1.19703 ± 0.60884	-1.966	0.0513
β_1	0.03036 ± 0.07331	0.414	0.6794
β_2	-0.04580 ± 0.06570	-1.697	0.0869
β_3	-0.09700 ± 0.08043	-1.206	0.2299
β_4	-0.09140± 0.08038	-1.837	0.0475
β_5	-0.06095 ± 0.06536	-0.932	0.3527
ω_t	ARIMA(0,0,0)	σ^2=0.432	

Table 4.5: Stationarity results and χ^2 test

	IT Simulation	IND Simulation
KPSS ~ 3	0.211	0.234
$p(K)$	0.33	0.23
ADF, $\sim$5	-4.71	-5.71
$p(A)$	0.01	0.01
χ^2	28382	27823
$p(\chi^2)$	0.2387	0.2502
height		

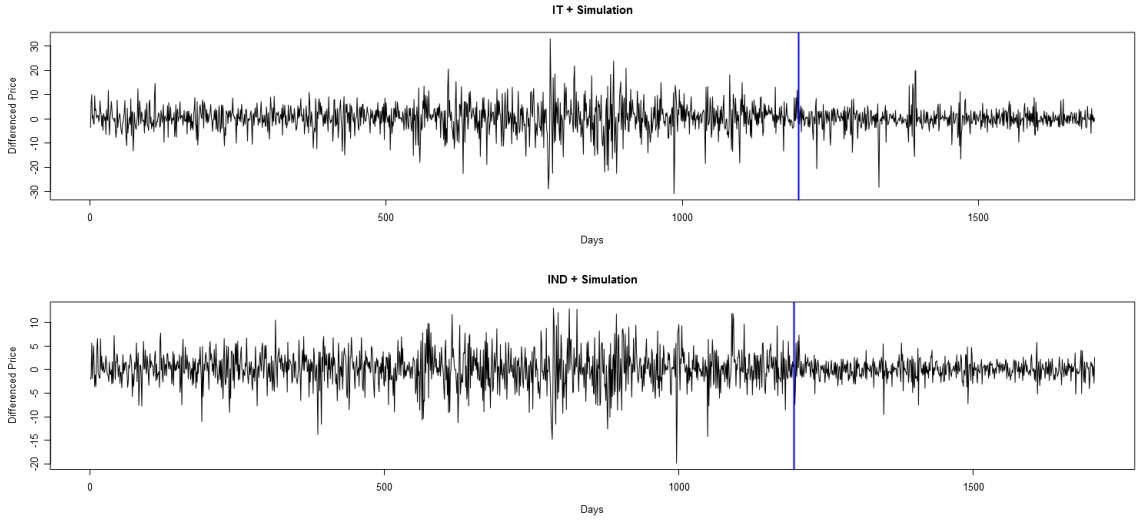

Figure 4.8: Original and simulation of differenced IT (top) and IND(Bottom), the blue vertical line shows where the simulation begins.

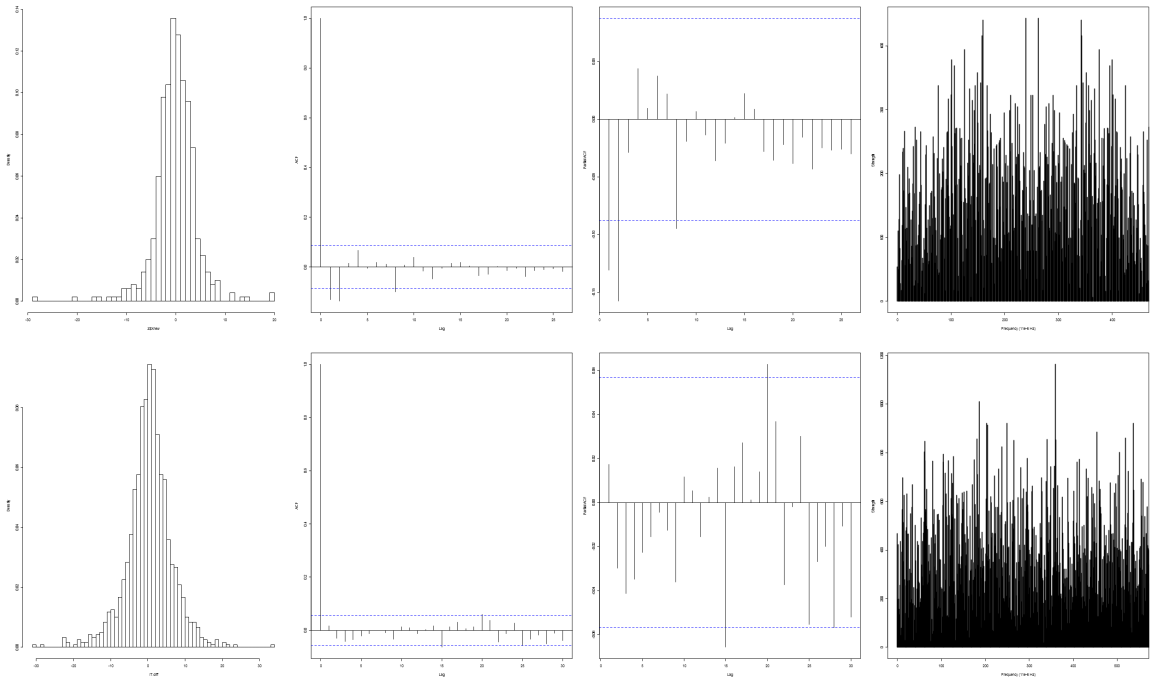

Figure 4.9: Characteristics of simulation (Top) vs IT (Bottom). In order: Probability density histogram, ACF, PACF and DFT

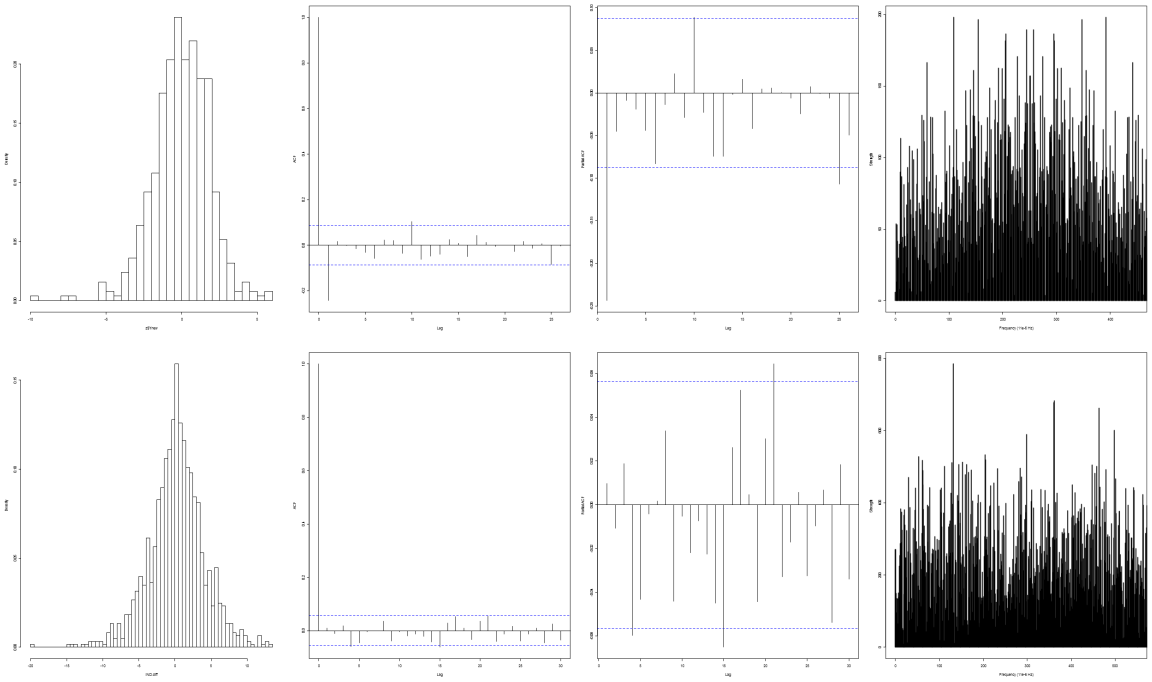

Figure 4.10: Characteristics of simulation (Top) vs IND (Bottom). In order: Probability density histogram, ACF, PACF and DFT

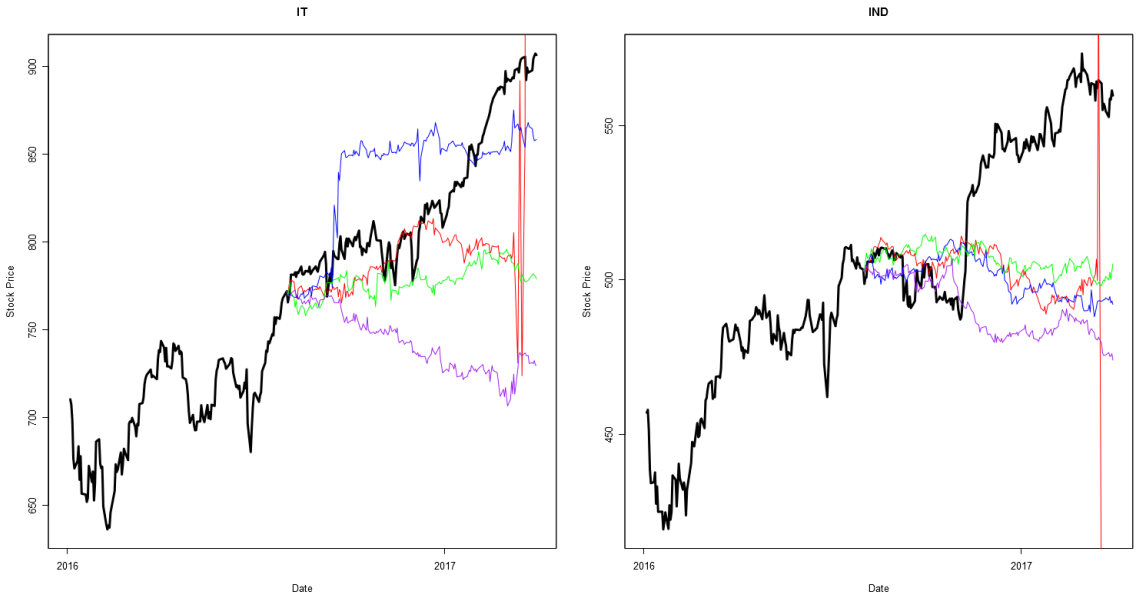

Figure 4.11: Several simulation forecasts of IT (left) and IND (Right).

Chapter 5

The Effect of News & Sentiment Analysis

Our study of financial markets so far has looked at relationships within the closed system of the market; in this chapter we extend the study by looking at how world events cause changes in the market. In this chapter we look at collective behaviour (after an world events) and the effect on the market, followed by a sentiment analysis on news data from a period of 8 months in order to see the relationship between stock price. Our study of the S&P500 continues here as we use the IT sector and its constituents (table 1) as our series'.

5.1 Collective Attention: Analysis of Google Trends

As a preliminary, we needed to prove simply that there is a correlation between the market and world events, Heiberger [39] showed that using the Google Trends[**?**] tool from Google inc. a trading strategy could be created during times of high stock volatility.

Google Trends (GT) is a publicly available tool from Google that shows how often a particular term is searched on any of Google's engines (Web search, Images, News etc.) by assigning a normalised popularity score $\approx \frac{\text{No. of Term searched}}{\text{Total Volume of Searches}}$. Google keeps a record of historical data overtime. We use this to make the assumption that data gathered from Google Trends from the *news* search engine is a measure of collective behaviour of US population and therefore a direct reflection of world events. We used the top 10 constituent tickers in the S&P500 IT sector as our search terms.

The data was gathered from 2016-01-01 to 2016-08-01. The data from GT included weekend data where our stock market series did not; to account for this we removed the weekend data from our analysis and assumed, for this preliminary experiment, that the effect is negligible. Figure 5.1 shows an example of the gathered data, for the search term "AAPL". We looked at the correlation between the trends and both trading volume and stock price. Table 5.1 shows the results of the correlation coefficient. We find that for stock price the Pearson's null hypothesis of 0 correlation cannot be rejected and therefore we cannot say GT is linearly related to stock price. Correlation is significant for trading volume, however, in almost ever instance the null hypothesis is rejected for 95% critical value, except for FB and INTEL which were very close to rejection; showing that there is indeed some linear relationship between trading volume and GT. We plot this relationship in Figure 5.2 for every stock. Our result shows that there is indeed some relationship with an outside system and the 'closed' market system. This result makes sense, as agents get word of events that may effect certain stocks, they use sites such as Google to gather further information and make trade decisions, the more agents that are searching the more number of trades that may occur. What this doesn't tell us is overall nature of trades, whether agents are pushing to sell stocks due to a an event that favours the stock badly, driving the price down, or pushing to buy stocks after news that favours stocks well, increasing the price. To find this out, we need a measure of news sentiment, i.e. a quantity that determines how favourable a world event is.

Table 5.1: Correlation Results, Stock Volume, Stock Price and GT data. N = 144.

	ρ_{price}	$p(\rho_{price})$	ρ_{Vol}	$p(\rho_{Vol})$
AAPL	0.0115	0.85	0.8240	2.2e-16
CSCO	0.0230	0.78	0.5145	3.0e-11
FB	0.0130	0.87	0.1550	0.06185
GOOGL	0.0826	0.318	0.4905	3.3e-10
IBM	0.1330	0.112	0.46256	4.15e-9
INTEL	0.0500	0.52	0.1555	0.055
MSFT	0.0810	0.33	0.6480	2.2e-16
ORCL	0.1230	0.131	0.4212	1.2e-7

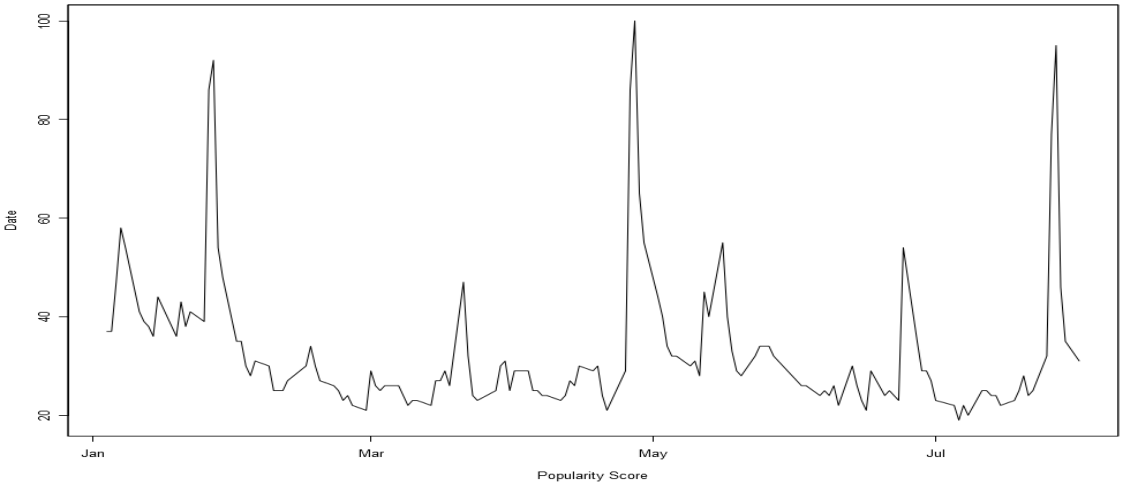

Figure 5.1: Data gathered from Google Trends news for the ticker: AAPL, weekend data has been omitted.

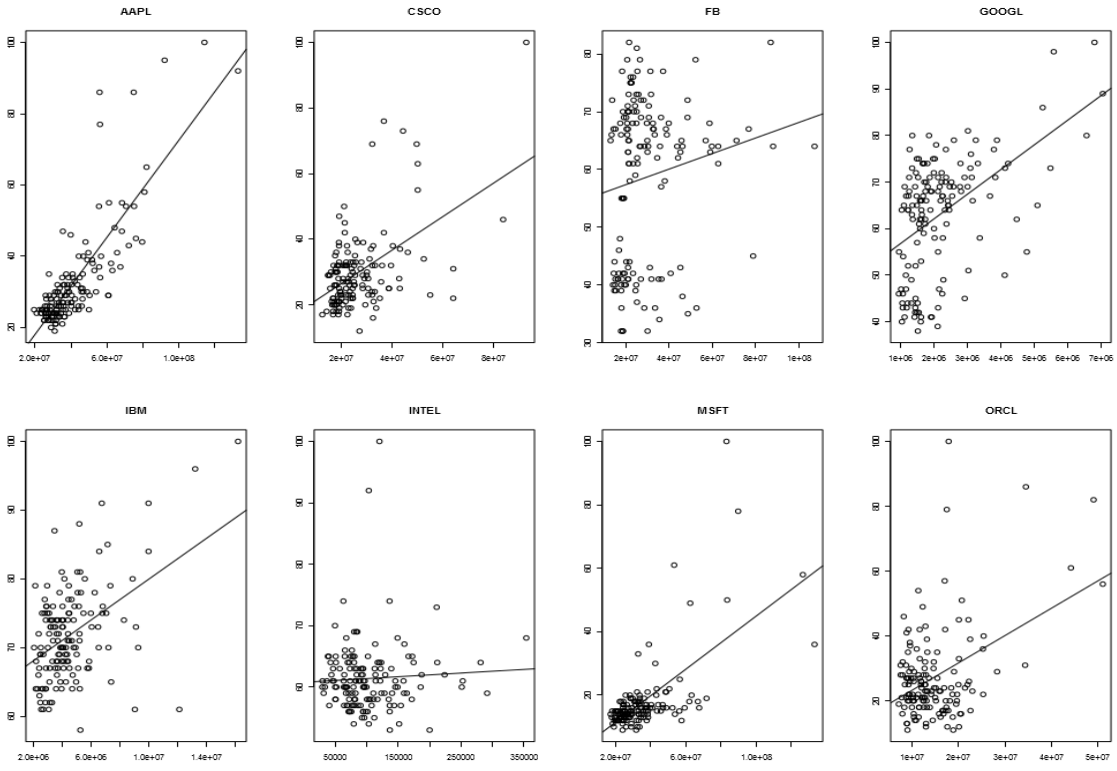

Figure 5.2: Multiple plots of S&P500 IT sector constituent trading volume vs GT score with linear trendline included.

5.2 News Sentiment Analysis

Yue Xu (2009) [40] used an arbitrary measurement system of news articles from the Yahoo finance in order to test the linear correlation between stock markets and the news; it was found that at certain time differences there was a small but significant correlation. The shortcomings of their experiment was that their systematic quantification of news was done arbitrarily and that Yahoo finance news website was not finely tuned for investors, news data was very mixed and full of irrelevant information. We looked at sentiment analysis as a more scientifically sound method of quantification. O'Connor et al (2010). [41] showed that very basic "bag-of-words" sentiment analysis models could be used to show very significant cross-correlation results with opinion poll results. Alanyali et al (2013) [42] showed that using the Financial Times as a news source, by analysing the number of mentions of a stock, showed how the dynamics of stocks and trade decisions change together by studying trading volumes. Here, we take inspiration from all three papers and develop a model sentimental analysis technique. Our proposal is that agents read news articles that give insights to world events which may have consequences to a particular investments, from the overall sentiment of the article they make a decision to continue their investment in the stock, buy or sell. The Financial Times is a source attuned for active investors, with articles and features written by experts in the field who give insights on the implications news articles when required, we hypothesised that the sentiment of these articles has a direct effect on the agents' decisions.

5.2.1 Methodology

The source of news used was directly from the Financial Times (FT.com) website, we used their press cuttings service[43] to look for news articles between the dates 01-01-2016 and 01-08-2016. For each day, we searched for each ticker in the S&P500 IT sector and stored all the relevant news articles (ignoring non-news topics such as 'Home and Lifestyle') as an accumulated corpus. The FT paper is published daily, excluding Sundays, this posed a problem for comparison with stocks as the market is closed on weekends. Our hypothesis is that news may take time to effect the market, so we approximated the Saturday results by counting them as released on Friday.

We used the text mining libraries "tm" and "RWeka" available in R for the analysis. Each daily corpus was preprocessed by removing special characters such as

"-","?","/" and convert all letters into lower case. We separated the corpus in to 1-length vectors for each word, a process called *tokenization*, and analysed the resultant vectors.

We arrive at a sentiment using two different models. The *geometric* polarity is a measure of the ratio of frequency of positive corpus vectors and negative ones

$$g_t = \frac{\text{Count(Positive Vector)}}{\text{Count(Negative Vector)}} = \frac{P(\text{Positive Vector } |t)}{P(\text{Negative Vector } |t)} \tag{5.1}$$

The *arithmetic* polarity

$$a_t = Count(\text{Positive Vector}) - \text{Count(Negative Vector)} \tag{5.2}$$

The arithmetic polarity is very much an unnormalised model, while the geometric is normalised, this allows us to see whether more overall news has an effect on the market series.

We used the subjectivity lexicon developed by OpinionFinder [44], under GNU public licence, to determine the counts in sentiment, we align the vectors with the lexicon of positive and negative words, and count the number of matches. Polarity is determined from a list of 1300 negative and 1600 positive words.

5.2.2 Resultant Series

Figure 5.3 shows the resultant series for both the geometric and arithmetic polarity, stationarity was found through both KPSS and ADF tests, shown in table 5.2. The resultant series shows a very random and fluctuating pattern, similar to a stationary market time series; at certain dates, spikes in polarity exist. These can occur due to a very large/ important event or a series of smaller news articles, comparing both models can allow us to see the distinction. In Figure 5.4 we plotted the time series characteristics. For the arithmetic polarity, we observe that the distribution is asymptotically normal from the histogram and Q-Q plot, with a $\mu = -2.8$ and $\sigma = 20.4$. We see no significant autocorrelation however there is a single significant partial autocorrelation at $k = 12$.The DFT shows a very high amplitude at lower frequencies, telling us there is periodicity in news polarity every few days, to distinguish peaks in other frequencies we would require a filter to the spectra. For the geometric polarity, the distribution is no longer normal, but a right-skewed curve, fat tails are confirmed by the Q-Q plot, there is no significant autocorrelation in the series; the

spectra now shows a very high peak at the lowest frequency - we know this is not due to non-stationarity, so the series is periodic at the lowest frequency.

Table 5.2: Stationarity Tests

	$KPSS$	$p(KPSS)$	ADF	$p(ADF)$
a_t	0.42789	0.065	-3.6344	0.03259
g_t	0.25395	0.13	-3.7296	0.02441

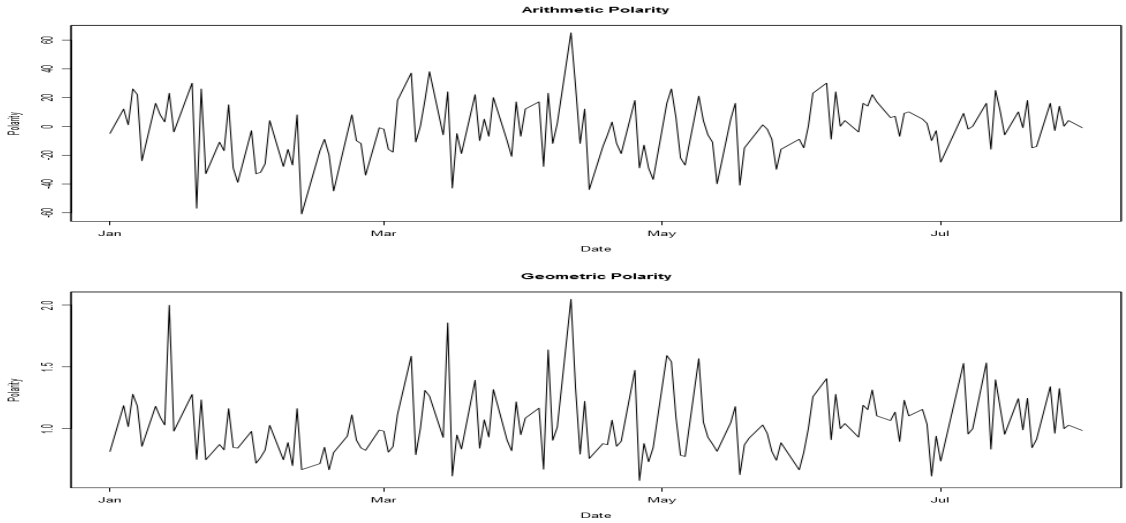

Figure 5.3: Arithmetic (top) and Geometric(bottom) Polarity.

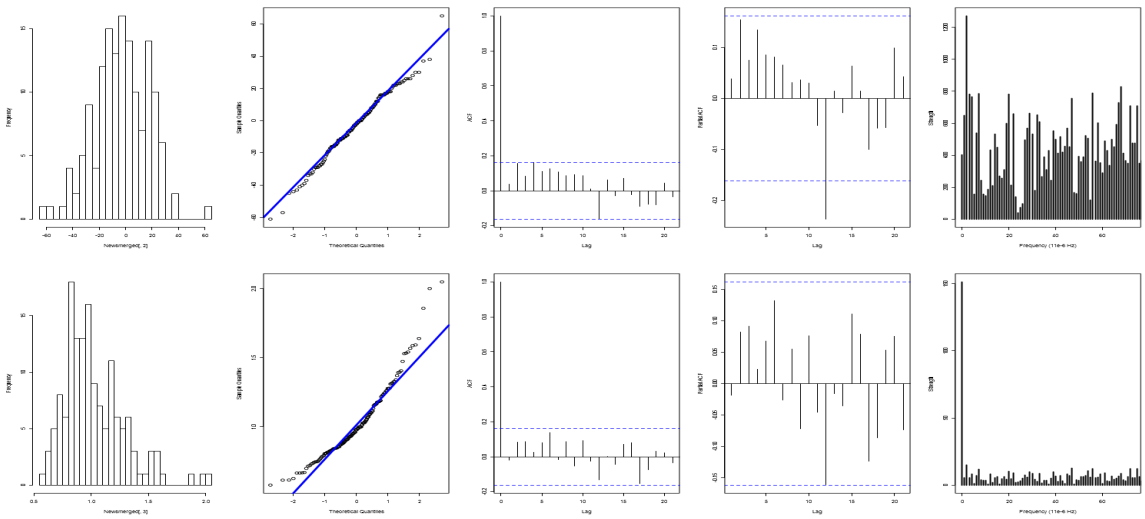

Figure 5.4: Time series characteristics, distribution, Q-Q plot, ACF,PACF and spectra. Arithmetic (top) and Geometric(bottom) Polarity.

5.2.3 Stock Market Correlation

We calculated the correlation coefficient with the IT index at a number of different lags of the sentiment ,k, shown table 5.3. There are number of significant results

Table 5.3: Correlation Results :stock price and sentiment data. N = 144.

k	ρ_a	$p(\rho_a)$	ρ_g	$p(\rho_g)$
1	-0.098	0.239	-0.091	0.273
2	0.030	0.720	0.0679	0.415
3	0.151	0.081	0.0881	0.291
4	-0.152	0.079	-0.072	0.389
5	0.148	0.079	0.183	0.0287
-1	0.003	0.975	-0.116	0.161
-2	-0.061	0.463	-0.041	0.623
-3	-0.132	0.114	-0.0156	0.852
-4	0.103	0.220	0.0450	0.593
-5	-0.020	0.812	-0.071	0.402

at above 80% interval, however a very the null hypothesis is rejected at above 97% confidence for $g_t : k = 5$. We plot these correlations (akin to a CCF) in Figures 5.5 and 5.6.

The regression relation for $g_t : k = 5$ is

$$\Delta IT_t = (-4.927 \pm 0.200) + (5.41 \pm 0.31) g_{t-5} + \epsilon_t \tag{5.3}$$

Where the residuals are found to be $NID(0, 7.9)$, independence is shown by the 0 autocorrelation, Figure 5.7. This result essentially states that news polarity, on average, has the highest contribution into the market after 5 business days (1 week). We do see some other notable correlations, and this hints that the models used here require further development; we discuss improvements to the model in the final chapter.

We simulate the IT time series using equation 5.3 a number of times, shown in Figure 5.7, applying thr χ^2 test with $df = 19596$ we find $\chi^2 = 19932$ which corresponds to a p value of 0.0457, allowing us to reject the null hypothesis of Independence.

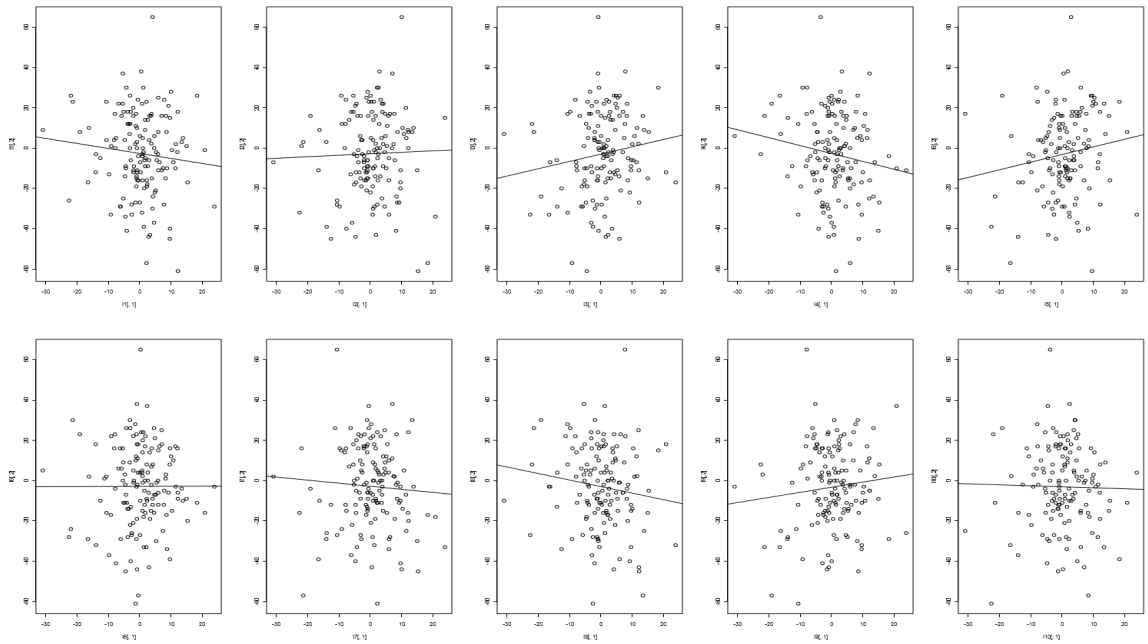

Figure 5.5: Lag regression plots of arithmetic polarity vs IT. Top: $k = 1 : 5$. Bottom: $k = -1 : -5$

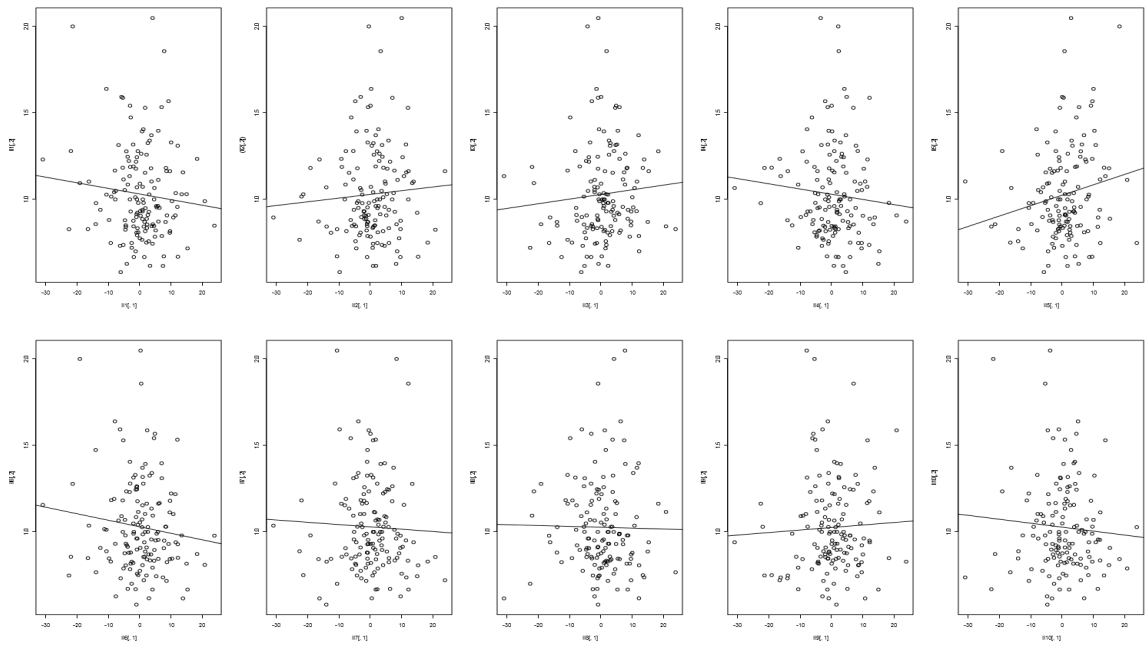

Figure 5.6: Lag regression plots of geometric polarity vs IT. Top: $k = 1 : 5$. Bottom: $k = -1 : -5$. The most significant is $k = 5$ top right.

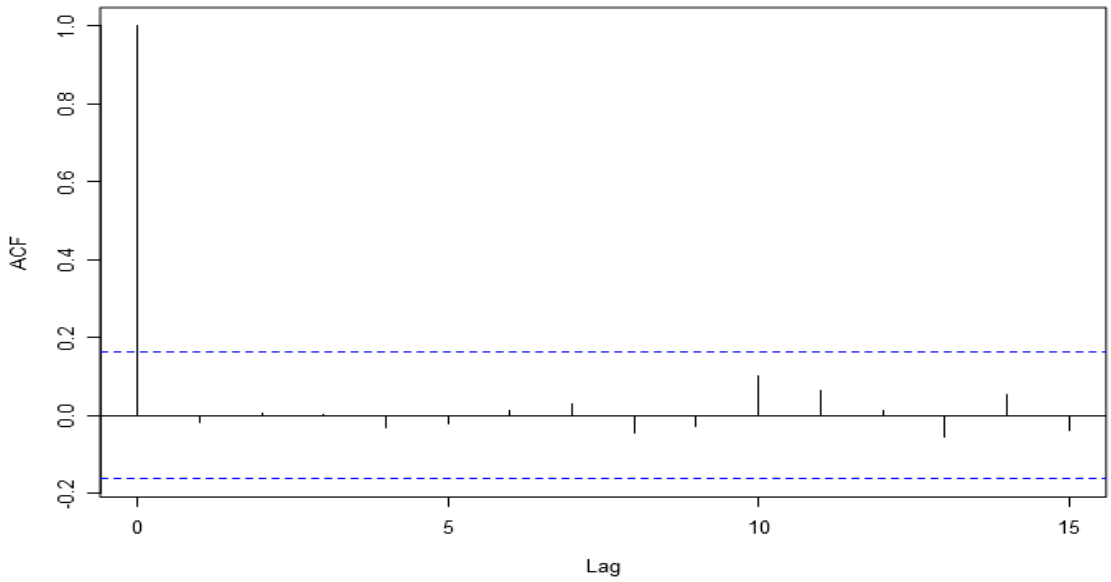

Figure 5.7: Autocorrelation of the residuals of $k = 5$ regression.

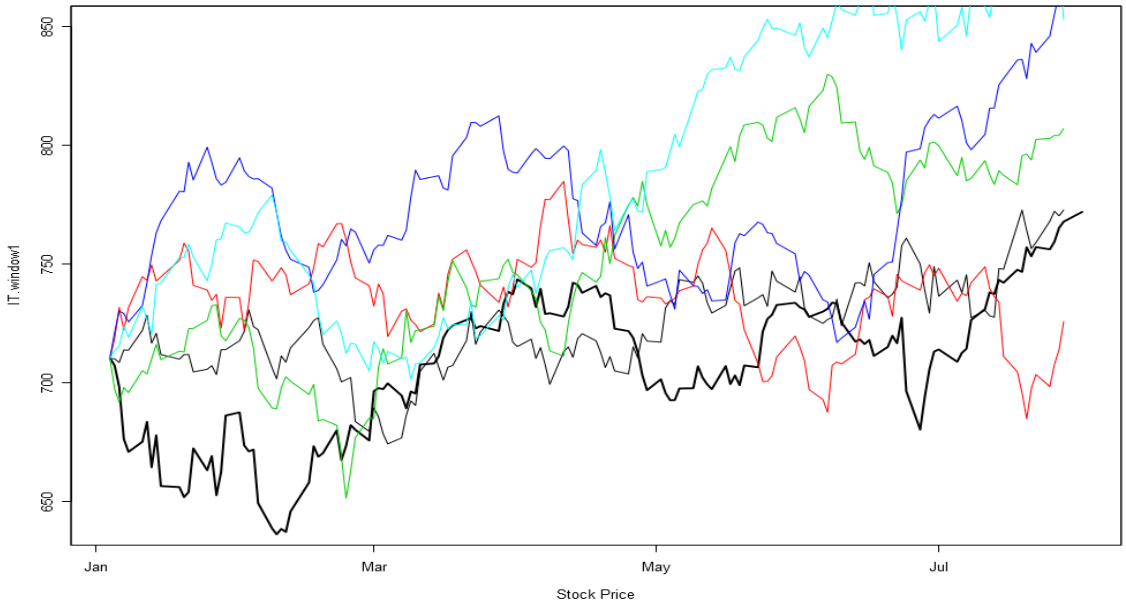

Figure 5.8: Simulations of IT index using geometric polarity as sentiment with $k = 5$.

Chapter 6

Discussion and Conclusion

We set out withe the goals of analysing financial markets through the point of view of a multi-layered complex system with clear aims set from Chapter one. Here we will conclude out findings and discuss how well our aims were met. We began by applying common time series analysis methods, commonly utilised in both statistics and physics. We showed, by use of statistical tests, that differencing a financial time series more sufficiently leads to stationarity over linear detrending. We introduced Fourier transforms and found the frequency spectra of multiple financial time series and simulations throughout the project using the discrete Fourier transform. The Fourier transform was useful as a measure of time series characteristic when comparing with other models and simulations. Similarly, we calculated the autocorrelation and partial autocorrelation function (ACF, PACF) as another characteristic measure. We found, for the S&P500 IT sector, that statistically small autocorrelation exists that decays with time; partial autocorrelation was also found at one month intervals. Autocorrelation was more significant in financial series with lower trading volume such as the cotton or copper commodity, all of these make for evidence against the efficient market hypothesis. We introduced the mathematics of ARIMA models and found that for system with significant autocorrelation such as the copper series, the linear models provided a reasonable explanation of their dynamics. We found that for more complex systems such as the IT index, a random walk was the best fit model. Random walk as a null hypothesis was rejected using the BDS test for non-linearity in both the IT and the IND index, suggesting a non-linear system. We then used the correlation dimension to look for low dimensional chaos within both sectors and found that there was little evidence for chaos using this method. However this is not to say chaos does not exist within the stock market, or even within our studied system; studying daily

data, we were restricted by the number of data points to accurately calculate the correlation dimension at high embedding dimension. Typically, scientists who study chaos work with data with lengths of over 100,000, ours had only 1200 - perhaps the the use of intraday data would lead to different results, this is certainly an avenue to we could explore in the future. Furthermore, other analysis methods may show evidence in chaos, utilising methods such as largest Lyapunov exponent or Kolmogrov entropy are other common methods of measuring chaos, which could be tested for in the future. We used the inferences made in the previous chapter to propose a potential model for the relationship between two markets in the form of a simultaneous recurrence relation. We used the IT and IND data to find parameters of the model in the deterministic regime and were unsuccessfully in defining stable trajectories that did not converge. We found that the application of ARIMA models on the residuals of parameter estimations showed some promise in a stochastic approximation. Our simulated data agreed with a number of characteristics from the original series but the independence between the original and simulations could not be rejected even after modification to include previous lags. There are multiple avenues we could take in the future with regards to the model, one example is the accounting for non-linearity in the variance of the series (also known as heteroskedasticity) by adding a variance term into the model, akin to those termed as GARCH [45].It is also obvious that a set of constraints need to be added to the model to ensure stability, this would be another topic of research in the future. We may also look at intraday data as they have different characteristics, or perhaps the model could be extended to include a larger number of series than 2. Finally, we looked at the effects outside of the system by quantifiying the power of news using a simple, novel sentiment analysis model. We found that a simple linear correlation model describes the relationship between the IT sector and news sources when the news is described using a geometric polarity. It is not known, however, if the model will be successful as a trading strategy, perhaps this is a topic for further investigation. The use of language models as indicators for systems is a very new field, future prospects could look at machine learning as a more accurate method of quantifying news data.

Appendix A

Personal Reflective Statement

I came into fourth year with the impression that I was an especially gifted student and that I could sail through the project with ease. Dr Matthai showed me almost instantly that neither was not the case and for that I cannot thank him enough. Sobering me of my arrogance has been the main factor in allowing me to complete such a challenging project, that I'm immensely proud of.

And this project was indeed challenging, in fact the most challenging thing I have ever done, but with all challenges comes great reward and at this point I feel very rewarded.The first challenge was the decision to use a combination of the programming language R and Python to carry out my analysis as I realised R had potential where Python did not in time series analysis. This meant that I had to self teach a new programming language, thankfully this did take much away from the scope of the project and as a result I am now very well versed in R. Then there was the need to learn the vast plethora of financial economic theory that I had never been exposed to previously, as interesting as these turned out to be, it was rather difficult treading through to learn subjects at a fourth year level. The mathematical methods of time series analysis were also a field I had never been exposed to as it is not a widely taught subject at undergraduate physic. There was also the limitations of the data I was able to find, the availability of financial data is often hidden behind very large premiums and many interesting avenues of analysis, such as intraday data are largely expensive to obtain, luckily I was able to find good quality, well sourced data. But the biggest challenge of all has been the concept of building a mathematical model from scratch. So much of my time was spent going back and forth looking for different ways to express my findings.

However I was also well equipped, thanks to lessons in data analysis and the general mathematics I learnt in fourth year, I believe I made it through with some success! The final difficulty came in how this project was to be presented, having to explain a topic within a different discipline and at the same time at a Masters level is no easy task. I decided that It would be very beneficial to the reader to build up the more basic things in relatively high detail and move on with the more complex more quickly as the reader got the hang of things. The consequence of this was that the dissertation had a very large number of pages as I used figures to explain many of the findings.

Thank you for taking the time to read this dissertation, I hope that it was as enjoyable to read as it was to write.

Arman Tadjrishi

Bibliography

[1] Philipp Hartmann, Florian Heider, Elias Papaioannou, and Marco Lo Duca. The role of financial markets and innovation in productivity and growth in europe. *Social Science Research Network, European Central Bank*, SEPTEMBER 2007.

[2] Amitava Sarkar. Financial markets as complex systems. *West Bengal University of Technology*, 2009.

[3] S&P500. hhttps://www.britannica.com/topic/SandP-500. Full Description of the SP500 Index.

[4] S&P500 IT Sector. https://markets.ft.com/data/indices/tearsheet/constituents?s=SP500-45:IOM. Accessed: Throughout 2016/2017.

[5] S&P500 IND Sector. https://markets.ft.com/data/indices/tearsheet/constituents?s=SP500-20:IOM. Accessed: Throughout 2016/2017.

[6] Butterfield N. *Determinism and indeterminism*. 2005.

[7] Eugene F. Fama. The behavior of stock-market prices. *The Journal of Business*, 38(1):34–105, 1965. ISSN 00219398, 15375374. URL http://www.jstor.org/stable/2350752.

[8] George Box and Gwylim Jenkins. *Time series analysis: Forecasting and Control*. Prentice-Hall, 1994. ISBN 0130607746.

[9] Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637–54, 1973. URL http://EconPapers.repec.org/RePEc:ucp:jpolec:v:81:y:1973:i:3:p:637-54.

[10] Michael Stecher. Converting the black-scholes pde to the heat equation. *Spring 2012 Math 425, Texas AM University*. URL http://www.math.tamu.edu/~stecher/425/Sp12/blackScholesHeatEquation.pdf.

[11] Friedrich Wagner Matthias Raddant. Phase transition in the sp stock market. *Journal of Economic Interaction and Coordination, 11(2), 229-246, 2016.* doi: 10.1007/s11403-015-0160-x.

[12] J. Jurczyk I. Morgenster A. Eckrot. Ising model of financial markets with many assets . *Department of Physics, University of Regensburg, D-93040 Regensburg, Germany*, 17 June 2016. doi: http://doi.org/10.1016/j.physa.2016.06.045.

[13] Tetsuya Takaishi. Multiple Time Series Ising Model for Financial Market Simulations. *Hiroshima University of Economics, Hiroshima 731-0192, JAPAN*, 2015. doi: 10.1088/1742-6596/574/1/012149.

[14] Amitava Sarkar. Financial markets as complex systems . *West Bengal University of Technology*, 2009.

[15] PELN ÖZKAN. Analysis of stochastic and non-stochastic volatility models. SEPTEMBER 2004.

[16] Charlotte Werndl. Are deterministic descriptions and indeterministic descriptions observationally equivalent? *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 40(3): 232 – 242, 2009. ISSN 1355-2198. doi: https://doi.org/10.1016/j.shpsb.2009.06.004. URL `http://www.sciencedirect.com/science/article/pii/S135521980900032X`.

[17] Edward Lorenz. Deterministic nonperiodic flow. 1962.

[18] Benoit Mandelbrot. *Chaos and Fractals: New Frontiers of Science*. Springer, February 1993. ISBN 0387979034. URL `http://www.worldcat.org/isbn/0387979034`.

[19] David Hsieh. Chaos and nonlinear dynamics: Application to financial markets. *Fuqua School of Business, Duke University*, 1990.

[20] Blake LeBaron. Chaos and nonlinear forecastability in economics and finance. *Philosophical Transactions: Physical Sciences and Engineering*, 348(1688):397–404, 1994. ISSN 09628428. URL `http://www.jstor.org/stable/54216`.

[21] Quandl. `https://www.quandl.com/`. Accessed: Throughout 2016/2017.

[22] Terence C. Mills and Raphael Markellos. *The Econometric Modelling of Financial Time Series*. Cambridge University Press, 2008. URL `http://EconPapers.repec.org/RePEc:cup:cbooks:9780521710091`.

[23] Phillips P. Schmidt P. y Shin Y. Kwiatkowski, D. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, 1992.

[24] J. Durbin. Wavelets. *American Scientist*, Vol. 82, No. 3 (MAY-JUNE 1994), pp. 250-255.

[25] Jeff Fessler. Fft algorithms. *University of Michigan*. URL `https://web.eecs.umich.edu/~fessler/course/451/l/pdf/c6.pdf`.

[26] Trading volumes of gold and copper, .

[27] Trading volumes of cotton, .

[28] Andres M. Alonso Carolina Garcıa-Martos. Lectures in time series analysis. *Universidad Carlos III de Madrid Universidad Politecnica de Madrid June – July, 2012*.

[29] Bachioua Lahcene. On pearson families of distributions and its applications, 2013.

[30] Cai R Zhou R. Applications of entropy in finance: A review. *Entropy 15: 4909–4931. 2013*.

[31] Roncalli T (2004) Jouanin JF, Riboulet G. Risk measures for the 21st century. *zego PG, editor, Social Goals and Social Organiza*.

[32] Ph.D William J. Egan. The distribution of sp 500 index returns.

[33] W. A. Broock, J. A. Scheinkman, W. D. Dechert, and B. LeBaron. A test for independence based on the correlation dimension. *Econometric Reviews*, 15(3): 197–235, 1996. doi: 10.1080/07474939608800353. URL `http://dx.doi.org/10.1080/07474939608800353`.

[34] David Hsieh. Implications of Nonlinear Dynamics for Financial Risk Management. *Journal of Financial and Quantitative Analysis, Duke University*, 9, 1993.

[35] P. Grassberger and I. Procaccia. Measuring the strangeness of strange attractors. *Physica D Nonlinear Phenomena*, 9:189–208, October 1983. doi: 10.1016/0167-2789(83)90298-1.

[36] Paul R. Cohen. Getting what you deserve from data. *Department of Computer Science Lederle Graduate Research Center University of Massachusetts, Amherst MA 01003*.

[37] Solutions to model 1 ($x_0 : -1 \rightarrow +1$, . URL `http://docdro.id/6UxVZNz`.

[38] Solutions to model 1 (multiple decimal places), . URL `http://docdro.id/6UxVZNz`.

[39] Raphael H. Heiberger. Collective attention and stock prices: Evidence from google trends data on standard and poor's 100. *PLOS ONE*, 10(8):1–14, 08 2015. doi: 10.1371/journal.pone.0135311. URL `https://doi.org/10.1371/journal.pone.0135311`.

[40] Selene Yue Xu. Stock price forecasting using information from yahoo finance and google trend. *Thesis Submitted to UC Berkley*. URL `https://www.econ.berkeley.edu/sites/default/files/Selene%20Yue%20Xu.pdf`.

[41] Bryan R. Routledge Noah A. Smith Brendan O'Connor, Ramnath Balasubramanyan. From tweets to polls: Linking text sentiment to public opinion time series. URL `http://brenocon.com/oconnor_balasubramanyan_routledge_smith.icwsm2010.tweets_to_polls.pdf`.

[42] Helen Susannah Moat Tobias Preis Merve Alanyali. Quantifying the relationship between financial news and the stock market. *Scientific Reports 3, Article number: 3578 (2013)*. doi: 10.1038/srep03578.

[43] The Financial Times, press cuttings. URL `http://presscuttings.ft.com/`.

[44] Subjectivity lexicon. URL `http://mpqa.cs.pitt.edu/opinionfinder`.

[45] Robert Engle. Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business and Economic Statistics*, 20:339–350, 2002.