LUND UNIVERSITY
Faculty of Science

A model of stock price movements

Johan Gudmundsson

Thesis submitted for the degree of Master of Science
60 ECTS Master Thesis

Supervised by Sven Åberg

Department of Physics
Division of Mathematical Physics
May 2016

Abstract

The goal of the thesis is to model stock prices as a stochastic process which exhibits reversion towards an equilibrium point, where the equilibrium point is set by fundamental data points of the company. The stochastic model is compared to the standard approach of using Geometric Brownian motion to simulate stock prices.

The autocorrelations of a group of stocks are investigated. This has lead to the development of a method of modifying stochastic models of stock movements to include autocorrelation, by introducing an autoregressive term.

A method to achieve an index behaviour for a group of simulated stocks is developed, by the introduction of an index term. This can be added to stochastic models of stock movements.

Acknowledgements

I would like to express my sincere gratitude to Professor Sven Åberg, for his guidance, useful discussions and allowing me to explore different topics.

Contents

Acronyms

,

VRM	Value Reverting Model
GBM	Geometric Brownian motion
OU	Ornstein-Uhlenbeak process
GOU	Geometric Ornstein-Uhlenbeak process
AR	Autoregressive Model
PMOD	Probability of movement in opposite direction
CLT	Central limit theorem
V_t^k	Value of company k at time t
S_t^k	Share price of company k at time te
E_t	Earnings per share at time t
B_t	Book value at time t
g_t	Growth of earnings per share at time t
SEC	U.S. Securities and Exchange Commission
DCF	Discounted Cash Flows
NYSE	New York Stock Exchange
NASDAQ	National Association of Securities Dealers Automated Quotations

1. Introduction

The mathematical formalization of a path consisting of a succession of discrete random steps is known as a random walk. Random walks are used in many different areas such as physics, economics, computer science, chemistry, and biology. A rigorous mathematical framework for random walks has been developed to allow for applications in many different areas. A standard random walk has discrete time steps, the continuous-time analogue to a random walk is called Brownian motion.

There are a lot of different applications for Brownian motion, as well as extensions of the standard Brownian motion to apply it to additional situations. One example of this is geometric Brownian motion, in which the logarithm of a random varying quantity follows Brownian motion with the introduction of a drift term. The drift term represents the rate at which the average of the process changes. One of the places where Geometric Brownian motion is applied is in the modelling of stock price movements, as it has similar characteristics as stock prices. For example, the movements of Geometric Brownian motion are independent of the previous value; it produces only positive values; Geometric Brownian motion creates jumps in the value similar to what is seen in the stock market. The model of geometric Brownian motion simulates the stock movement without incorporating what a stock is.

But a stock is a security that represents a partial ownership in a corporation, where it accounts for a claim on a part of the company's assets and earnings. Shares are bought and sold on stock markets and it is this buying and selling that gives rise to stock prices. The fundamentals of a company, such as earnings and assets are commonly used in the analysis of stocks by investors but is ignored in stochastic models of stock movements. The intrinsic value of a stock refers to the shares claim on the value of the underlying business, taking into account both tangible and intangible factors. So in an ideal world, the price of the stock and its intrinsic value should be the same. But this is not the case as the stock price is constantly changing.

In this thesis, the idea of modelling the stock price as a stochastic process which exhibits properties of reversion towards the intrinsic value is explored, where the intrinsic value is calculated using the companies earnings, assets, etc. This can be thought of as if the stock price is connected to the intrinsic value with a spring and the spring pulls the stock price towards the intrinsic value. This new model is then compared to the method of simulating the price development of a stock using Geometric Brownian motion. The models are then explored further to find the limitations and how they can be corrected.

Section 2 explains the principles of what a stock is and how the stock market works, as well as useful stochastic processes and statistical tools. In section 3 are stock prices simulated using stochastic processes and properties such as autocorrelations and correlations are studied. The results from section 3 are discussed in section 4. In section 5 further

developments are discussed and in section 6 conclusions are given.

2. Background

2.1 Stock

A stock is a type of security that signifies a partial ownership in a corporation and represents a claim on a part of the company's assets and earnings. So if a company has 1,000 outstanding shares and one person owns 100 shares, that person would own and have a claim to 10% of the business's assets and earnings. The company's outstanding shares are the total number available shares in the market of that company.

A market is a place or an environment where the traders meet to exchange assets[4], where a trader refers to everyone who buys or sells an asset. Stocks are bought and sold on stock exchanges, which is a regulated marketplace where shares are traded. For a stock to be traded on an exchange, it needs to be listed on that exchange. Different exchanges have their own regulations and requirements that must be meet in order for the companies to have their shares traded on the exchange. The requirements can include conditions such as minimum annual income, minimum number of shares outstanding and minimum market capitalization.

Markets provide liquidity, which means that the shares can easily be bought and sold. The more buyers and sellers there are the more liquidity there is. The price of a share is quoted with a Bid-Ask spread. The bid (the lowest) is the price at which the stock can be sold, and the ask (the highest) is the price at which the share can be bought. The difference between the ask and bid is known as the spread. The width of the spread will be different for different companies and changes over time. The spread usually reflects the liquidity of the share and it is the price at which a transaction between buyers and sellers occurs that is the quoted stock price.

Stocks are only traded on days called 'trading days' which correspond to weekdays. Thus, there is no trading on weekends or holidays. So there can be a different number of trading days per year, but there are approximately 250 trading days per year. In this paper any references to a year refer to the trading days within the year and the weekends and holidays are ignored.

The price of a stock changes during the day, but is commonly quoted in data sets of historical stock prices with the price it has at the end of the trading day. This price is known as the end of day stock quote and is what is referred to as the price of a stock in this thesis.

2.1.1 Index

To measure the performance of a group of stocks a stock index ($I(t)$) is used, which calculates the collective movements of the group of stocks. The index is calculated at discrete times t_i with a fixed time step $\Delta t = t_{i+1} - t_i$. In this case, Δt will be equal to a trading day as the data that are used are only reported once a day.

The price of a stock at time t of company k is denoted with either S_t^k or $S^k(t)$, both notations will be used interchangeably trough the thesis. The index k=1..K denotes the compamies which are included. The index ($I(t)$) is commonly calculated as

$$I(t) = I_0 \sum_{k=1}^{K} m_k(t) S^k(t), \tag{2.1.1}$$

where I_0 is a constant used to set the price of he index at time $t = 0$, $m_k(t)$ is the number outstanding shares for company k at time t. So $m_k(t)S^k(t)$ is the market capitalization of company k at time t i.e. the market value of the whole company. So this type of index is called a market capitalization weighted index.

In an index weighted by market capitalization, a single company can come to dominate the entire index. This effect can be minimized by using an equally weighted index. It calculates the index ($I(t)$) by setting the price of the stocks at $t = 0$ equal to one,

$$I(t) = I_0 \sum_{k=1}^{K} \frac{1}{S^k(t=0)} S^k(t), \tag{2.1.2}$$

where $S^k(t = 0)$ is the price of company k's stock at time $t = 0$ and I_0 is a constant used to set the price of he index at time $t = 0$.

2.1.2 Fundamental data points

The intrinsic value of a company is the value of the business as a whole with all aspects of the company included, regarding both tangible (earnings, assets, etc.) and intangible (business model, governance, etc.) factors. The intrinsic value is also referred to simply as the value. The value of a company does not have to be equal its share price.

The value of a company is calculated independently of its market value, and it is assumed it can be calculated for any company or business. Given the fact that a share is not just a tradable price of paper but represents a fractional ownership of that business, the share of a company should then be valued proportional to its claim on the intrinsic value.

The fundamentals of a company refer to qualitative (intangible) and quantitative (tangible) information that affects the value of a company. Since intangible factors are unquantifiable, they are not included in standard valuation models. Valuation models which estimate the

intrinsic value rely instead on the fundamental data points that are easily quantifiable and standardized, thus making it possible to compare different companies.

Fundamental data points refer to data points that are connected to the company such as earnings, revenue, assets, liabilities, and growth in earnings. It does not include quantities that relate to the trading patterns of the stock itself, such as volatility of the stock price. The share price and the fundamental data points need to have the same scale in order to compare the results from different companies. This can be done by dividing the fundamental data points such as earnings, equity, etc. with the total number of shares for that company. Thus the fundamental data points will become earnings per share, equity per share etc. The company reports its earnings, equity, etc. quarterly with a special emphasis on the last rapport each year, which summarizes the annual results.

So earnings per share (E_t) is the portion of a company's profit allocated to each share at time t. Thus, it serves as an indicator of the company's profitability and, as such earnings per share is a key driver of the stock price. E_t is calculated by dividing net income earned in a given reporting period (quarterly or annually) by the total number of outstanding shares.

Earnings growth (g_t) is a measure of the growth in a company's earnings per share over a particular period, where t is the time of the last E_t in the period. g_t is calculated by fitting the function

$$E_t = c(1 + g')^t \tag{2.1.3}$$

to E_t during the selected period using the least mean square method. c and g' are constants and $g' = g_t$ where t corresponds to the time of the last E_t in the period used in the calculation. Three or five data points are commonly used in the calculations of g_t. In this thesis g_t is calculated using three data points; this means that the data points E_{t-2}, E_{t-1} and E_t are used in the calculation of g_t.

A company has assets such as cash and inventory, as well as equipment, buildings and real estate that are subject to depreciation according to accounting standards. The company also has liabilities such as loans, accounts payable and mortgages. Assets and liabilities are combined into a measure called equity, where the equity is calculated by subtracting the liabilities from the assets. It is similar to the book value per share (B_t), which is calculated by subtracting the liabilities from the assets and dividing by the number of outstanding shares. The book value per share can thus be considered a measure of the amount of money a shareholder would receive for each share if the company were to liquidate.

Accounting standards make it possible manipulate the earnings (E_t) and book value (B_t) to some degree, both by creating lower and higher values. For example, current assets such as receivables and inventories are usually worth close to the reported value, but plants and equipment may be outmoded or obsolete and thus worth less than carrying value. On the other hand, a company with fully depreciated plant and equipment could have a reported value considerably below the real value. Things like real estate purchased decades ago with a reported value equal to the original cost decades ago may be worth considerably more

today. Thus, a precise determination of the value with an algorithm or function using variables such as E_t, B_t is not possible; rather an estimate of the intrinsic value is the goal.

There are many investment theories which outline a framework how to pick a winning stock. But there are primarily two of them that have proven successful over long time periods. Those are value investing and growth investing: in value investing the goal is to buy stocks trading below their intrinsic value, and profit as the stock price increases. In growth investing is the goal to pick a stock that will grow fast and thus get a higher intrinsic value in the future which will result in a higher stock price in the future.

2.1.3 Valuation model

A common valuation procedure in finance is performed by discounting cash flows(DCF). The DCF analysis is a very adaptable tool, which can be used to estimate the intrinsic value of companies, determining the price of initial public offerings and other financial assets. It estimates the present value (V) of its future cash flows by discounting the future cash flows with a discount rate, so that cash flows such as future earnings are worth more if they will be earned next year compared to if they are earned 10 years into the future. The DCF method is subject to assumption bias and small changes in the underlying assumptions of the analysis will alter the valuation results. The value of future cash flows is given by

$$V = \sum_{t=1}^{N} \frac{F_t}{(1+d)^t} \tag{2.1.4}$$

where V is the present value, F_t is the future cash flow at time t, N is the number of years and d is the discount rate[2]. The choice of discount rate (d) is different in different implementations of DCF models[10], but it is assumed to be a constant.

The future cash flows for companies are estimated by assuming the future earnings per share are growing according to

$$E_t = E_{t-1} \cdot i_t, \tag{2.1.5}$$

where i_t is an interest corresponding to time t and E_t is the earnings per share at time t. The interest rate must go towards zero as $t \to \infty$, as it is not possible for a company to grow its earnings indefinitely. As a first approximation a step function with i_t equal to the growth in earnings per share(g) for the first five years and then drops to zero is commonly used. In more details models an exponential decay of g is used as the interest, which is calculated according to

$$i_t = ge^{-u \cdot t}, \tag{2.1.6}$$

where g is the growth in earnings per share and u is a constant that determines how fast the interest goes towards zero.

As the book value per share (B_t) can be thought of as the liquidation value or as the accumulated value from the past, it should also be included in a valuation model. This gives a formula to estimate the intrinsic value(V) of a company, which is given by

$$V = \sum_{t=t_A}^{N} \frac{E_f \cdot g_f e^{-u \cdot t}}{(1+d)^t} + c \cdot B_f, \tag{2.1.7}$$

where d, u and c are constants that need to be determined for the specific company, and E_f is the earnings per share at time f, B_f is the book value at time f and g_f is the growth rate of earnings per share at time f. E_f, B_f and g_f correspond to fundamental data points that are reported annually at time f.

The valuation model (eq. (2.1.7)) can be used when the data points E_f, B_f and g_f are known. Since the data set used only contains annual information on the companies and it is only possible to calculate the value in these points, thus the value can only be calculated once per year. So in order to compare the value(V) with stock prices which are quoted daily, the data points in between were interpolated using a straight line given by

$$V(t) = \frac{V(f_{i+1}) - V(f_i)}{T} \cdot t + V(f_i). \tag{2.1.8}$$

Where $V(f_i)$ is the calculated value using the annual data points from f and T is the number of trading days between f_i and f_{i+1}.

It is possible to obtain the constants d, u and c for the eq. (2.1.7) by assuming that the share price ($S(t)$) oscillates around $V(t)$ during the analysed period. Thus it is possible to fit $V(t)$ to $S(t)$ using a least mean square procedure over the desired period, and this gives the parameters d, u and c. By assuming that the parameters are the same for other time periods it is possible to use the calculated parameters to calculate the $V(t)$ for other data points outside the given period.

2.1.4 Data set

The fundamental data points were sourced from the U.S. Securities and Exchange Commission(SEC) directly in combination with data derived from SEC sourced data. It consisted of ten years (2002-2012) of annual fundamental indicators and financial ratios for active and inactive US companies.

The historical stock price data obtained for stocks from the New York Stock Exchange(NYSE) and The National Association of Securities Dealers Automated Quotations(NASDAQ), which is the largest and second largest stock exchanges in the world ranked by market capitalization[7]. The stock price data were validated against prices published on Yahoo Finance.

2.2 Stochastic processes

In its most basic form the random walk model is defined as a process where the current value is composed of the past value plus a randomly drawn number (ϵ_t). ϵ_t is drawn from a distribution with zero mean and variance one. A one dimensional random walk can be expressed as the quantity X, which is composed of N random steps(ΔX_n)

$$X = \sum_{n=1}^{N} \Delta X_n. \tag{2.2.9}$$

ΔX_n is drawn from a distribution $q(\Delta X_n)$ [4], which is normalized and symmetric

$$\int_{-\infty}^{+\infty} q(\Delta X_n) d\Delta X_n = 1, \qquad q(-\Delta X_n) = q(\Delta X_n). \tag{2.2.10}$$

Random walk can be generalized as a Wiener process, given by

$$dX_t = \sigma\varepsilon\sqrt{dt}, \tag{2.2.11}$$

σ sets the scale and has the dimension of X_t. ε is a dimensionless random number distributed according to the standard normal distribution ($N(0,1)$).

A commonly used distribution when generating random numbers is a Gaussian, which is given by the formula [8]

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-(x-\mu)^2}{(2\sigma^2)}\right] \tag{2.2.12}$$

and is commonly denoted $N(\mu, \sigma)$. The standard normal distribution is denoted $N(0,1)$ and is obtained when $\mu = 0$ and $\sigma = 1$, which gives the formula

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-x^2}{2}\right]. \tag{2.2.13}$$

For the case where $\sigma = 1$ the Wiener processes is denoted dW_t, making it possible to express a Wiener process as

$$dX_t = \sigma dW_t \tag{2.2.14}$$

where σ is a constant and σ^2 is the volatility. This can be generalized by introducing a drift term, which adds a general trend to the process X_t. This gives the expression

$$dX_t = \mu dt + \sigma dW_t \tag{2.2.15}$$

where μ is a constant, this is known as Brownian motion with drift.

2.2.1 Geometric Brownian motion

If a stochastic process X_t satisfies the flowing stochastic differential equation

$$dX_t = \mu X_t dt + \sigma X_t dW_t, \tag{2.2.16}$$

it is said to follow Geometric Brownian Motion(GBM), where μ and σ are constant and dW_t is an increment of a Wiener process or an increment of Brownian motion. The term $\mu X_t dt$ corresponds to the drift and represents the "trend", the term $\sigma X_t dW_t$ represents the random noise and corresponds to a random walk. So σ gives the order of the random movement and is the variance per unit time, while μ controls the drift and can thus be considered to be the expected return per unit time. GBM is also often written as

$$\frac{dX_t}{X_t} = \mu dt + \sigma dW_t, \tag{2.2.17}$$

note that dX_t in all the stochastic differential equations is not an exact differential.

It is possible to estimate the parameters μ and σ by using the Euler-Maruyama discretization on eq. (2.2.16), which results in the expression

$$X_t = X_{t-1} + \mu X_{t-1}\Delta t + \sigma X_{t-1}\varepsilon_t\sqrt{dt}. \tag{2.2.18}$$

It can be rewritten as

$$\frac{X_t - X_{t-1}}{X_{t-1}} = \mu\Delta t + \sigma\varepsilon_t\sqrt{dt}, \tag{2.2.19}$$

then using linear regression the parameters μ and σ can be estimated. It is also possible to use eq. (2.2.18) to simulate GBM.

Applying GBM to the stock movements gives the expression

$$dS_t = \mu S_t dt + \sigma S_t dW_t, \tag{2.2.20}$$

where μ will correspond to the drift in stock prices commonly observed during long time periods. σ^2 corresponds to the volatility of the stock price S_t.

The expectation value of S_t at time t depends only on the initial price at $t = 0$, the drift and volatility parameters[9], and is given by

$$\langle S(t)\rangle = S_0 \exp\left[t(\mu + \sigma^2/2)\right]. \tag{2.2.21}$$

2.2.2 Ornstein–Uhlenbeck process

The Ornstein–Uhlenbeck process(OU) is a stochastic process that can be used as an alternative to Brownian motion when a tendency of reversion towards an equilibrium point is required. The process is stationary[3] and given by

$$dX_t = \alpha(\mu - X_t)dt + \sigma dW_t, \tag{2.2.22}$$

where X_t is a stochastic variable, α, μ and σ are constants. α is larger than zero and determines the rate at which the process reverts towards μ, σ gives the amplitude of the stochastic movements and dW_t is an increment of a Wiener process.

From eq. (2.2.22) it can be inferred that X_t will revert to the constant level μ when $\alpha > 0$. If $X_t > \mu$, the coefficient of the drift term $\alpha(\mu - X_t)dt$ will be negative. So X_t will tend to move downwards, with the reverse happening if $X < \mu$. When $\alpha = 0$ eq. (2.2.22) becomes Brownian motion with no drift.

2.2.3 Geometric Ornstein-Uhlenbeck process

Their is also a counterpart to GBM that exhibits reversion towards an equilibrium point. It is given by

$$dX_t = -\alpha(X_t - \mu)dt + \sigma X_t dW_t, \tag{2.2.23}$$

and is known as the Geometric Ornstein-Uhlenbeck process(GOU). For $\alpha = 0$, the process is equivalent to GBM with the drift parameter equal to zero. Eq. (2.2.23) can be approximated in discrete time using the Euler-Maruyama discretization, which gives the expression

$$X_t = X_{t-1} - \alpha(X_{t-1} - \mu)\Delta t + \sigma X_{t-1}\sqrt{\Delta t}\varepsilon_t \tag{2.2.24}$$

where $\varepsilon_t \sim N(0, 1)$.

The estimation of the parameters α and σ for GOU can be done using a linear regression[5] by writing eq. (2.2.24) as

$$\frac{X_t - X_{t-1}}{X_{t-1}} = -\alpha\Delta t + \frac{1}{X_{t-1}}\alpha\mu\Delta t + \sigma\sqrt{\Delta t}\varepsilon_t. \tag{2.2.25}$$

Setting $R_t = \frac{X_t - X_{t-1}}{X_{t-1}}$ gives

$$R_t = c(1) + c(2)\frac{1}{X_{t-1}} + e_t, \tag{2.2.26}$$

where $c(1) = -\alpha\Delta t$, $c(2) = \alpha\Delta t\mu$ and $e_t = \sigma\sqrt{\Delta t}\varepsilon_t$.

Generalizing eq. (2.2.23) so that the equilibrium point is function of time ($F(t)$) instead of a constant value μ, gives the expression

$$dX_t - dF_t = -\alpha(X_t - F_t)dt + \sigma X_t dW_t, \tag{2.2.27}$$

where α, σ are constants and dW_t is an increment of a Weiner process. Similarly, this can be approximated in discrete time as

$$X_t - X_{t-1} - (F_t - F_{t-1}) = -\alpha(X_{t-1} - F_{t-1})\Delta t + \sigma X_{t-1}\varepsilon_t\sqrt{\Delta t} \tag{2.2.28}$$

where $\varepsilon \sim N(0,1)$ and α and σ are constant.

In the same way as for GOU, α and σ be estimated by rearranging eq. (2.2.28) to

$$\frac{X_t - X_{t-1} - (F_t - F_{t-1})}{X_{t-1}} = -\alpha\left(1 - \frac{F_{x-1}}{X_{t-1}}\right)\Delta t + \sigma\varepsilon_t\sqrt{\Delta t}. \tag{2.2.29}$$

Setting $R_t = \frac{X_t - X_{t-1}}{X_{t-1}}$ gives

$$R_t = c\left(1 - \frac{1}{X_{t-1}}\right) + e_t, \tag{2.2.30}$$

where $c = -\alpha\Delta t$ and $e_t = \sigma\sqrt{\Delta t}\varepsilon_t$, which can be solved using linear regression.

2.3 Correlations

Correlation is a statistical measure of how two time series move in relation to each other. The product-moment correlation coefficient $\rho_{X,Y}$ is given by

$$\rho_{X,Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X\sigma_Y}, \tag{2.3.31}$$

where E is he expected value operator, σ is the variance and μ is the mean of the time series. The correlation coefficient ranges from $+1$ to -1 with three special values,

$$\rho_{ij} = \begin{cases} +1 & \text{completely correlated,} \\ 0 & \text{completely uncorrelated,} \\ -1 & \text{completely anticorrelated.} \end{cases}$$

The closer $\rho_{X,Y}$ is to $+1$ or -1, the more closely the two time series are related. If it is close to zero, it indicates that their are no relationship between the two time series.

For a sample consisting of n data points, $x_1, ..., x_n$ and $y_1, ..., y_n$ the correlation coefficient is calculated using

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.3.32}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

The stock price of different companies can be correlated, uncorrelated and anticorrelated. For example, companies that are in the same industry are more likely to be correlated as

they are subject to similar economic influences compared to two companies that are in different industries and operate on different continents.

The synchronous time evolution of two stocks can be studied using the correlation coefficient ρ_{ij} of the daily changes in return for the two stocks i and j. If the stock price for company k at time t is given by $S^k(t)$, then the return for company k is given by

$$r^k(t) = \frac{S^k(t) - S^k(t - \Delta t)}{S^k(t - \Delta t)}. \tag{2.3.33}$$

$r^k(t)$ makes it possible calculate the correlation between two companies without being affected by differences in the size of $S^k(t)$ by looking at the relative amplitude of the movement for the stocks. The logarithmic differences in price

$$G^k(t) = \ln S^k(t) - \ln S^k(t - \Delta t), \tag{2.3.34}$$

are also commonly used instead of r^k when looking at high-frequency data, as it is possible to approximate

$$\ln S^k(t) - \ln S^k(t - \Delta t) = \ln \left[1 + \frac{S^k(t) - S^k(t - \Delta t)}{S^k(t - \Delta t)}\right] \approx \frac{S^k(t) - S^k(t - \Delta t)}{S^k(t - \Delta t)} = r^k(t) \tag{2.3.35}$$

when Δt is small and $|S^k(t) - S^k(t - \Delta t)| << S^k(t)$. But using $G_k(t)$ is not always suited for data sets consisting of daily data points but rather intra day data.

This makes it possible to calculate the correlation coefficient for the stocks i and j according to

$$\rho_{i,j} = \frac{\left\langle r_t^i r_t^j \right\rangle - \left\langle r_t^i \right\rangle \left\langle r_t^j \right\rangle}{\sqrt{\left\langle r_t^i \cdot r_t^i - \left\langle r_t^i \right\rangle^2 \right\rangle \left\langle r_t^j \cdot r_t^j - \left\langle r_t^j \right\rangle^2 \right\rangle}}. \tag{2.3.36}$$

It is often of interest to study how set of K stocks are correlated by calculating all the possible correlation coefficients for the time series of the companies, $S^k(t)$, $k = 1, ..., K$ where K is the number of stocks. The correlation coefficients are commonly arranged in a correlation matrix C, which is a $K \times K$ square matrix consisting of the elements $C_{i,j} = \rho_{i,j}$ and is real and symmetric[4].

Thus, the elements of C fulfil $C_{i,j} = C_{j,i}$ so the diagonal elements of the matrix will be equal to one, as they represents the correlation of a stock with itself. There will be $(K \times K)/2$ different ρ_{ij} for the set of K stocks but as the diagonal elements that are equal to one, and not of interest. So there will be $(K \times (K - 1))/2$ different ρ_{ij} that are of interest.

So the correlation matrix is calculated using K time series of length T, and if T is not very large compared to K the correlations coefficients will be noisy. Thus the correlation matrix is to a large extent random and random matrix theory can be applied. Using random matrix theory it is possible to predict how the eigenvalues will be distributed.

The eigenvalues of the correlation matrix are obtained by solving the equation

$$det(C - \lambda I) = 0, \tag{2.3.37}$$

where the eigenvalues λ_j, $j = 0, ..., N-1$, are the ones that solve the equation. In the limit $N \longrightarrow \infty$, $K \longrightarrow \infty$ and $Q = T/K \geqslant 1$ eigenvalues of a random matrix C, will be distributed according to

$$\rho_C(\lambda) = \frac{Q}{2\pi\sigma^2}\frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda}, \tag{2.3.38}$$

$$\lambda_-^+ = \sigma^2(1 + 1/Q \pm 2\sqrt{1/Q}), \tag{2.3.39}$$

with $\lambda \in [\lambda_-, \lambda_+]$, and where σ^2 is equal to the variance of the elements in the timeseries used to calculate C [6]. So with the condition $Q > 1$ in eq. (2.3.39) λ will be larger than zero.

Note that this distribution is valid in limit $K \longrightarrow \infty$. So when looking at data from stocks the edges of the distribution become blurred. If $Q < 1$, the distribution will be even more blurred as a lot of the eigenvalues will be close or equal to zero. When looking at a correlation matrix then the eigenvalues corresponding to noise will be distributed according to eq. (2.3.38). A first approximation of the eigenvalues corresponding to the noise will be the bulk of the eigenvalues. Thus it is possible to distinguish information from noise by looking at the eigenvalues outside of the bulk of eigenvalues.

However, if $Q < 1$ then a fraction of the eigenvalues $1 - Q$ will be zero i.e. $(1 - Q)K$ eigenvalues will be equal to zero, the remaining eigenvalues will follow $\rho_C(\lambda)$. In the case were there exists an eigenvalue λ_0 which is larger than the bulk such as $\lambda_0 >> \lambda_n$, $n \neq 0$ then $\sigma^2 = \frac{\lambda_0}{K}$. This can be used to clean up the distribution of eigenvalues and find which eigenvalues corresponds to the bulk.

Let λ_+ be the largest eigenvalue of the bulk and the eigenvalues λ_i be the eigenvalues outside of the bulk. Thus $\lambda_i > \lambda_+$ are the eigenvalues of interest. The largest eigenvalue (λ_0) corresponds to the correlation of the market itself and the other eigenvalues outside of the bulk represent the different sectors, and $\lambda < \lambda_+$ is noise. The stocks that make up the components of the eigenvector for λ_i are the stocks that make up that sector.

2.3.1 Autocorrelation

The correlation between a time series X_t and a lagged version of the same time series $X_{t+\tau}$ over successive time intervals is given by the autocorrelation. The autocorrelation is calculated in the same way as a correlation coefficient using eq. (2.3.31), but with the same mean and variance. Giving the expression

$$R(\tau) = \frac{E[(X_t - \mu)(X_{t+\tau} - \mu)]}{\sigma^2}, \tag{2.3.40}$$

where μ is the mean and σ is the variance of the time series X_t. In the same way as for correlation coefficients, the autocorrelation can range from +1 to -1, where an autocorrelation of +1 represents complete correlation, -1 represents complete anticorrelation and 0 represents no correlations between the time series and a lagged version of itself.

In the same way as the correlation between stocks are calculated using the return, so should the return be used in the calculation of autocorrelation. Giving the expression

$$R(\tau) = \frac{\left\langle r_t^k \cdot r_{t+\tau}^k \right\rangle - \left\langle r_t^k \right\rangle \left\langle r_{t+\tau}^k \right\rangle}{\sqrt{\left\langle r_t^k \cdot r_t^k - \left\langle r_t^k \right\rangle^2 \right\rangle \left\langle r_{t+\tau}^k \cdot r_{t+\tau}^k - \left\langle r_{t+\tau}^k \right\rangle^2 \right\rangle}}. \tag{2.3.41}$$

Note that when using daily data of stock prices i.e. one data point per day, τ will be an integer.

2.4 Autoregressive Models

An autoregressive model(AR) is a random process that depends linearly on its previous values with a stochastic term. An autoregressive model of order p can be written as

$$X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} + dW_t, \tag{2.4.42}$$

where c is a constant and dW_t is an increment of a Wiener process. The parameters $\phi_1, ..., \phi_p$ are used to control the time series pattern. This type of model is referred to as an AR(p) model.

AR models can be used to create autocorrelation in the simulated time series X_t. The autocorrelation for each time step is controlled by the choice of parameters $\phi_1, ..., \phi_p$. The order of the process determines for how high value of τ autocorrelations will exist.

When simulating a time series with the same autocorrelations as real data, the coefficients $\phi_1, ..., \phi_p$ and the order p needs to be selected with care. The coefficients $\phi_1, ..., \phi_p$ can be estimated using the standard least mean square procedure or using Yule–Walker equations[11]. The order p needs to be selected in such a way that it agrees with the actual data.

2.5 Central limit theorem

The central limit theorem (CLT) states that if X_n is the sum of n independent random variables, then the distribution function of X_n will be a Gaussian when n is large.

The CLT thus shows what will happen with a sum of a large number of independent random variables, where each variable contributes with small amount to the total. It is also closely related to the law of large numbers, which states that the mean of the sample converges to the distribution mean as the sample size increases.

3. Results

3.1 Value reverting model

The basic OU process given by eq. (2.2.22) are well suited for applications in finance where reversion towards an equilibrium point is desired. Here μ represents an equilibrium point supported by a fundamental property, σ describes the volatility and α is the rate at which rate the variable reverts to the equilibrium point. Since stocks represent a fractional ownership of a company, it would be possible to say that there exist a value that is possible to calculate from the company fundamental data points such as earnings, equity, etc.

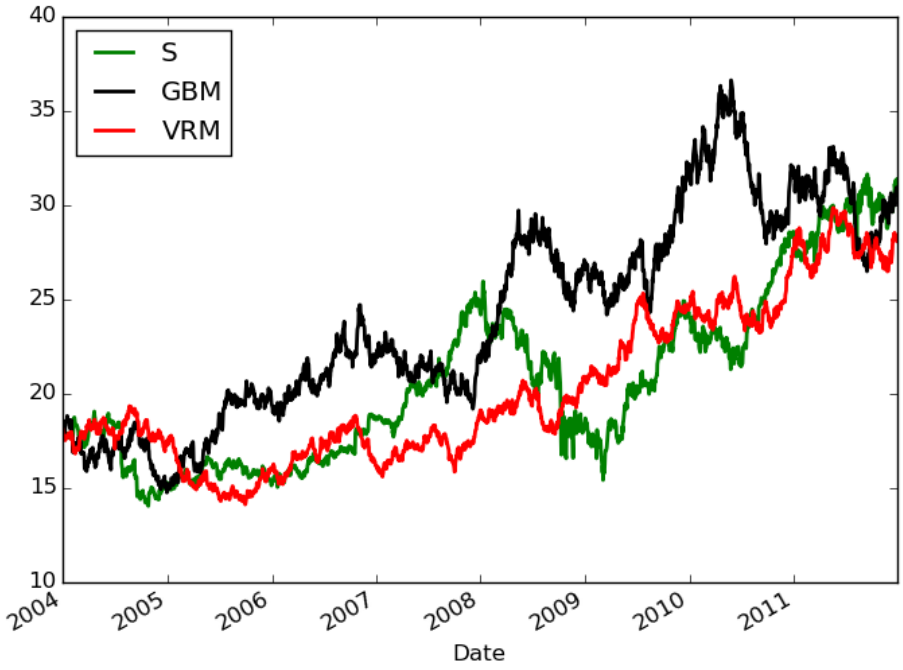


Figure 3.1: The stock price for 'The Coca-Cola Company, (NYSE: KO)' simulated using the value reverting model(VRM) is plotted in red and the stock price simulated using Geometric Brownian motion(GBM) is plotted in black. The realized stock price(S) is plotted in green. The parameters used in the simulations were calculated using the data from the period 2002-2007.

But since a company consists of multiple parts, and there are big differences between different companies it is not possible to perform an exact calculation of value using a simple

algorithm. Instead eq. (2.1.7) was used to estimate the value (V_t), based on fundamental data points. The fundamentals of a company change with time as the company changes, which means that the value of the company will also change with time.

So the basic OU process will not function as desired, given the fact that μ needs to be constant. Instead the modified geometric Ornstein-Uhlenbeck process eq. (2.2.27) can be used, where the share price (S_t) exhibits reversion towards the value (V_t). This gives the expression

$$dS_t - dV_t = -\alpha(S_t - V_t)dt + \sigma S_t dW_t, \tag{3.1.1}$$

This can be discretized as

$$S_t - S_{t-1} - (V_t - V_{t-1}) = -\alpha(S_{t-1} - V_{t-1})\Delta t + \sigma S_{t-1}\varepsilon_t\sqrt{\Delta t}, \tag{3.1.2}$$

where S_t is the share price at time t, V_t is the estimated value per share at time t, α is the rate at which the stochastic process exhibits reversion towards V_t, σ is the volatility and dW_t is an increment of a Wiener process. This will be referred to as value reverting model(VRM) in the rest of this thesis.

By using both GBM and VRM to simulating the stock price path for 'The Coca-Cola Company, (NYSE: KO),' it is possible to compare the differences between modelling stock price movements using VRM compared to the standard approach of GBM.

The parameters for d, u and c for the valuation model (eq. (2.1.7)) were estimated with data from the period 2002-2007 using the least mean squares method. Then V_t were calculated using eq. (2.1.7) with the fundamental data points from the period 2002-2012. Note that three consecutive fundamental data points for the company are needed to calculate the growth rate (g) which means that it was possible to calculate V_t for the period 2004-2012.

The parameters used for VRM and GBM were estimated using data from the period 2002-2007 to see how well suited the models are to both simulate the stock during the period used to fit parameters and to extrapolate future stock movements beyond the period used to estimate the parameters. So the parameters α, σ_{VRM} for VRM (eq. (3.1.1)) and μ and σ_{GBM} for GBM (eq.(2.2.20)) were estimated using the least mean squares method.

The fitting for 'The Coca-Cola Company, (NYSE: KO)' resulted in the parameters $r = 9.64 \cdot 10^{-2}$, $u = 3.25$, $c = 8.05 \cdot 10^{-1}$ for the valuation model. The parameters for VRM were $\alpha = 7.93 \cdot 10^{-3}$, $\sigma = 8.55 \cdot 10^{-3}$ and the parameters for GBM were $\mu = 7.63 \cdot 10^{-5}$ $\sigma = 8.55 \cdot 10^{-3}$. Then these parameters were used to simulate the stock path during the period 2004-2012 using the discretized VRM and GBM, given by eq. (3.1.2) and eq. (2.2.18). This means that the simulation during the period 2004-2007 is simulated with well fitted parameters and the period 2008-2012 acts as an extrapolation.

In fig. (3.1) the share price movement is simulated using VRM and GBM as well as the real stock path (S_t). But only looking at a single simulation of VRM and GBM does not show how the simulation will look when repeated. By doing the same simulation with the

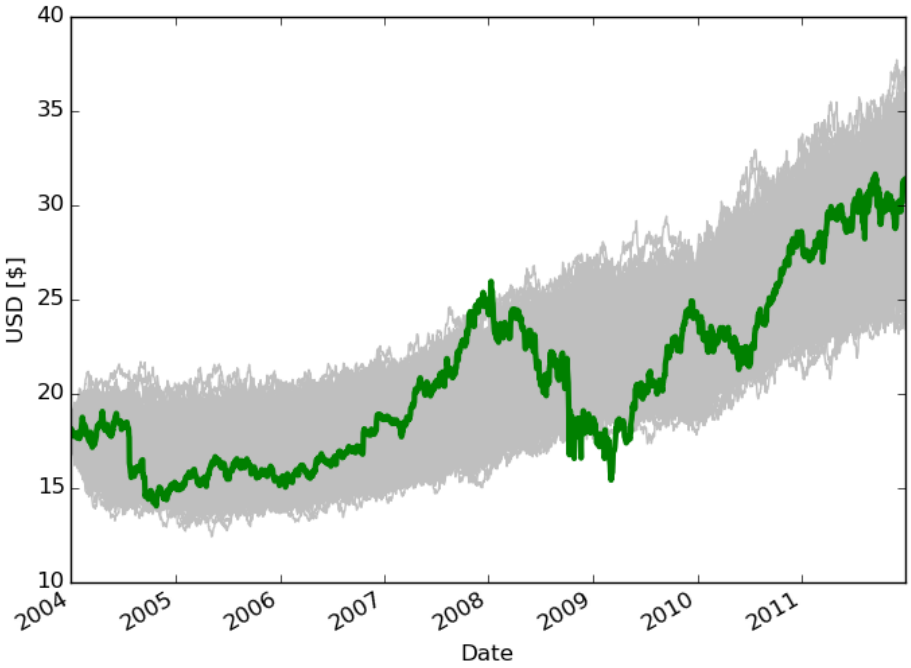


Figure 3.2: Simulation of the stock price of 'The Coca-Cola Company, (NYSE: KO)' during the period 2004-2012, using the value reverting model. The grey lines are 1000 simulations and the green line is the real stock price. The parameters used in the simulation were estimated using data from the period 2002-2007.

same parameters 1000 times, it is possible to observe how the different models behave. The simulations of VRM can be seen in fig. (3.2), where the simulated stock paths are plotted in gray.

Similarly the simulations using GBM can be seen in fig. (3.3). The distribution of the simulated stock prices for the day 2011-12-30 is shown in fig. (3.4). Note the difference in scale for the plots using VRM compared to GBM and also that the stock prices are distributed according to a lognormal distribution.

The inclusion of the intrinsic value in the model of stock price movements, provides a narrower band of possible stock prices compared to the standard model of GBM. This is seen when comparing fig. (3.1) and fig. (3.2). Thus, the inclusion of the fundamental datapoints of the company in a model of stock price movements provides a better prediction of the stock price movements.

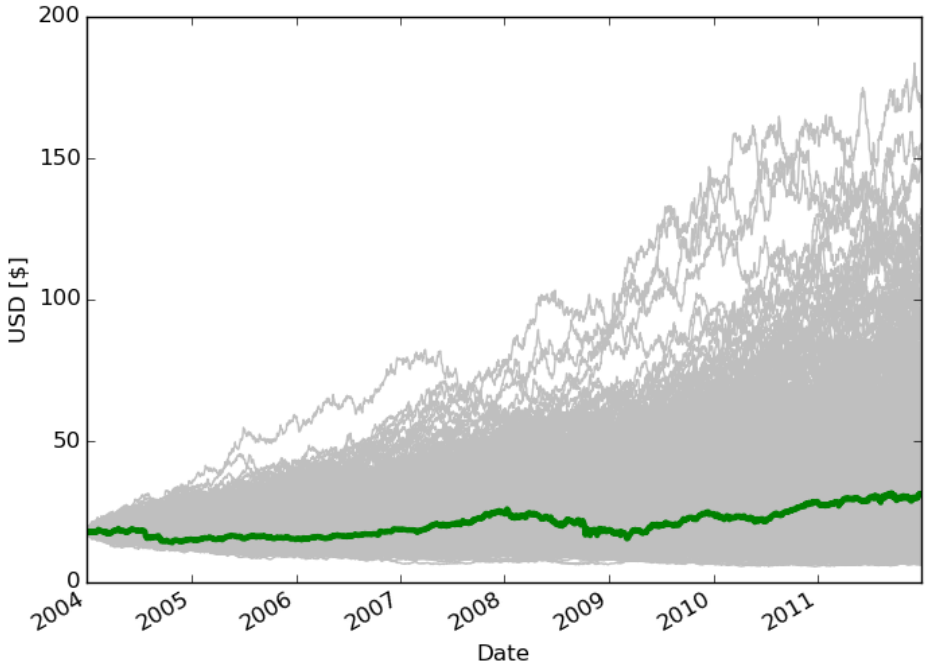
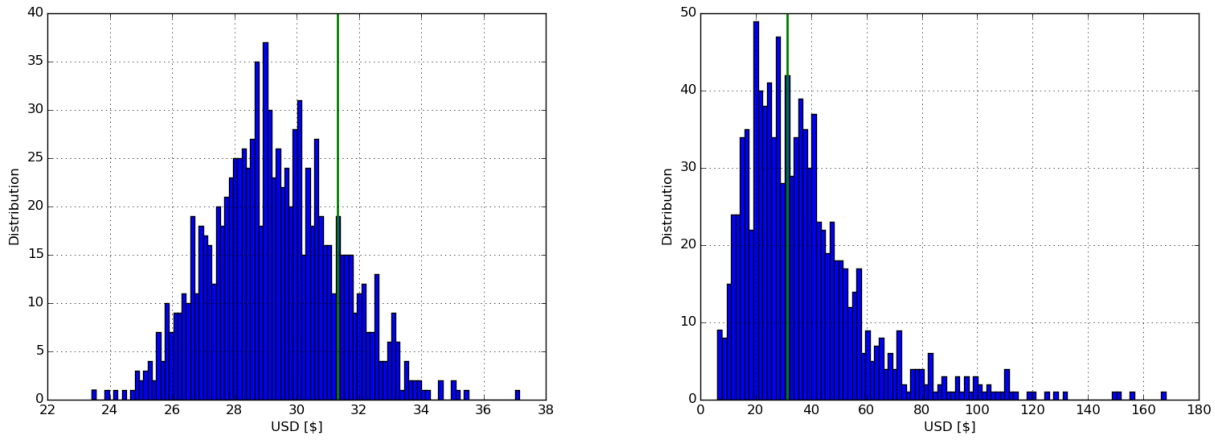


Figure 3.3: Simulation of the stock price of 'The Coca-Cola Company, (NYSE: KO)' during the period 2004-2012, using the Geometric Brownian motion. The grey lines are 1000 simulations and the green line is the real stock price. The parameters used in the simulation were estimated using data from the period 2002-2007.

3.2 Autocorrelation

By looking at the autocorrelation of a stock it is possible to see if the stock price movements are correlated with a lagged version of itself. If there exists noticeable autocorrelation, then it indicates a memory effect where the daily change in stock price depends on the previous changes in stock price. If this is the case, then this behaviour needs to be included in the models of stock price movements.

The autocorrelation for the stock 'The Coca-Cola Company, (NYSE: KO)' was calculated using eq. (2.3.40). In fig. (3.5) the autocorrelation coefficient is plotted as a function of number days the time series are shifted (τ). The $R(\tau)$ is calculated using stock price from the period 2000-2015. From what is seen in fig. (3.5), it would be reasonable to conclude that the autocorrelation is negligible as $R(\tau)$ is evenly distributed around zero for $\tau > 0$. One should keep in mind though that τ is an integer and $\tau = 0$ is simply the time series correlated with it self and by definition $R(0) = 1$ and thus not of interest.



(a) (b)

Figure 3.4: Distribution of the 1000 simulated stock prices at the day '2011-12-30' for 'The Coca-Cola Company, (NYSE: KO)'. In a) the value reverting model is used for the simulations and in b) the share price is simulated using Geometric Brownian motion. The green line is the real stock price at day '2011-12-30'.

However, the autocorrelation can not be considered negligible when considering a group of stocks. This can be seen by studying the mean value of autocorrelation as a function of τ i.e. taking the mean value of R_k of the group of stocks for a fixed τ. The mean value of R_k is calculated as

$$\langle R_k(\tau) \rangle_k = \frac{1}{K} \sum_{k=1}^{K} R_k(\tau) \tag{3.2.3}$$

where K is the number of stocks analysed, $R_k(\tau)$ is the autocorrelation of company k calculated using eq. (2.3.41).

The calculations of autocorrelation were performed using all the stocks listed on the NYSE and NASDAQ during the period 2000-2015. This corresponds to 7164 stocks that were used in the analysis of mean autocorrelation. By calculating $\langle R_k(\tau) \rangle_k$ for the stocks from NYSE and NASDAQ, it is evident that the autocorrelation is no longer negligible, which can be observed in fig. (3.6), where there is a larger amplitude of the mean autocorrelation for $\tau = 1, 2, 3$ compared to how $\langle R_k(\tau) \rangle_k$ behaves for larger values of τ.

It is also of interest to look at the distribution of $R_k(\tau)$ for a fixed τ as one would expect a symmetric distribution around zero. In fig. (3.7) the distribution of $R_k(\tau = 1)$ can be seen. There is an asymmetry in the distribution of $R_k(\tau)$, as well as a shift of $\langle R_k(\tau) \rangle_k$ away from zero for $\tau = 1, .., 4$. For larger values of τ the distribution of R_k is symmetric and the mean autocorrelation oscillate around zero.

There is a big difference in distribution of $R_k(\tau)$ for $\tau = 1$ in fig. (3.7) and larger τ values.

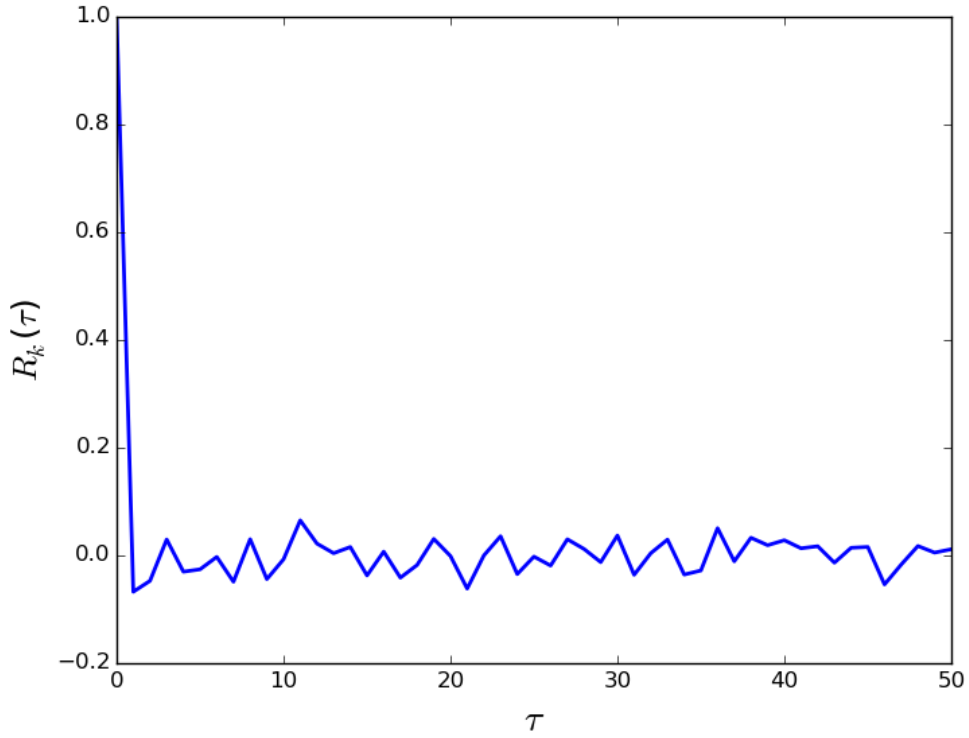


Figure 3.5: The autocorrelation ($R(\tau)$) for 'The Coca-Cola Company, (NYSE: KO)' as a function of τ. $R(\tau)$ is calculated using stock price from the period 2000-2015.

In fig. (3.8) the distribution of $R_k(\tau = 9)$ is plotted. For $\tau = 9$ there is a Gaussian distribution with small standard deviation compared with a case of $\tau = 1$ which has a unsymmetrical distribution with a large standard deviation.

As seen in fig. (3.6), there is noticeable autocorrelation for small values of τ when considering the mean value of the autocorrelation for the stocks trading on NYSE and NASDAQ during the period 2000-2015. So the autocorrelation of stock price movements is not negligible as commonly thought. This mean that there is a memory effect, where the stock price movement depends on the recent stock price movements.

3.3 Probability of movement in opposite direction

A simplified model was used to interpret how small autocorrelations affects the movement of stock prices. The simplified model looks at the probability of the stock to move in the opposite direction from a previous movement, by looking at $dS_t \cdot dS_{t-\tau} < 0$ compared to $dS_t \cdot dS_{t-\tau} \neq 0$. This model will be referred to as the probability of movement in opposite

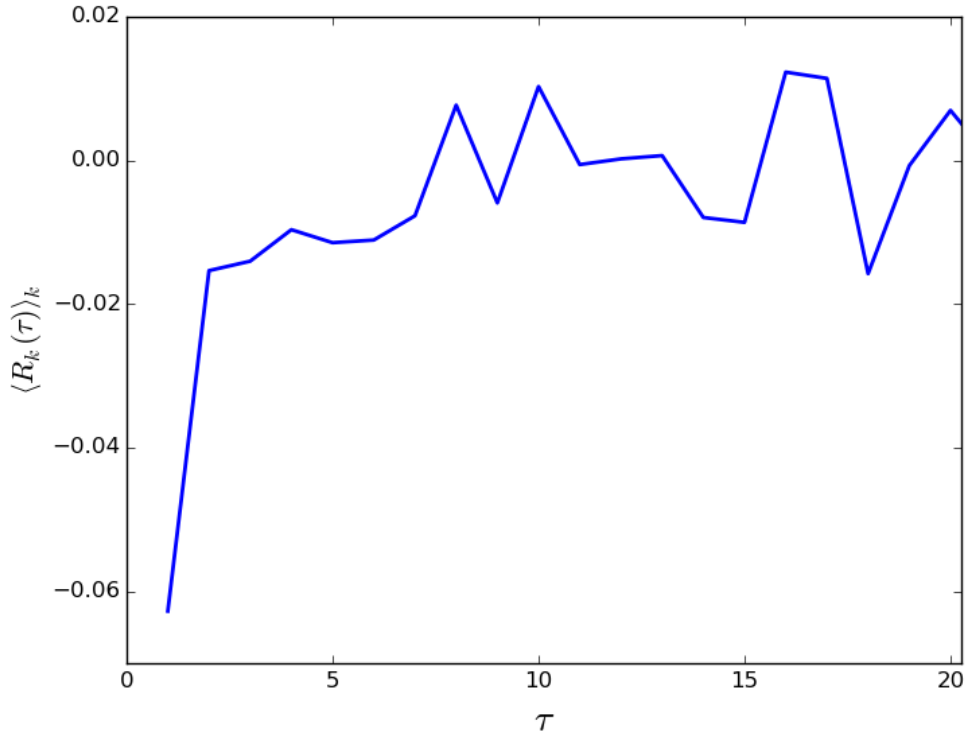


Figure 3.6: The mean autocorrelation ($\langle R_k(\tau)\rangle_k$) for stocks trading on NYSE and NASDAQ during the period 2000-2015 as a function of the time lag τ in days, starting from $\tau = 1$.

direction (PMOD).

The daily movement of a stock (dS_t) is mostly influenced by the stochastic term $\sigma S_t dW_t$, which will be distributed symmetrically around zero. So PMOD of stock movement will be analogous to a series of coin flips, where each day an increase in stock price ($dS_t > 0$) corresponds to heads and a decrease in stock price ($dS_t < 0$) corresponds to tails. Thus, it is the direction of the movement that is of interest and not the size of the movement. So the amplitude of the movement is not included in the analysis, as well as an inability to process $dS_t \cdot dS_{t-\tau} = 0$, which would be equivalent to the coin landing on its edge. So for the case of $\tau = 1$ it corresponds to the probability of two consecutive coin tosses to show different sides i.e. heads(H) and tails(T).

There are four possible outcomes for two consecutive coin flips, H - H, H - T, T - H and T - T. So from the law of large numbers one would expect to see the probability for H - T, T - H go towards .5, as the number of flips analysed increase. Let $P(\tau)$ be this probability of movement in opposite direction.

It is possible to study the how $P(\tau)$ should be distributed when looking at multiple series of coin flips. Let X be the outcome for the analysis of two consecutive flips ($\tau = 1$), where

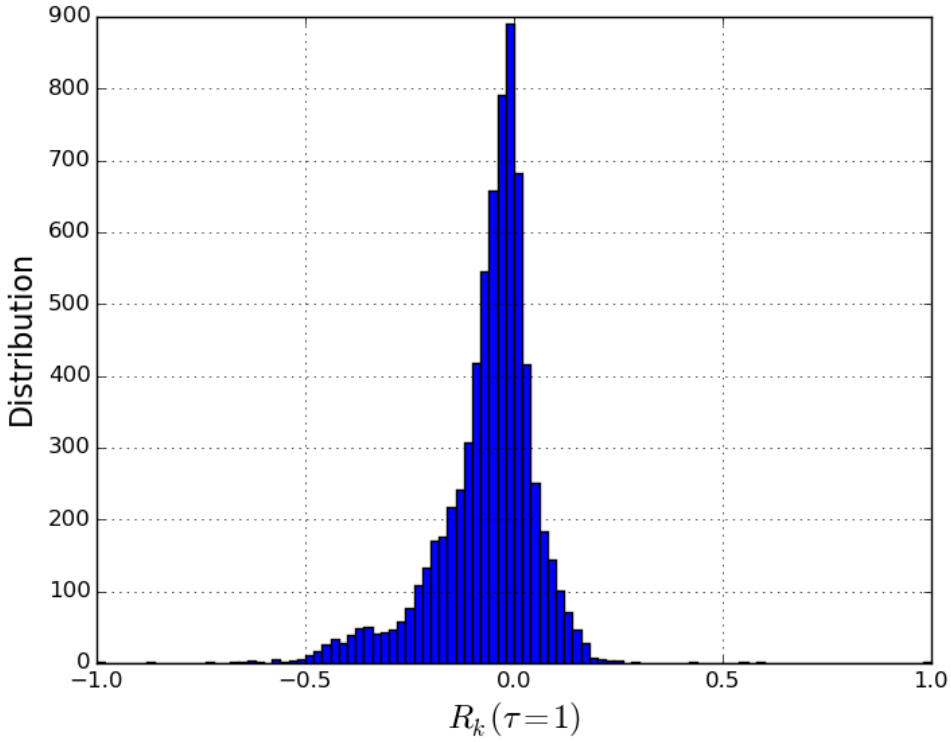


Figure 3.7: The distribution of autocorrelation coefficients (R_k) for $\tau = 1$. R_k is calculated for the stocks trading on NYSE and NASDAQ during the period 2000-2015, which resulted in 7164 autocorrelations coefficients.

we assign the value one if there are different sides of the coins H - T and T - H and the value zero if it is the same side twice i.e. H - H and T - T. Then X will have the expected value

$$E(X) = 1/2(1) + 1/2(0) = 1/2, \tag{3.3.4}$$

with the standard deviation

$$\sigma(X) = \sqrt{var(X)} = \sqrt{1/2(1-1/2)^2 + 1/2(0-1/2)^2} = 1/2. \tag{3.3.5}$$

According to the CLT $P(\tau)$ will be distributed according to a Gaussian with $\mu = 1/2$ and $\sigma = (1/2)/\sqrt{(N-1)}$, where N is the number of coin flips in one series, this assumes that the series of coin flips have the same length. This is not the case when looking at a group of stocks that forms an index, but the majority of the stocks will have approximately the same number of trading days during the analysed period. So a $P(\tau)$ is still expected to be distributed according to a Gaussian.

The calculation of the probability of movement in opposite direction for a stock, $S^k(t)$ with T^k number of data points for company k is performed by looking at the number of times

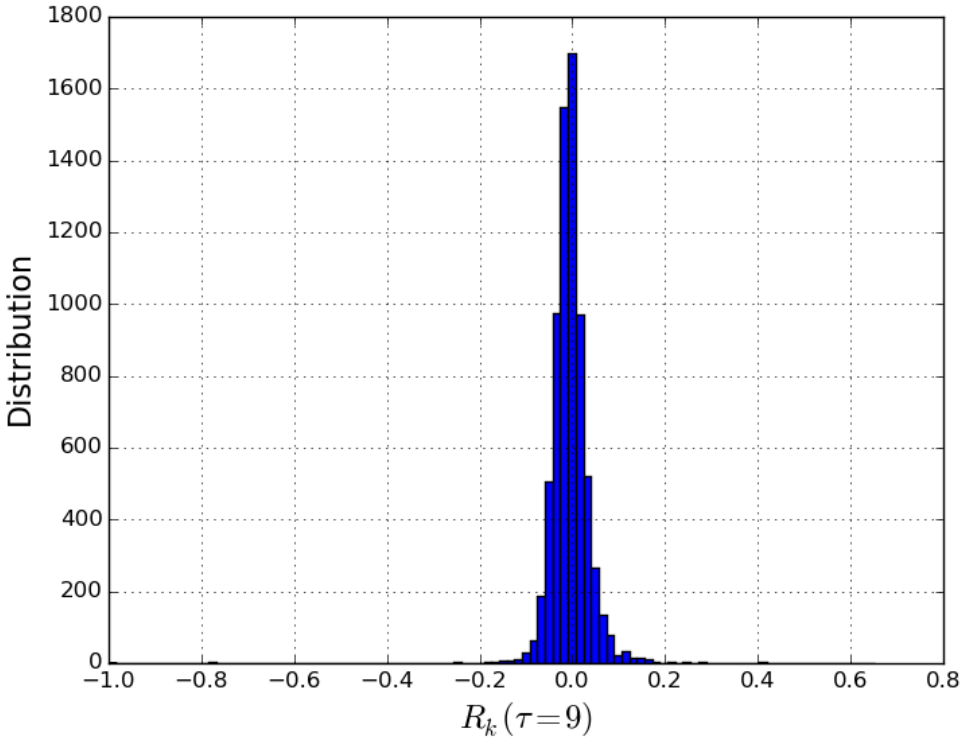


Figure 3.8: The distribution of autocorrelation coefficients (R_k) for $\tau = 9$. R_k is calculated for the stocks of NYSE and NASDAQ during the period 2000-2015, which resulted in 7164 autocorrelations coefficients.

the stock moves in the opposite direction

$$O^k(t,\tau) = \begin{cases} 1, & \text{if } dS_t^k \cdot dS_{t-\tau}^k < 0 \\ 0, & \text{otherwise} \end{cases}. \tag{3.3.6}$$

compared to the number of time their is actual change in the stock price

$$N^k(t,\tau) = \begin{cases} 1, & \text{if } dS_t^k \cdot dS_{t-\tau}^k \neq 0 \\ 0, & \text{otherwise} \end{cases}. \tag{3.3.7}$$

Using this it is possible to express $P^k(\tau)$ as

$$P^k(\tau) = \frac{\sum_{t=1}^{T^k} O^k(t,\tau)}{\sum_{t=1}^{T^k} N^k(t,\tau)}, \tag{3.3.8}$$

where $dS_t^k = S_t^k - S_{t-1}^k$ and T^k is the number of data points for company k.

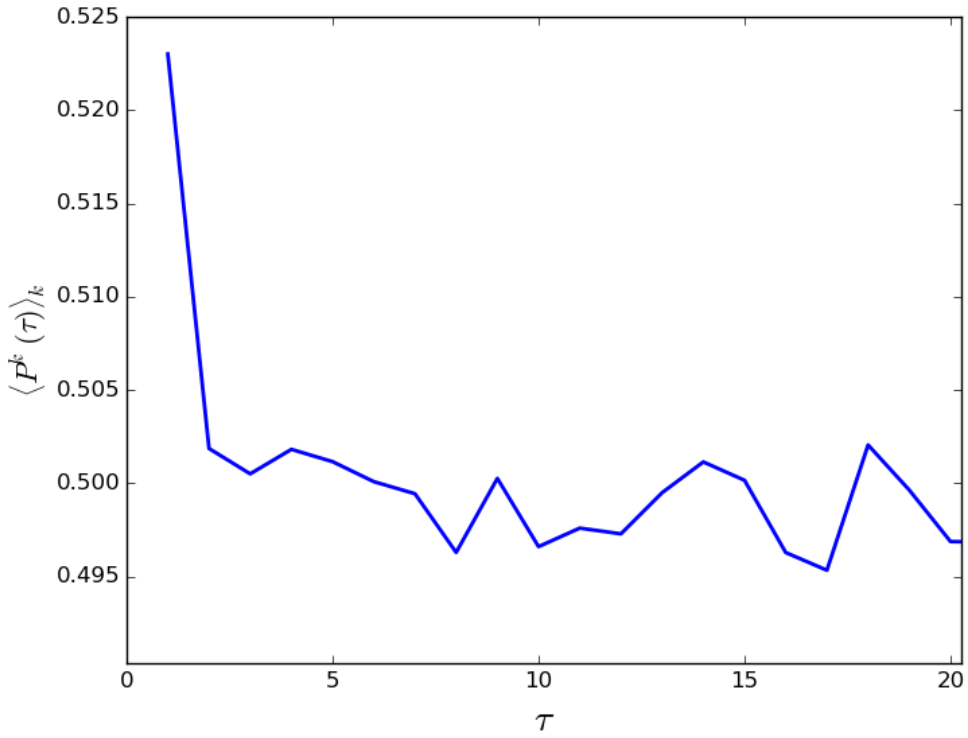


Figure 3.9: The mean probability of movement in opposite direction ($\langle P^k(\tau)\rangle_k$) as a function of τ in days. Where $\langle P^k(\tau)\rangle_k$ is calculated using the stocks trading on NYSE and NASDAQ during the period 2000-2015.

In fig. (3.9) the mean value of P^k is calculated for all companies in NYSE and NASDAQ during the period 2000-2015. The mean value of P^k for a given τ is given by

$$\langle P^k(\tau)\rangle_k = \frac{1}{K}\sum_{k=1}^{K} P^k(\tau) \tag{3.3.9}$$

where K is the number of companies.

The distribution of $P^k(\tau = 1)$ and $P^k(\tau = 9)$ for the stocks trading on NYSE and NASDAQ during the period 2000-2015 can be seen in fig. (3.10) and fig. (3.11).

Note that a negative autocorrelation would mean a PMOD larger than .5 and this is what is observed for small values of τ, as shown in fig. (3.9). This is not unexpected as PMOD looks at a similar property as autocorrelation, but PMOD does not include the amplitude of the movement. So PMOD agrees with the autocorrelation for small values of τ and also indicates a memory effect for stock price movements.

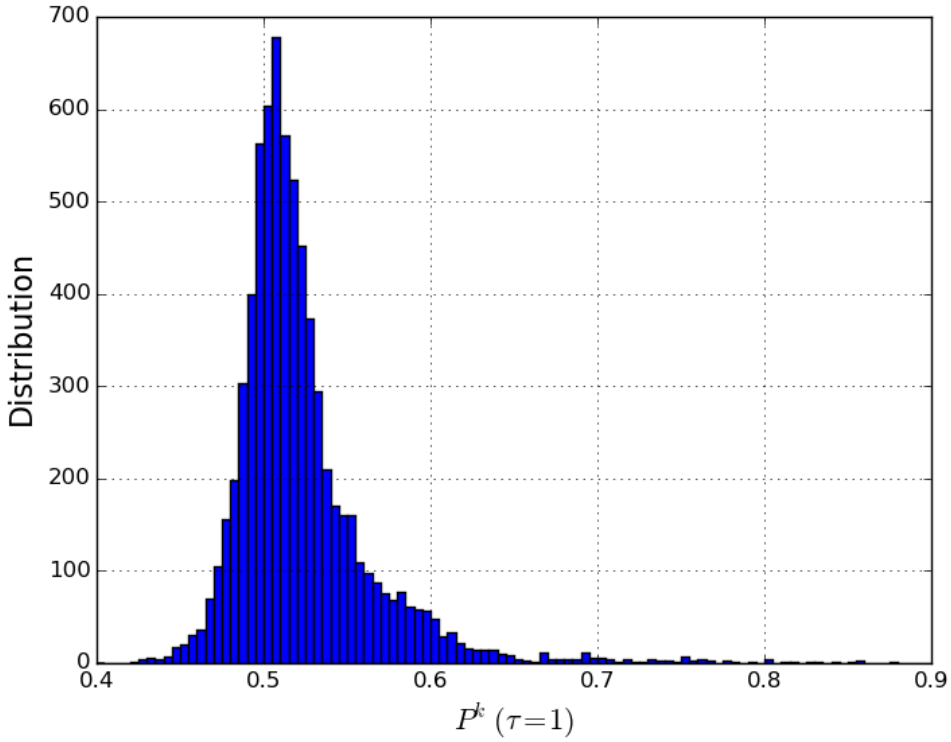


Figure 3.10: The distribution of probability of movement in opposite direction ($P^k(\tau)$) for $\tau = 1$. Where P^k is calculated using the stocks trading on NYSE and NASDAQ during the period 2000-2015, which are 7164 stocks.

3.4 Autoregressive Model

In order to include the memory effect of stock price movements that is observed when studying autocorrelation and PMOD, an autoregressive model is used. But to use an autoregressive model as eq. (2.4.42) to model a stock price movements as well as capture the behaviour of autocorrelations and PMOD, the stochastic term needs to be modified in such a way that it behaves as GBM. This is done by using the stochastic term $\sigma S_t^k dW_t^k$ instead of dW_t^k in eq. (2.4.42), giving the expression

$$dS_t^k = \sum_{i=1}^{p} \phi_i^k dS_{t-i}^k + \sigma^k S_t^k dW_t^k. \tag{3.4.10}$$

S_t^k is the stock price of company k at time t, σ^k is the volatility parameter for company k and the parameters $\phi_1^k, ..., \phi_p^k$ are used to create the behaviour of autocorrelation for company k and p is the order of the autoregressive term.

Eq. (3.4.10) does not explain the drift of stock prices, so a drift term needs to be added as

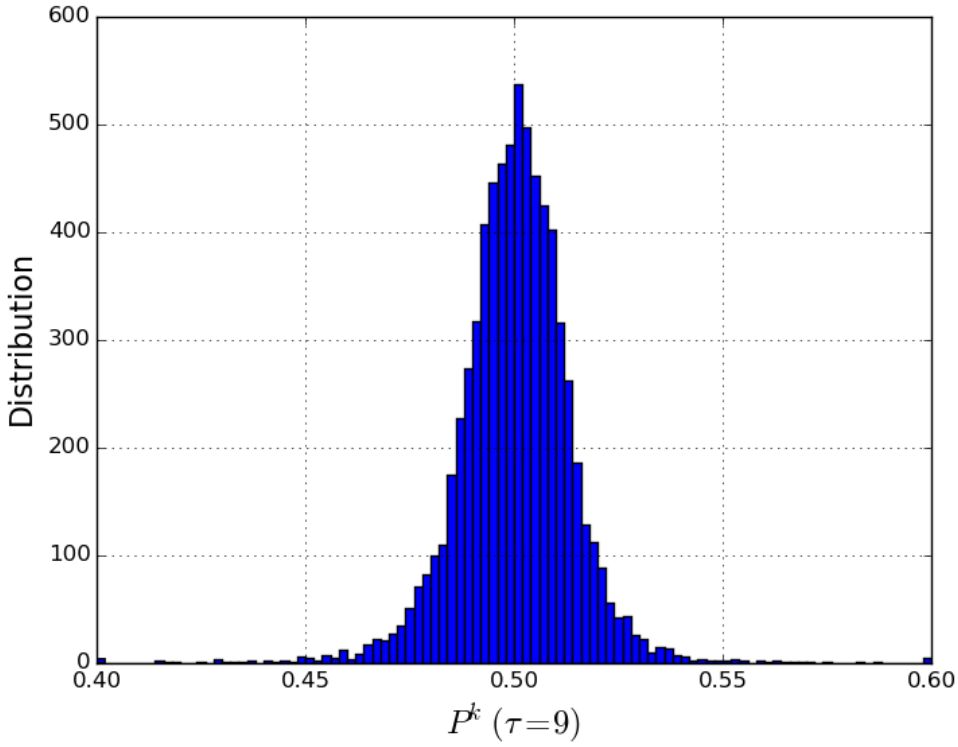


Figure 3.11: The distribution of probability of movement in opposite direction ($P^k(\tau)$) for $\tau = 9$. P^k is calculated using the stocks trading on NYSE and NASDAQ during the period 2000-2015, which are 7164 stocks.

well. For simplicity the drift term from GBM is used, as the autocorrelation only looks at effects from short time-spans. So it will not make a difference if the drift term from GBM or VRM is used. Combining an AR model with GBM gives

$$dS_t^k = \sum_{i=1}^{p} \phi_i^k dS_{t-i}^k + \mu^k S_t^k dt + \sigma^k S_t^k dW_t^k, \tag{3.4.11}$$

where S_t^k is the stock price of company k at time t, σ^k is the volatility parameter and μ^k is the drift parameter for company k. The parameters $\phi_1^k, ..., \phi_p^k$ are used to create the behaviour of autocorrelation for company k and p is the order of the autoregressive term.

Eq. (3.4.11) is used to simulate the stocks of NYSE and NASDAQ during the period 2000-2015. As there existed a deviation away from zero in the mean autocorrelation, as well as an asymmetry in the distribution of $R_k(\tau)$ for $\tau = 1, ..., 4$ in the real data, the autoregressive term was selected to be of the fourth order i.e. $p = 4$. The parameters σ^k, μ^k and ϕ_i^k where $i = 1, ..., 4$ are estimated for each k using a least mean square procedure fitting eq. (3.4.11) to each stock during the period 2000-2015.

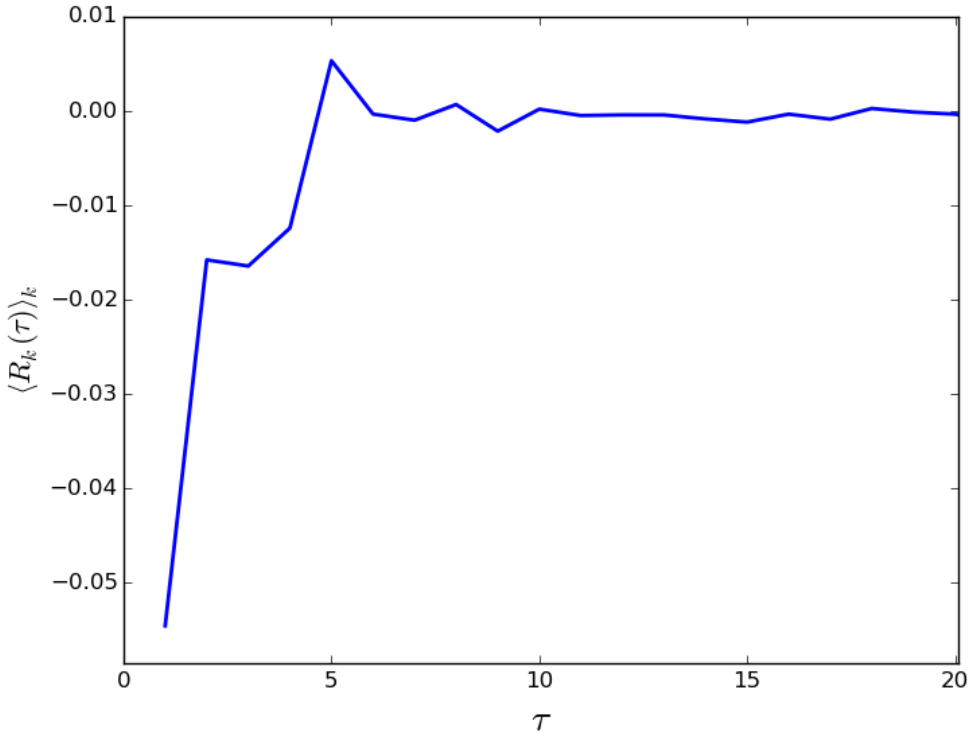


Figure 3.12: The mean autocorrelation coefficient ($\langle R_k(\tau)\rangle_k$) as a function of τ for simulated data. The parameters used in the simulations were estimated using the stocks trading on the NYSE and NASDAQ during the period 2000-2015, resulting in 7164 autocorrelation coefficients.

Eq. (3.4.11) was used with the estimated parameters to simulate a stock price for the stocks trading on NYSE and NASDAQ during the period 2000-2015. The simulated stock price was the same as for the real stock in the first time step and the same number of time steps as the real stock. Thus, a new data set was created that has the same number of stocks and with the same amount of trading days as for the each company as the data set previously used when looking at the autocorrelation and PMOD.

The simulated stocks were then analysed in the same way as before, by calculating the autocorrelation coefficient using eq. (2.3.41). The mean autocorrelation $\langle R_k(\tau)\rangle_k$ can be seen in fig. (3.12) plotted as a function of τ. The distribution of R_k for $\tau = 1$ can be seen in fig. (3.14). The same thing can be done for PMOD by calculating $P^k(\tau)$ using eq. (3.3.8) and calculate $\left\langle P^k(\tau)\right\rangle_k$ using eq. (3.3.9) as a function of τ, which is shown in fig. (3.13).

When the autoregressive term is of the first order, it is possible to calculate the expectation value of the share price. Start by considering eq. (3.4.11) with the autoregressive term of

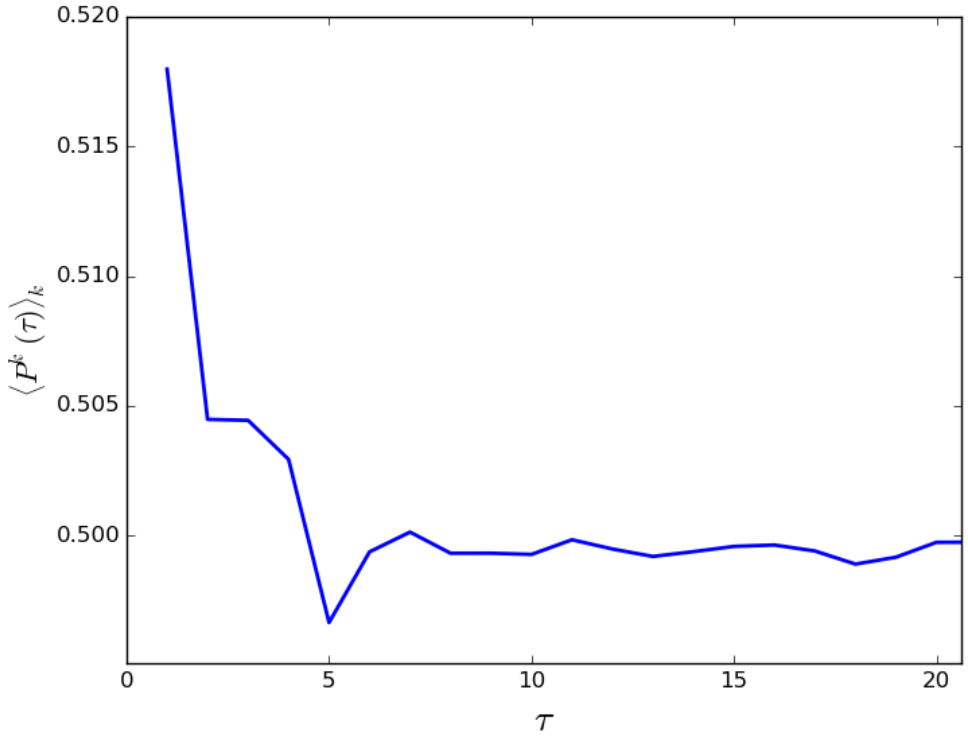


Figure 3.13: The mean probability of movement in opposite direction ($\langle R_k(\tau)\rangle_k$) as a function of τ, calculated for stocks simulated using GBM with an additional autoregressive term of the forth order. The parameters were estimated from the stocks trading on the NYSE and NASDAQ during the period 2000-2015, which corresponds to 7164 stocks.

the first order, given by

$$dS_t = \phi dS_{t-1} + \mu S_t dt + \sigma S_t dW_t. \tag{3.4.12}$$

Then the expectation value of the share price can be calculated by

$$dE[S_t] = \phi dE[S_{t-1}] + \mu E[S_t]dt + \sigma E[S_t]E[dW_t] \tag{3.4.13}$$

where E is the expectation value operator. Using $E[dW_t] = 0$ and $dE[S_{t-1}] = dE[S_t]$ gives the expression

$$dE[S_t] = \frac{\mu}{1-\phi} E[S_t]dt. \tag{3.4.14}$$

Since it is a deterministic differential equation it can be solved by looking for the function $a(t) = E[S_t]$ that solves the differential equation, $a'(t) = \beta a(t)$ with the condition $a(0) = S_0$ where $\beta = \frac{\mu}{1-\phi}$.

The solution is given by $a(t) = S_0 \exp[\beta t]$, using it one obtains the expression for the

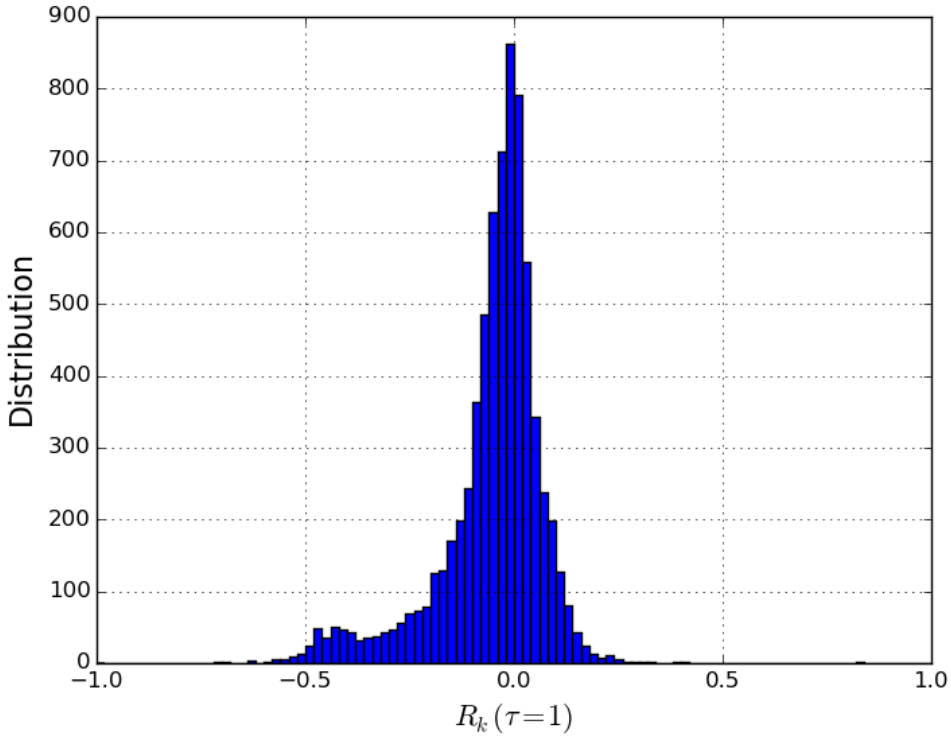


Figure 3.14: The distribution of autocorrelation coefficients (R_k) for the simulated data for $\tau = 1$. R_k is calculated for stocks simulated using GBM with an autoregressive term of the forth order. The parameters were estimated for the stocks from trading on NYSE and NASDAQ during the period 2000-2015, which corresponds to 7164 stocks.

expectation value of S_t

$$E[S_t] = S_0 \exp[\beta t] = S_0 \exp\left[\frac{\mu}{1-\phi} t\right]. \tag{3.4.15}$$

Note that when $\rho \in (0, 1)$, the autoregressive term will make the derivative of $E(S_t)$ larger, which will make $E(S_t)$ grow faster than the GBM in the case $\mu > 0$. The expectation value of the stock simulated with GBM is obtained when $\rho = 0$. When $\rho < 0$ the autoregressive term will reduce the second derivative of $E(S_t)$ and thus make it will grow slower compared to GBM when $\mu > 0$.

The addition of the autoregressive term of the fourth order to GBM captures the observed behaviour of autocorrelation and PMOD for a group of stocks for small values of τ. Thus it captures the observed memory effect for small values of τ. But both the mean value of the autocorrelation in fig. (3.12) and the mean value of PMOD in fig. (3.13) shows smaller volatility for large values of τ compared to what is seen for real data.

3.5 Index

A model of stock movements should also be consistent when looking at the movements of a group of stocks. This can be studied using a stock index, which calculates the collective movements of the group of stocks. No model of stock price movements can predict a stock market crash, but a model should allow market crashes to occur.

3.5.1 Equally weighted index

An equally weighted index using eq. (2.1.2) for a group of stocks is used to avoid a stock with a very high market capitalization being responsible for large movements of the index and thus obscuring the results, which could be possible if an index is weighted by market capitalization as in eq. (2.1.1).

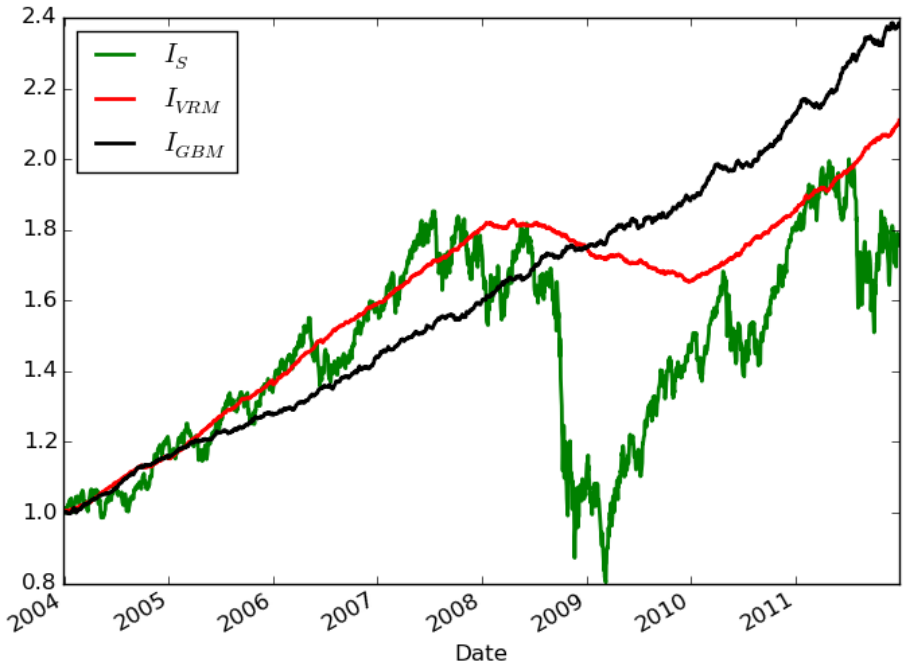


Figure 3.15: The simulation of an equally weighted index using value reverting model (I_{VRM}) and Geometric Brownian motion (I_{GBM}) as well as the real data(I_S), for the period 2004-2012. The indices were calculated with the stocks from the NYSE that had sufficiently many data points to estimate the parameters of GBM and VRM, which were 647 stocks.

For simplicity each stock is set equal to one at t_0^k and the entire index equal to one at t_0. t_0^k is the first trading day for the stock k during the selected period and t_0 is the first trading day of the selected period. This equally weighted index is calculated by the expression

$$I(t) = \frac{1}{K_t} \sum_{k=1}^{K_t} \frac{1}{S^k(t_0^k)} S^k(t), \tag{3.5.16}$$

where K_t is the number of stocks in the index at time t and $S^k(t)$ is the stock price of company k at time t.

The equally weighted index of NYSE for the period 2002-2012 was simulated using VRM and GBM. The value term (V_t^k) was calculated using eq. (2.1.7) using parameters estimated using data from the time period 2002-2007. The parameters of GBM and VRM were estimated using a least mean squares approach for eq. (3.1.1) and eq. (2.2.18) with the stocks trading on NYSE during the period 2002-2012. The parameters were then used to simulate the stock price movements of each company using VRM according to eq. (3.1.1) and GBM using eq. (2.2.18) where the simulated stock had the same initial value at t_0^k as the real stock. There were 712 stocks that traded on the NYSE during the period 2002-2012 where it was possible to calculate V_t during at least one year. As three fundamental data points are needed to calculate the V_t the first time, it was only possible to simulate the stock price during the period 2004-2012.

For the estimation of value to be good, the business needs to be engaged in similar operations during the time analysed. So for example, a company like 'The Coca-Cola Company, (NYSE: KO)' does the same thing without any big changes during the period, will make the value far more stable and predictable. However, the market crash 2008 caused some companies to alter the way they were doing business dramatically.

For example, during the crash 2008 American International Group, Inc (AIG) got a bailout of 85 billion dollars from the U.S. government and as a result, the U.S. government received nearly 80% of the firm's equity. This caused a major change in AIG's business and resulted in a drop in the AIG's share price from above 400 $ just before the crash to below 10$ a few months after the crash. The valuation model is not capable of taking this into account and thus gives a very bad estimate of the value (V_t), which then has an effect on VRM.

So in order to avoid cases like this a filter was applied to the simulated data. The filter removed any stock that became 100 times larger than the initial value during the simulation from both the GBM and VRM simulation. The number of companies that passed this screen was 647 and 65 stocks were removed.

The stocks that passed the filter were then used to calculate an equally weighted index using eq. (3.5.16) for the stocks simulated using VRM and GBM as well as the real stocks. The index for the real data is denoted I_S, the stocks simulated using the VRM denoted I_{VRM} and the stocks simulated using GBM are denoted I_{GBM}, the indices can be seen in fig. (3.15).

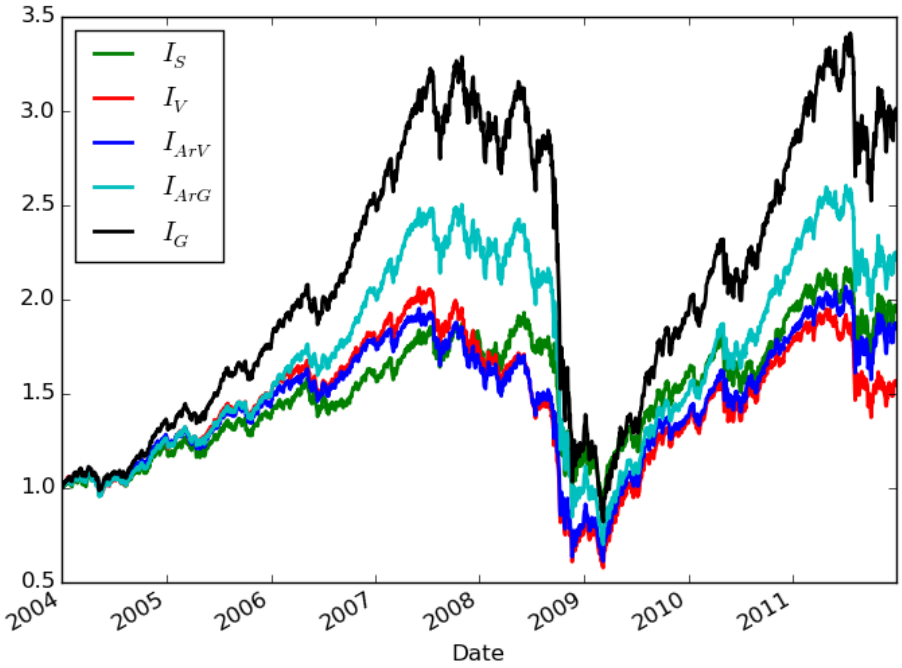


Figure 3.16: Simulated paths of equally weighted indices for the period 2004-2012, using value reverting model (I_V) and Geometric Brownian motion (I_G) with the addition of an index term. The indices are also simulated with an additional autoregressive term, I_{ArV} and I_{ArG}. The index calculated using the real data (I_S) is also included. The indices were calculated with the stocks from the NYSE that had sufficiently many data points to estimate the parameters of GBM and VRM, which were 622 stocks.

The fact that there is so small volatility of I_{VRM} and I_{GBM} is due to the stochastic part of GBM and VRM consisting of randomly drawn numbers $\varepsilon_t^k \sim N(0,1)$. ε_t^k is independent drawn for each company and for each time step. So if one considers the mean value of ε_t^k for the companies k in each time step,

$$\left\langle \varepsilon_t^k \right\rangle_k = \frac{1}{K}\sum_{k=1}^{K} \varepsilon_t^k \Longrightarrow \left\langle \varepsilon_t^k \right\rangle_k \in N\left(0, \frac{1}{\sqrt{K}}\right), \tag{3.5.17}$$

where K is the number of companies used to calculate the index at time t with the companies $k = 1, ..., K$. So $\left\langle \varepsilon_t^k \right\rangle_k$ will follow the distribution $N\left(0, \frac{1}{\sqrt{K}}\right)$ [1]. The stochastic term $\sigma^k S_t^k \varepsilon_t^k$, consist of two additional variables. This means that the variance of each time step may not distributed exactly according to $N\left(0, \frac{1}{\sqrt{K}}\right)$ as $\sigma^k S_t^k$ will make the influence the impact of the weighting when calculating the index. But the variance of the index will

decrease as K increase according to $\frac{1}{\sqrt{K}}$.

3.5.2 Inclusion of an index term

The difference in volatility for both the simulated I_{VRM} and I_{GBM} compared to the actual data (I_S) indicates that there may be some correlations in the daily stock moment for the stocks. The inclusion of an autoregressive term to the model of stock prices as in eq. (3.4.11) does not change the fact that there is no index behaviour.

So by introducing a stochastic variable that is the same for all the stocks will potentially create an index behaviour. The index term should have a similar structure as the stochastic term of GBM in order to allow different degrees of influence of this term on the stock price. With the same structure as GBM, the index term can be expressed as

$$\sigma_I^k S_t^k dW_t \tag{3.5.18}$$

The movement of the index dW_t is the same for all the stocks. S_t^k is the stock price of company k at time t and σ_I^k is a constant for company k. The value of for σ_I^k can be both positive and negative; a negative value will create anti-correlation with dW_t and if it is positive the stock will be positively correlated with dW_t. So this new index term makes correlations between the stocks possible, and an index behaviour can occur.

So adding the index term to eq. (2.2.20) gives GBM with the additional index term,

$$dS_t^k = \mu^k S_t^k dt + \sigma_k^k S_t^k dW_t^k + \sigma_I^k S_t^k dW_t. \tag{3.5.19}$$

Including the autoregressive term and the index term to GBM gives

$$dS_t^k = \sum_{i=1}^{p} \phi_i^k dS_{t-i}^k + \mu^k S_t^k dt + \sigma_k^k S_t^k dW_t^k + \sigma_I^k S_t^k dW_t. \tag{3.5.20}$$

S_t^k is the stock price of company k at time t and dW_t is an increment of a Wiener process and is the same for the group of companies analysed at time t and σ_I^k is a term corresponding to the index term. Just as for GBM μ^k is a drift term, σ_k^k is the volatility and ϕ_i^k are constants for the autoregressive term where p gives the order of the autoregressive term.

The same thing can be preformed for the VRM by adding the index term to eq. (3.1.1), which gives

$$dS_t^k = -\alpha^k(S_t^k - V_t^k)dt + \sigma_k^k S_t^k dW_t^k + dV_t^k + \sigma_I^k S_t^k dW_t. \tag{3.5.21}$$

Adding the autoregressive term and the index term to VRM gives

$$dS_t^k = \sum_{i=1}^{p} \phi_i^k dS_{t-i}^k - \alpha^k(S_t^k - V_t^k)dt + \sigma_k^k S_t^k dW_t^k + dV_t^k + \sigma_I^k S_t^k dW_t, \tag{3.5.22}$$

where α^k is the rate of reversion towards the value, S_t^k and V_t^k are the share price and value of company k at time t. dW_t is an increment of a Wiener process, σ_k^k gives the volatility of the share price that is independent of the index behaviour and σ_I^k is the volatility associated with the index behaviour. Just as before ϕ_i^k are the constants of the autoregressive term and p gives the order of the autoregressive term.

To compare the simulated stocks with real data, dW_t needs to be the same in the simulations as for the real data. So for the simulations, the stochastic part of the index term is estimated using real data by calculating the mean value of the return in each time step for the group of companies

$$dI_t = \left\langle r_t^k \right\rangle_k = \frac{1}{K} \sum_{k=1}^{K} \frac{S_t^k - S_{t-1}^k}{S_{t-1}^k} \tag{3.5.23}$$

where K is the number of stocks being analysed, S_t^k is the share price of company k at time t. This means that dI_t is used in stead of dW_t in the simulations.

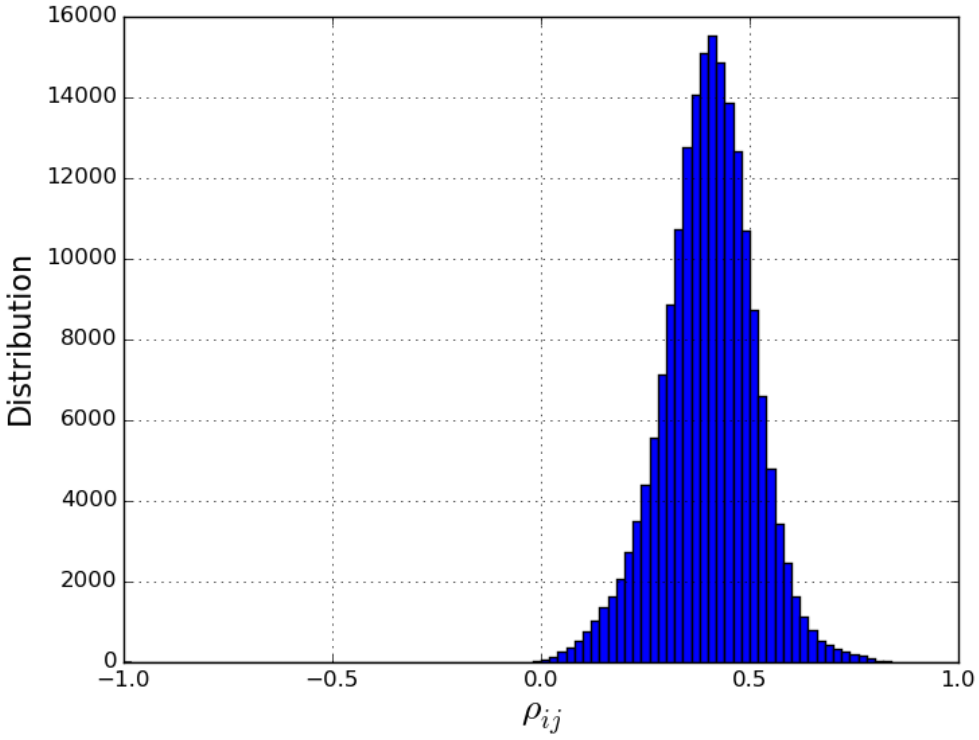


Figure 3.17: Distribution of the correlation coefficients (ρ_{ij}). ρ_{ij} were calculated using the stocks from NYSE during the period 2004-2012 that passed the selection criteria. This resulted in 193131 correlation coefficients.

The equally weighted index of NYSE is simulated for the period 2004-2012, by simulating the individual stocks from the NYSE using the four models eq. (3.5.19), eq. (3.5.20), eq.

(3.5.21) and eq. (3.5.22). The value term (V_t^k) was calculated using eq. (2.1.7) using parameters estimated using data from the time period 2002-2007.

The parameters for eq. (3.5.19), eq. (3.5.20), eq. (3.5.21) and eq. (3.5.22) were estimated using data from the period 2002-2012. The autoregressive term was selected to be of the fourth order to agree with the observed behaviour for autocorrelation and POD.

There were 712 stocks that traded at least four consecutive years during the period 2002-2012 for which all the data points required were accessible. As before a filter was applied to the simulated data. The filter removed any stock that became 100 times larger than the initial value during the simulation from both the GBM and VRM simulation. The number of companies that passed this screen was 622 out of 712 companies.

The stocks that passed the filter were then used to calculate an equally weighted index using eq. (3.5.16). The index calculated using real data is denoted $I_S(t)$, the index calculated with stocks simulated using VRM with the additional index term (eq. (3.5.21) is denoted $I_V(t)$. When the index is calculated using stocks simulated using VRM with the additional index term and an autoregressive term (eq. (3.5.22)) it is denoted I_{ArV}.

When the stocks used to calculate the index are simulated using GBM with the additional index term (eq. (3.5.19)) it is denoted $I_G(t)$ and when the stocks are simulated using GBM with the additional index term and an autoregressive term (eq. (3.5.20)) it is denoted I_{ArG}. The indices are plotted in fig. (3.15).

The inclusion of the new index term resulted in an index behaviour and the market crash that occurred in late 2008 is captured in this model. The crash is captured due to the fact that dI_t is calculated with real data. The inclusion of the autoregressive term increased the agreement of both VRM and GBM to the real data, which can be seen in fig. (3.15), but it is VRM with the additional index term and the autoregressive term (I_{ArV}) which is the closest to the index calculated with real data ($I_S(t)$).

3.5.3 Correlations

Now with the inclusion of the index term, an index behaviour is achieved. So now it is of interest to study how the simulated stocks correlate with each other and how this compares with the real data. The data simulated with VRM with the additional index term and the autoregressive term is used for the analysis of correlations between the stocks, as it had the best agreement with the index calculated with real data.

Using eq. (2.3.36) the correlation coefficients ρ_{ij} are calculated using the returns the entire time interval. Fig. (3.17) shows the distribution of the unique correlations coefficients for the real data where the stocks correlation with itself is not included. This can be compared with fig. (3.18) where the unique correlations coefficients (ρ_{ij}) are plotted for the stocks simulated using VRM in combination with an autoregressive term and an index term according to eq. (3.5.22).

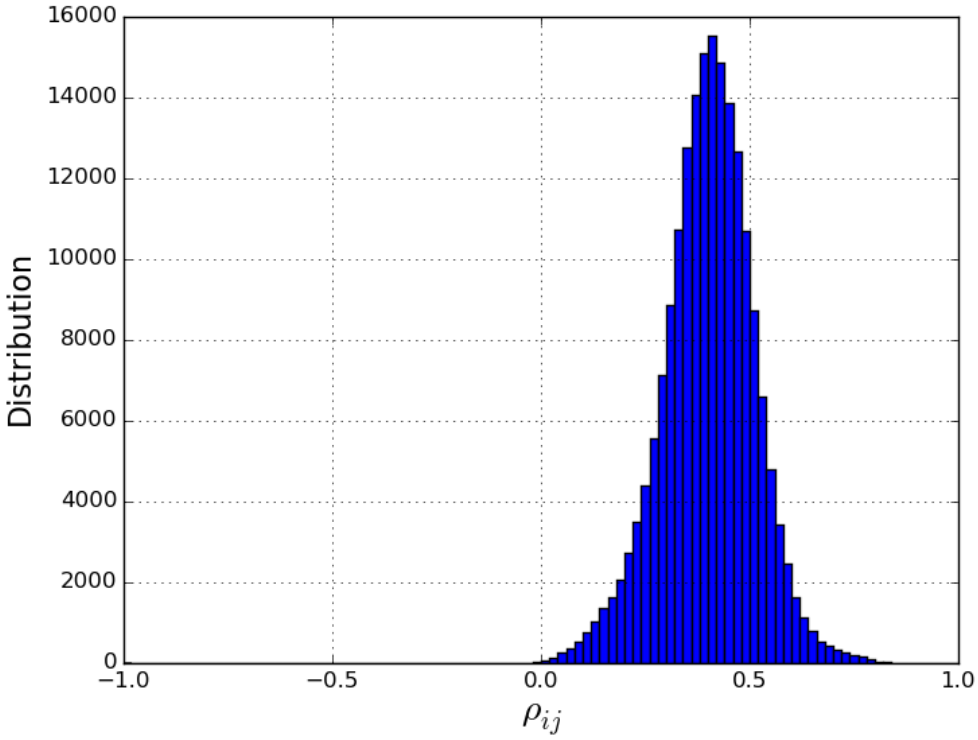


Figure 3.18: Distribution of the correlation coefficients (ρ_{ij}) for simulated stocks from NYSE that passed the selection criteria. The stocks were simulated using a value reverting model with the an index term and an autoregressive term of the fourth order, which resulted in 193131 correlation coefficients.

Similarities in the distribution of correlation coefficient do not show how close the real and simulated correlation matrices agree. By comparing the correlations coefficients for the stocks i and j for the real data (ρ_{ij}) and the simulated data (ρ_{ij}^{sim}) it is possible to see how well the simulation and real data agree. So by looking at the ratio

$$D_{ij} = \frac{\rho_{ij}^{sim}}{\rho_{ij}}, \tag{3.5.24}$$

this can be studied. If the simulated and real correlations coefficients are close to each other D_{ij} will be close to one.

By looking at D_{ij} on a year by year basis it is possible to see how it changes with time. The mean value and standard deviation of D_{ij} were used to see how it D_{ij} changes over time for such a large sample. Where the mean value (m) is given by

$$m = \langle D_{ij} \rangle , \tag{3.5.25}$$

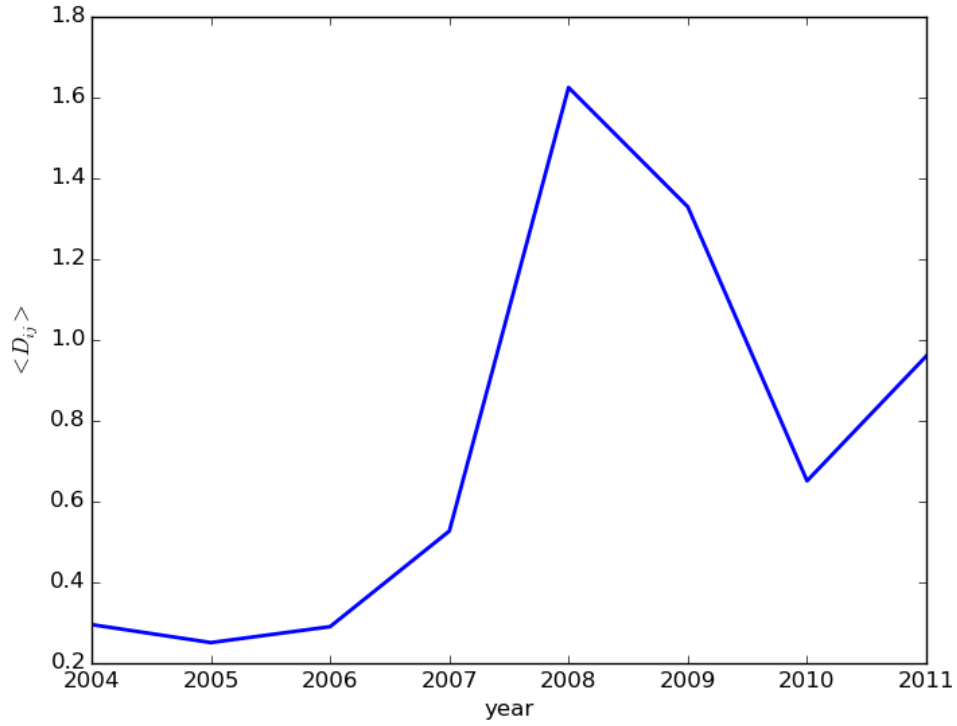


Figure 3.19: The mean value (m) of the ratio between correlation coefficients calculated using simulated data and the correlation coefficient calculated using real data is plotted as a function of time. The stocks were simulated using a value reverting model with the an index term and an autoregressive term of the fourth order. The ratio was calculated using the stocks from NYSE that passed the selection criteria, which were 622 stocks.

and the standard deviation(σ) is calculated according to

$$\sigma = \left\langle (D_{ij} - \langle D_{ij} \rangle)^2 \right\rangle . \tag{3.5.26}$$

If ρ_{ij}^{sim} is close to ρ_{ij} it would result in m oscillating around one and with a small σ. In fig. (3.19) m is plotted as a function of time and in fig. (3.20) σ is plotted as a function of time where m and σ are calculated using data for a year.

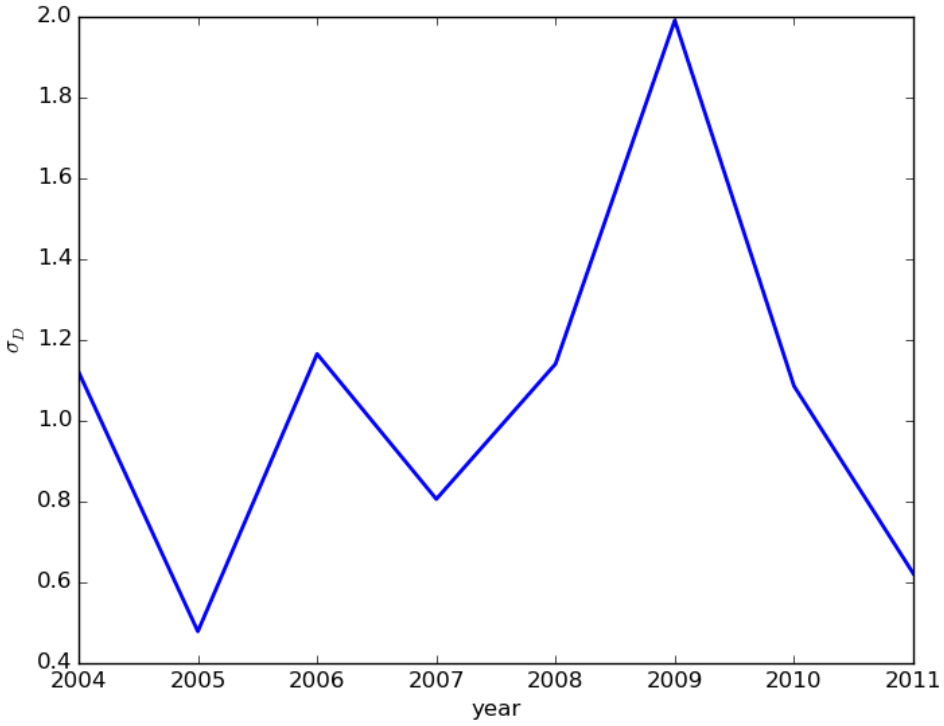


Figure 3.20: The standard deviation (σ) of the ratio between correlation coefficients calculated using simulated data and the correlation coefficient of the real data, which is plotted as a function of time. The stocks were simulated using a value reverting model with the an index term and an autoregressive term of the fourth order. The ratio was calculated using the stocks from NYSE that selection criteria, which were 622 stocks.

The analysis of correlation coefficients is continued by looking at the eigenvalues of the correlation matrices. The correlation matrices are calculated using the data from the simulation using eq. (3.5.22) as well as the real data.

In fig. (3.21) the normalized distribution of eigenvalues λ_i is plotted, where the fraction $1 - Q$ of eigenvalues equal to zero is not included. The eigenvalues were calculated for the stocks trading on NYSE during 2005. The eigenvalues λ_i $i = 5, ..., K - 1$ are included in the plot but the five largest eigenvalues λ_i $i = 0, ..., 4$ are not included in the plot. In addition to the eigenvalues in the plot, the theoretical distribution is given by eq. (2.3.38) included in the plot.

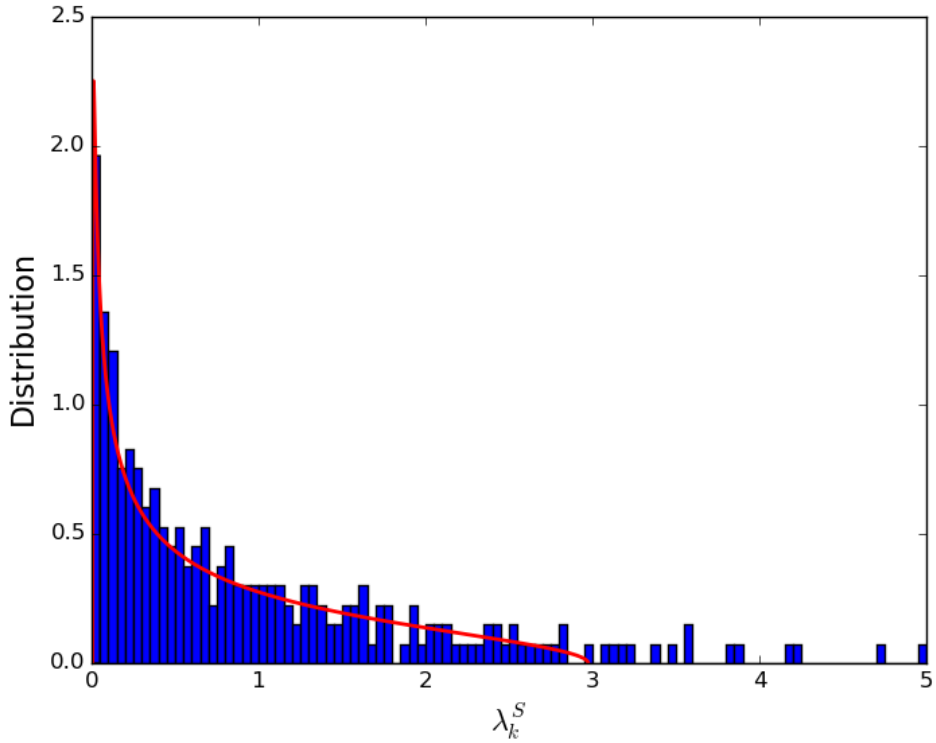


Figure 3.21: Distribution of the eigenvalues (λ_i) for the correlation matrix. Note that the largest eigenvalue λ_0 is not included. The correlation matrix is calculated with the 622 stocks that traded on NYSE during 2005, and that passed the selection criteria. The theoretical distribution of the bulk of eigenvalues given by eq. (2.3.38) is plotted in red.

The ratio $Q = T/K$ will be smaller than one for this data set as T is equivalent to the number of trading days of a year which are approximately 250 days. K is the number of companies that passed the selection criteria($K = 622$). Since $Q < 1$ there will be a lot of eigenvalues at or close to zero. Then choosing to focus on the three largest eigenvalues outside the bulk, λ_i $i = 0, 1, 2$ for further analysis.

λ_0 is the largest eigenvalue and corresponds to correlations caused by the market. λ_1 and λ_2 are the second and third largest eigenvalues and thus corresponds to the two largest sectors of the stock market. So by calculating the λ_i $i = 0, 1, 2$ on year by year basis, it is possible to see how they change with time. This is done for both the data simulated with eq. (3.5.22) and for the real stock price data. In fig. (3.22) λ_i $i = 0, 1, 2$ is plotted as a function of time; λ_i^S are the eigenvalues calculated using real stock price data and plotted as a line. λ_i^{VRM} are eigenvalues calculated for the stocks simulated using eq. (3.5.22) and are plotted as a dashed line.

The mean value of the ratio between the simulated correlation coefficients and the correlation coefficients calculated with real data ($\langle D_{ij} \rangle$) shows that there is a bad agreement

between the correlation coefficients for the simulated data and the real data. This is seen in fig. (3.19) where $\langle D_{ij} \rangle$ is plotted as a function of time and $\langle D_{ij} \rangle$ is not close to the expected value of one, which would indicate a good agreement.

The further analysis of the three largest eigenvalues from the correlation matrix also shows that the model does not capture the correlations caused by different sectors in the stock market. This can be seen by looking at the bad agreement of the second and third largest eigenvalue between for the real and simulated data in fig. (3.22).

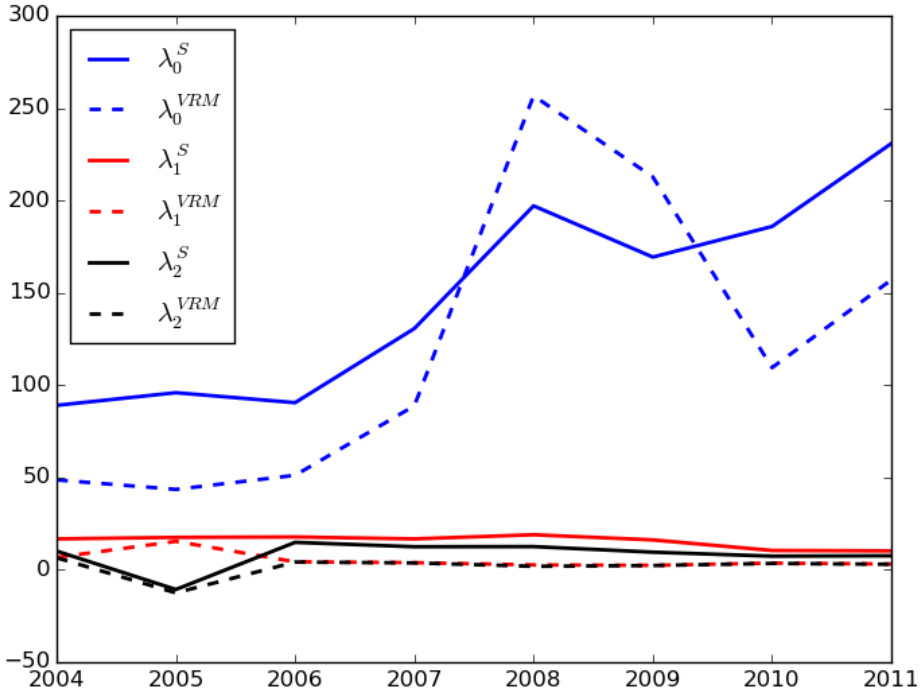


Figure 3.22: The tree largest eigenvalues (λ_i, $i = 0, 1, 2$) of the real and simulated correlation matrix plotted as a function of time. The eigenvalues calculated using real stock price are plotted as a full line and the eigenvalues calculated from simulated data are plotted as a dashed line. The stocks were simulated using a value reverting model with the an index term and an autoregressive term of the fourth order. 622 stocks from NYSE that passed the selection criteria were used in the calculations

4. Discussion

The stock market is a man-made system, that changes with time. So a model that describes the stock market perfectly may never be found. The big challenge is that there is only one real stock price path during a period for each company. So the stock price path of a company can be seen as a non repeatable experiment, as it is impossible to replicate all the initial conditions and repeat the initial public offering of the same company. Instead the stock path of multiple companies, all with different parameters and initial conditions have to be used to determine what is a good model of the stock price movements.

When looking at the simulation of the stock price for 'The Coca-Cola Company, (NYSE: KO)' using VRM and GBM which is shown in fig. (3.1), it is not possible to distinguish between the simulations and the real data. But this inability does not mean that both of the models are good at simulating the stock price. When looking at multiple simulations of VRM and GBM, using the same parameters as shown in fig. (3.2) and fig. (3.3). It is clear that VRM creates a narrow band of share prices compared to GBM that has a wider range of possible share prices.

This difference is due to the difference in how the drift term in the two models functions. In VRM, the drift term $-\alpha(S_t - V_t)dt$ acts as a spring and results in a stronger force the further share price deviates from V_t and thus creates a narrow of possible share prices. In contrast, GBM does not have a "corrective" action, but rather the drift term $\mu S_t dt$ simply adds to the share price independently of the size of the share price. The drift term for GBM gets bigger and crests a higher share price if $\mu > 0$ when S_t gets big, where VRM would create a larger downwards force if the share price saw an increase in S_t above V_t.

So by using fundamental data points of the underlying business, the simulation of the stock prices can be improved compared to GBM. But in order to utilize VRM effectively in applications, a fair estimate of the value is needed. A problem with this estimation of $V(t)$ is the assumption that the business will behave similarly when extrapolating $V(t)$ beyond the period used to estimate the parameters. It does hold true and is a good assumption in most cases. But several companies had to do big changes in their business after the 2008 crash. In those cases, the model did not work well to extrapolate $V(t)$ with the calculated parameters from 2002-2007 beyond 2008. Similar cases must exist in other time periods for some companies. This problem could potentially be addressed with better data points or more elaborate selection criteria, but this was not possible with data points used.

The estimation of value was done with fundamental data points that were reported annually. These annual data points were then used to calculate the value at one time during the year. To create a time series (V_t) with data points on all trading days, a line was fitted between the points calculated annually. A straight line was used instead of a step function, as companies report their earnings quarterly in addition to press releases, etc. Which means that there will in reality be multiple sources of information from the company

during a year, that need to be included in the stock price. So using a straight line between the annual data points is an attempt to use the data available as well as well as possible. Of course, a better dataset containing quarterly data would be preferable, but this could not be accessed.

The value reverting model can be used to select a stock that will increase in price. One way is to select a stock that is trading at a significantly lower level than the estimated value. Then relying on the property of reversion towards the value to see an appreciation in the stock price. This method is refereed to as value investing by investors. Another way is to predict the future value and to buy a stock that will have a greater value in the future than the current share price, and relying on the reversion towards value to create an appreciation of the stock price over time. This is refereed to as growth investing by investors as the future value is highly dependent on the growth of the company. So VRM is consistent with the investment frameworks of value investing and growth investing. While GBM is not, as the only way with GBM to pick a successful stock is to look at the drift parameter μ.

The presence of autocorrelations when looking at the stocks from an entire index could potentially also be used to profit in the stock market. The mean autocorrelation coefficient as a function of τ is shown in fig. (3.6), where it is shown that there are negative autocorrelations for small values of τ. The mean value of PMOD ($\langle P^k(\tau) \rangle_k$) as a function of τ and the distribution of P^k for fixed τ is similar to what is observed when studying the autocorrelation. The mean value of the autocorrelation as seen in fig. (3.6) compared to and mean value of PMOD in fig. (3.9) behave similarly in that they deviate from the expected value of 0 and .5 respectively for low values of τ. This similarity is to be expected as they look at a similar property, but where autocorrelation also captures the amplitude of the movements whereas PMOD only look at the direction of the movement. It is possible to conclude that autocorrelation is not negligible as the mean autocorrelation for NYSE and NASDAQ is negative for small values of tau.

The expected Gaussian behaviour of probability of movement in the opposite direction is not observed in the case $\tau = 1$ as seen in fig. (3.11). But it is observed for $\tau = 9$ as seen in fig. (3.11). This indicates that the assumption of the daily movement is independent on the previous days movement is wrong, but for larger values of τ it is a valid assumption. So when studying a group of stocks, the autocorrelation and PMOD are not negligible for the $\tau = 1, ..., 4$, this could be interpreted as the stock price have a "memory" of the last four trading days. This behaviour can be included in models of stock price movements by adding an autoregressive term.

When comparing the mean autocorrelation ($\langle R_k(\tau) \rangle_k$) for the real data in fig. (3.6) with the simulated data using GBM with an autoregressive term of the forth order in fig. (3.12). It can be seen that there is a similar behaviour for $\tau = 1, ..., 4$. But the mean autocorrelation has larger volatility in the real data when $\tau > 4$. This could be corrected by using a larger order of the autoregressive term. An order larger than five for the autoregressive term

would mean a memory of past movements larger than a week. Which would only model the behaviour without an adequate explanation of why it is present in the data. It would also be computationally expensive in the estimation of the autoregressive coefficients.

So the autoregressive term may not be the perfect description of the behaviour of autocorrelation and PMOD. However it provides a simple way of modifying models like VRM and GBM to include the autocorrelation behaviour for small values of τ. As seen in fig. (3.6) the mean autocorrelation will oscillate around zero for larger values of τ, so it will not have a noticeable effect on the simulations. So the autoregressive term is suitable to use given the fact that is simple to include in models.

When considering an index where each stock is simulated using VRM or GBM with no additional terms, which is seen in fig. (3.15). The desired index behaviour does not occur, but rather a smooth line. This is due to the fact that the stochastic part (dW_t^k) of the stochastic term $\sigma^k S_t^k dW_t^k$ will cancel each other out. This will result is a smaller volatility than observed for real data. The introduction of an index term which has the same stochastic variable (dW_t) for all the companies, allows for correlation between the companies to occur, which creates the desired index behaviour.

The addition of an autoregressive term to the model of GBM with the index term (I_{ArG}) have a noticeable effect compared to GBM with only the index term (I_G) as seen in fig. (3.16). I_G is larger than the real value (I_S), the addition of the autoregressive term creates a dampening effect that makes I_{ArG} closer to I_S. This can be due to the autoregressive terms adds additional parameters, which will create a better fitting. Or as seen when looking at the expectation value of the share price when the autoregressive term is added to GBM, it will act dampening when the drift is positive, and the autocorrelation correlation is negative.

However, the damping is not observed for the equally weighted index simulated using VRM with the index term (I_V) and with the additional autoregressive term (I_{ArV}). But as seen when looking at the simulations of KO, VRM provides a narrow range of the share price, and the additional parameters may not have as much influence on the result as it have on GBM. Which is why I_V and I_{ArV} is closer to the real data (I_S). In 2008 I_V and I_{ArV} started to decline before I_S, this is due to the extrapolation of V_t from annual point to daily points. The crash causes a decrease in earnings in 2009 which influences V_t during the period 2008-2009 which in turn creates a downward trend during this period, which could be improved with the use of fundamental data points reported quarterly.

The combination of an autoregressive term and an index term to VRM gives a similar distribution of correlation coefficients as for the real data which can be seen in fig. (3.18) and fig. (3.17). But this does not mean that the simulated correlation coefficients agree with the real correlation coefficient.

The ratio of correlations coefficients for the stocks i and j for the real data (ρ_{ij}) and the simulated data (ρ_{ij}^{sim}) given by $D_{ij} = \rho_{ij}^{sim}/\rho_{ij}$ as a function of time as shown in fig. (3.19), which shows that there is a bad agreement between the two. The agreement in

the distribution of ρ_{ij}^{sim} in fig. (3.18) and ρ_{ij} in fig. (3.17) can be due to the fact that the parameters for the model are fitted during the entire period and it is this period that is used in when looking at the distribution of the correlation coefficients. While D_{ij} is analysed on a year by year basis,. The bad agreement of D_{ij} could be improved if the parameters used in the simulations were time-dependent and for example estimated on a year by year basis instead of estimated during the whole period.

However, when looking at the largest eigenvalues of the correlation matrices for the real data and simulated data in fig. (3.21), there is almost no agreement for λ_1 and λ_2, where λ_1 and λ_2 represent the two largest sectors. But there is some agreement for λ_0, which represents the market. The lack of agreement is a result from how dI_t in the index term is calculated. It treats all the sectors and the whole market as one variable. The different sectors and the market should be broken down into unique variables and treated separately, in order to achieve a better model. This could be solved by expanding the index term

$$\sigma_I^k S_t^k dW_t \tag{4.0.1}$$

in such a way that it have additional terms that correspond to different sectors. Such as

$$\sum_{m=1}^{M} \sigma_m^k S_t^k dW_t^m \tag{4.0.2}$$

where M is the number of sectors as well as the market as a whole, m is a sector or market that corresponds to the eigenvalues outside of the bulk.

The inclusion of sectors in the model could provide a more accurate model of how the stock market functions. But it does not provide a way in predicting how the market and sectors will move each day as this is modelled as a random process, which means that it does not provide a framework to simulate individual stocks into the future with any precision. So the best option in predicting future stock prices is by using VRM, with the possible inclusion of an autoregressive term. The value reverting term provides a framework to predict the direction of how the stock price will move, but it introduces the need of predicting the future value.

5. Further developments

The theoretical framework of VRM can be expanded further, to the point where it is possible to to find the theoretical band of possible stock prices. This can be done by looking at the expectation value of the VRM process and looking at the theoretical variance of the process. When looking at how VRM could be used, the limiting factor was the size of the data set with fundamental data points that were used in the estimation of V_t. It would be interesting to use a larger data set both of more stocks and over a longer period. This would allow for a more detailed analysis of the limits of how VRM can be used as well as comparing different time periods to each other.

This could involve using fundamental data points reported quarterly instead of annually and would make it possible to use a more detailed valuation model with four times as many data points during the same period. More detail data points reported quarterly would also make it possible to drop or alter the assumption of value (V_t) being a continuous function. By letting $V(t)$ being non-continuous and making the parameter of reversion towards the equilibrium (α) in VRM be time dependent, could provide a framework to describe jumps in stock prices. By looking at a jump in V_t in combination with an increase in α during a short period.

The analysis of PMOD and autocorrelations could be expanded by looking at how they change over time. This could entail comparing different time intervals as well as comparing stocks from different indices. It is also well established that dS_t^k are distributed with fat tails, but in this thesis a normal distribution is used in the simulations instead of a distribution with fat tails. So the models used in this thesis can be improved upon by introducing distributions with fat tails.

The index term $\sigma_I^k S_t^k dW_t$ can also be expanded in such a way that it has additional terms that correspond to different sectors as outlined in eq. (4.0.2). This could improve the agreement of the correlation coefficient for simulated and real stock price data.

6. Conclusions

VRM introduces a different type of drift term compared to GBM, that is more accurate in the simulation of stock price movements and relates the drift of stock prices to the value given by the fundamental data points of the company. Compared to GBM, that simply observes the existence of drift in the share price and introduces a term to model it, without providing a mechanism that describes why the drift exits. So VRM provides a better prediction of the future share price compared to GBM but also introduces new difficulties in the calculation V_t, such as the need for fundamental data points of the company. VRM is not the perfect model, but fundamental data points of the company should be included in the model in order to achieve a accurate model of the stock market.

As an initial model, GBM behaves similar to a stock price and is a functional first approximation of stock movements. GBM can easily be used to extrapolate into the future by using the estimated parameters. However just because it is easy to use this does not mean that GBM has a high degree of accuracy in the prediction.

When looking at a single stock both GBM and VRM can provide a good model of stock movements. But when considering a group of stocks and how they create an index VRM and GBM are no longer adequate, as GBM and VRM do not create an index behaviour but rather a smooth line. By introducing an index term, it is possible to simulate an index that behaves similar to actual data.

It is commonly thought that the autocorrelations of stocks are negligible and can be ignored. But when looking at the mean value for the autocorrelation coefficient for a group of stocks, it is clear that autocorrelation actually is present and should not be ignored. It is possible to include the behavior of autocorrelation in stochastic models using an autoregressive term.

Bibliography

[1] Gunnar Blom. *Sannolikhetsteori och statistikteori med tillämpningar*. Studentlitteratur, 2005.

[2] Eugene F. Brigham and Louis C. Gapenski. *Financial Management: Theory and Practice*. The Dryden Press series in finance. Dryden Press, 1997.

[3] Joseph L. Doob. The Brownian Movement and Stochastic Equations. *Annals of Mathematics*, 43:351–369, 1942.

[4] Thomas Guhr. *Econophysics*. Matematisk Fysik, LTH, Lunds Universitet, 2007.

[5] Margaret Insley and Kimberly Rollins. On solving the multirotational timber harvesting problem with stochastic prices: A linear complementarity formulation. *American Journal of Agricultural Economics*, 87(3):735–755, 2005.

[6] Laurent Laloux, Pierre Cizeau, Marc Potters, and Jean-Philippe Bouchaud. Random matrix theory and financial correlations. *International Journal of Theoretical and Applied Finance*, 03(03):391–397, 2000.

[7] World Federation of Exchanges. Wfe statistics, monthly-reports, 2015.

[8] John A. Rice. *Mathematical Statistics and Data Analysis*. Cengage Learning, 2006.

[9] Sheldon M. Ross. *An Elementary Introduction to Mathematical Finance: Options and Other Topics*. Cambridge University Press, 2003.

[10] Florian Steiger. The validity of company valuation using discounted cash flow methods. Papers, arXiv.org, 2010.

[11] Petre Stoica and Randolph L. Moses. *Introduction to spectral analysis*, volume 1. Prentice hall Upper Saddle River, 1997.