NTNU
Norwegian University of
Science and Technology

# A random Matrix Approach to collective Trends of falling and rising Stock Markets

**Christoffer Berge Hansen**

Master of Science in Physics and Mathematics
Submission date: June 2011
Supervisor: Ingve Simonsen, IFY

Norwegian University of Science and Technology
Department of Physics

NTNU
Faculty of Natural Sciences
and Technology
Department of Physics

# Master's Thesis for

**Christoffer Berge Hansen**

Master of Science in Technology, Applied Physics and Mathematics,
Study Direction Technical Physics

*A random Matrix Approach to collective Trends of falling and rising Stock Markets*

This work has been carried out at the Department of Physics, NTNU,
under the supervision of Ingve Simonsen.

Trondheim, 27.06.2011

**Abstract**

An inverse statistics analysis of one minute stock quotes from 492 large European companies has revealed the existence of a gain-loss asymmetry in the following index. The gain-loss asymmetry differs from that observed for daily closure prices of the Dow Jones Industrial Average [38], as the probability of the optimal investment horizon for a gain is *higher* than that of a loss. For individual stocks, the gain-loss asymmetry was observed to *only* appear for significantly larger return-levels. To the best of our knowledge, this is the first time such an analysis has been performed on high-frequency data.

A principal component analysis was done by performing an eigenvalue decomposition of the correlation matrix from a sliding time-window. The first principal component was observed to describe the market excellently. Its corresponding eigenvalue was observed to be significantly larger than theoretical predictions from random matrix theory, implying that the eigenvalue carries information common to all stocks. Using this eigenvalue as an index measuring the collectivity in the market has revealed the existence of collective trends that appear to be *stronger* during falling than rising markets. This has been observed for two different datasets, the above described one minute stock quotes and daily closure prices from 29 stocks composing the DJIA late February 2008. The observation is in accordance with results of Balogh et al. [40], and provides further support to the speculation of Johansen et al. [37] that a difference in collective trends is the reason behind the gain-loss asymmetry observed in indexes and not for individual stocks for the same return-level.

The key idea behind the fear factor model of Donangelo et al. [42] has been strongly supported by the observation that collective trends appear to be stronger during sharp index drops. As the collectivity increment has been observed to be dependent on the size of the index drop, it is suggested that the model should incorporate also *individual* fear factors for economic sectors, in addition to the global fear factor governing the market as a whole. Periods exhibiting a rising index *positively* correlated to the strength of collectivity has indicated the presence of an *optimism factor* that also should be incorporated in the fear factor model [42], forcing stocks to rise synchronously.

# Contents

# List of Figures

# Preface

This thesis is the final work of my Master's Degree in Applied Physics at the Norwegian University of Science and Technology (NTNU). The reason why I started at NTNU in the first place was my interest for physics and mathematics, as well as the fact that NTNU is a good scientific university in Norway with possibilities of taking parts of the education abroad. After studying for some years, finance attracted a lot of my interest, and especially the modeling of stocks and financial markets in general. The autumn of 2008, stock exchanges exhibited synchronized falls all over the world, and what we today know as the financial crisis had started. The index representing companies noted at the Oslo stock exchange fell from a value above 500 in September 2008 to a value of less than 200 March 2009, where the worst intraday return was 8.30%! According to the primitive model often used in finance to model stock fluctuations, such an event is impossible. This is also why such events have got the name *black swans*. To model stocks as well as trying to understand and maybe predict the behavior of individual stocks and more complex markets is therefore very interesting, especially from a physicist's point of view.

Working on my specialization project last term as a member of the soft and complex matter group at NTNU gave me a deep insight into self-assembling nanoparticles made up of clay platelets. However, after working with this for a while, I was ready to do something new. Even if the analysis of financial data is not very different from that of turbulence, I did not know that there were people working with this also at the Department of Physics at NTNU. Two doors down the corridor from where I mostly worked on my project, my former professor in Electromagnetic Theory had his office. The last month before Christmas, he introduced me to the field of econophysics, as he was a member of a group working on what is known as *inverse statistics* among other things. They observed, based on empirical data from the Dow Jones Industrial Average [38], that the expected time to achieve a gain on your investment is longer than that of a loss, an effect speculated to arise from a collective trend in the financial markets. In May 2011, one of their models

[42] to understand this behavior led to an article in the Danish version of Financial Times. I guess my former professor must have understood that I found this very interesting, as he agreed to be my supervisor when I asked after having handed in my project.

While working on my master's thesis, I realized that I had to learn a lot of finance to get a deeper understanding of the models used today. I have only had one course connected to finance, but thanks to several mathematics courses at NTNU this was not as hard as expected. In addition, I had to do a lot of programming, and chose to use Python for this purpose. Python is very intuitive to use (especially after having had C++ a few years ago), and is in addition free, compared to e.g. MatLAB that is extremely expensive when used outside universities. It follows that my learning curve has been very steep, but both my educational and personal outcome have been great.

It is no surprise that I want to thank my supervisor Ingve Simonsen for providing me with the possibility of writing this thesis, as well as helping me whenever I had questions. Even though Simonsen is one of the busiest professors at NTNU, he always has time and always makes sure that you understand. In addition, I would not been able to write this thesis if not his earlier student, Peter Ahlgren, had provided me with the necessary data. Finally, I must thank my fellow students Beate Cappelen and Hege Knutsen for providing a positive learning environment with yatzi breaks whenever needed.

Trondheim, 27.06.2011 Christoffer Berge Hansen

# Chapter 1

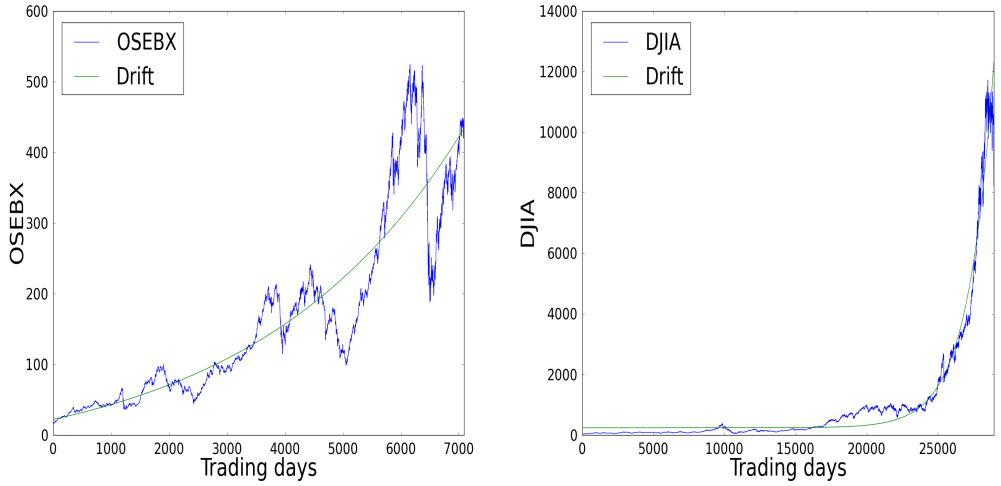
# Introduction

Financial markets allow people and institutions to buy or sell financial securities such as stocks and bonds, but also commodities and other fungible items. Such markets are very important, as they allow buyers and sellers to find each other. An idea needs capital, and someone with an idea may find the needed capital in the financial markets. Countries, companies and individuals invest money in the financial markets which are believed to have a close to exponential drift, resulting in a larger gain as compared to placing money in the close to risk free bank. It is a question of risk versus return. The drifts of two well known indexes are presented in figure 1.1, where the historic daily closure prices of the Oslo Stock Exchange Benchmark Index and the Dow Jones Industrial Average are shown with a fit to an exponential function.

Financial time series are interesting as they have been recorded and studied for many decades. The appearance of computers caused an acceleration of this development, and today large amounts of high frequency data are recorded daily. In 1900, Bachelier [13] suggested that drift-corrected stock fluctuations behave as a Brownian motion. Einstein published his first and famous article on Brownian motion [10], independently of Bachelier, 5 years later. The fact that both movements of Brownian particles and movements of stocks were (and still are) modeled this way[1] proves that the fields of finance and physics are deeply connected. Financial data have a relatively high frequency of events termed *black swans*[2], or in other words surprising events with major impact such as the financial crisis starting the autumn of 2008. Such events are not predicted by any random walk processes and have

[1]Note that the common model now for stocks is not standard Brownian motion, but rather the model known as geometric Brownian motion.

[2]More about black swans can be read in *The Black Swan* [3], a New York Times bestseller.




(a) (b)

Figure 1.1: (a) Blue line: Historic daily closure prices of the Oslo Stock Exchange Benchmark Index (OSEBX) over the period January 3rd, 1983 to March 29th, 2011. (b) Blue line: Historic daily closure prices of the Dow Jones Industrial Average (DJIA) over the period May 26th, 1896 to June 5th, 2001. The green lines are fits of an exponential function of the form $f(x) = a + b\ \exp(cx)$ to the empirical data, and just an illustration of the drift in stock markets. The OSEBX data and the DJIA data are obtained from [2] and [21] respectively, and the resulting fit parameters can be found in appendix B.1, table B.2.

caused stock dynamics to be compared with turbulence, exhibiting similar behavior with unpredictable spikes [5]. In 1973, the economist Scholes and the physicist Black together published their famous formula for pricing of options and derivatives [19]. In 1997, they were awarded the Nobel Prize in Economics[3] [1] for this formula, as it still is indispensable. A correct pricing of derivatives is important, and this is still done using the Black and Scholes formula or other physics-related models, even though it is clear that many of the assumptions they are based on are incorrect. The conclusion is that banks do not only need economists, but also physicists and mathematicians — which are nicknamed *quants* by many economists. One of the aspects of financial markets to be investigated in this thesis is whether there is a significant difference in the collective trends of share prices during stock index rising and falling periods. This has been indicated by findings of Balogh et

[3]Even though Black himself died 2 years earlier.

al. [40], but will be approached differently in this thesis. An understanding of such trends is important for portfolio management, as the difference in collective trends implies that risks are not symmetric measures for positive and negative returns.

The management of risks is an important aspect of finance. The optimal portfolio maximizes the return for a given level of risk. This is also a field where so-called *quants* are important, modeling future developments of interest rates or prices of stocks and commodities. Even currency developments are modeled, as these play a significant role for exporters, importers and individual investors. By use of empirical data, models trying to predict the future can be constructed. Even though no such models can predict future prices with certainty, they can be used to minimize the risk of losses. Physicists and mathematicians are well trained on doing research, and this is a field where their analytical skills are needed.

One of the techniques used in this thesis is called *random matrix theory*, and was actually developed by Wigner [47] in 1955 to describe the energy levels of atomic nuclei and their fluctuations. The theory is important as it is can be used for cleaning random correlations from empirical correlation matrices [25], which are among the corner stones of todays risk management. This enables the possibility of extracting true correlations between financial assets (and not the purely random correlations arising between stocks from time to time), which again can be used when building the optimal portfolio [24]. Due to the great importance of such models, the field of financial mathematics is an important one, growing at a high rate and employing many physicists.

# Chapter 2

# Background

Mathematical finance is often said to start with the publication of Louis Bachelier's doctoral thesis *Thèorie de la Spèculation* in 1900 [13]. Bachelier formulates several important hypothesis about the market, where especially one is of great importance: *"At a given instant, the market believes neither in a rise nor in a fall of the true price"*. Equivalently, *"the expectation of the speculator is zero"*, and thus Bachelier implicitly assumes that the market evaluates assets using a martingale measure (a zero drift stochastic process). Bachelier tried to model the drift-corrected asset prices in a mathematical fashion, and through three different approaches he concluded that the price changes of assets are Gaussian distributed and that the price follows a Brownian motion. As this in principle opens for a non-zero possibility of negative prices, the model has later been replaced with what is now known as the geometric Brownian motion model.

Geometric Brownian motion is a process where the logarithm of the randomly varying quantity follows a Brownian motion. In this case, it is the asset price that is randomly varying, leading the logarithmic return[1] to follow a Brownian motion. It follows that the problem of a non-zero probability for negative asset prices vanishes. The model was used by Black and Scholes when deriving their famous formula for option pricing [19], and is still the standard assumption in finance. However, due to its underestimation of large fluctuations causing so-called fat tails in empirical financial data [9], it is an open question of which probability distribution that describe the empirical data best.

[1]The relative price increment, or the arithmetic return $\eta_k$ is defined $\eta_k = \Delta S_k/S_k = (S_{k+1} - S_k)/S_k$, where $S_k$ is the stock price at time $t_k$. Mostly in this work, the considered return is the so-called logarithmic return. The logarithmic return $r_k$ is defined $r_k = \ln(S_{k+1}/S_k) = \ln(\eta_k + 1)$, and $r_k \approx \eta_k$ for $\eta_k << 1$.

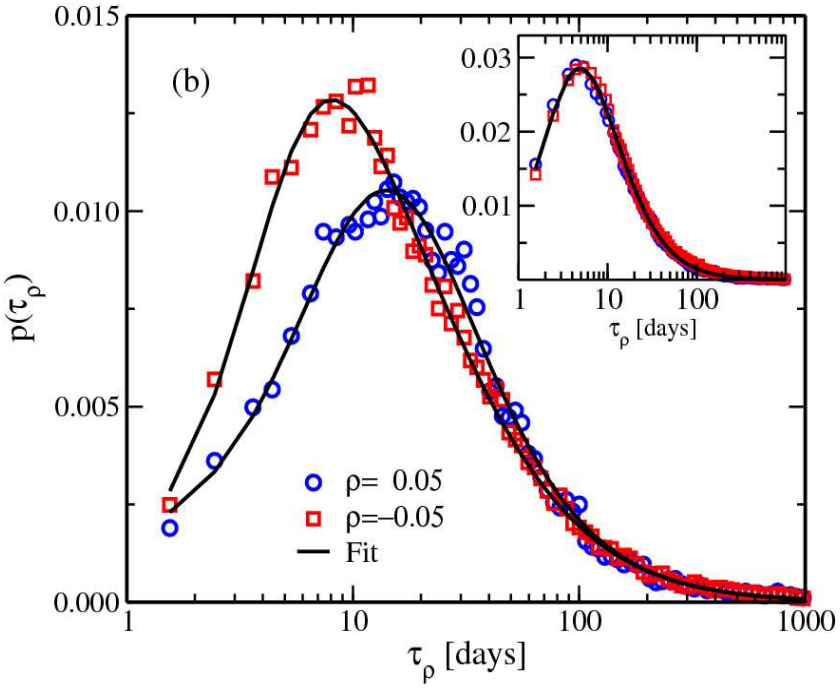



Figure 2.1: Inverse statistics distributions for the Dow Jones Industrial Average index for a return-level $|\rho| = 0.05$, based on empirical daily closure prices from 1896 to 2006. Red open squares correspond to negative returns and blue open circles to positive returns. Both distributions are normalized. The inset shows the gain and loss distributions obtained from averaging over the inverse statistics distributions of each of the constituent stocks of the index. It is observed a clear asymmetry in the index but not in the individual stocks. Figure adapted from [42].

In finance, the classical approach when studying time series has been what was denoted by Ahlgren et al. [34] as *forward statistics*. In this approach, a typical time-scale is chosen, and the typical return over the chosen time-scale is calculated based on historical data. What is now known as *inverse statistics* was introduced in 2002, when Simonsen et al. [36] asked the "inverse" question: *"What is the typical time span needed to generate a fluctuation or movement (in the price) of a given size?"*. The result was the observation of a *gain-loss asymmetry* in stock indexes, and that this asymmetry was small or not present in the individual stocks indexes are composed of [37, 38]. However, later research has concluded that also individual stocks exhibit a small gain-loss asymmetry, but that this only appears for a significantly larger return-levels [50]. An example of the gain-loss asymmetry is seen in figure 2.1, where the inverse statistics distributions calculated by

Donangelo et al. [42] are presented. The peaks of the distributions were coined *optimal investment horizons* by Simonsen et al. [36], and for mature and liquid western markets it has been observed that the optimal investment horizon always occur earlier for negative returns. This is also observed in figure 2.1. Another interesting observation from figure 2.1 is that there is a higher probability to find short investment horizons for negative return-levels compared to positive return-levels [37].

Johansen et al. [37] speculated that the origin of the gain-loss asymmetry was some kind of collective movement of the stocks of the index. This was the key idea behind the *fear factor* model by Donangelo et al. [42], constructed to explain the origin of the asymmetry. The model features short periods of synchronized dropping prices induced by a *fear factor*, and was found to reproduce the asymmetry in the index except from some minor differences especially for short waiting times. By construction, the model produced no asymmetry for the individual stocks, and the results are further discussed in section 3.8.2. However, the model caused the authors to speculate in the existence of an *optimism factor*, introducing synchronized upturns similar to the downturns introduced by the fear factor. After the introduction of the fear factor model, Balogh et al. [40] found empirical evidence indicating that collective trends during falling markets are stronger than during rising markets, supporting the fear factor hypothesis. It has also been speculated that the negative correlation between past returns and future volatility, known as the leverage effect, could be of the same origin as the gain-loss asymmetry [50, 51].

The strength of collective trends in the market is investigated in this thesis, using *principal component analysis* and *random matrix theory*. As the theory of random matrices gives a prediction for the eigenvalue spectrum of a very large random matrix, random and uncorrelated fluctuations between stocks this way can be excluded. This again leads to the possibility of extracting real information from financial correlation matrices.

# Chapter 3

# Theory

## 3.1 Econophysics

The term Econophysics was coined by the physicist Stanley at the second Statphys - Kolkata conference in Calcutta, India, in 1995 [4, pp. 225]. Mantegna and Stanley later defined the field of econophysics as *"a neologism that denotes the activities of physicists who are working on economic problems to test a variety of new conceptual approaches deriving from the physical sciences"* [5, pp. viii]. This definition is sociological, as it is not based on the type of problem or methods and theories for solving them, but rather who is doing the work.

The mathematician Bachelier [13] tried in the early 1900's to explain the fluctuations of commodity prices over time, using statistical physics. His explanation led to an introduction of the theory of random walks, a theory which was later developed independently by Einstein in his article titled *"On the Movement of Small Particles Suspended in a Stationary Liquid Demanded by the Molecular-Kinetic Theory of Heat"* published in 1905 [10]. Bachelier originally proposed that price changes were Gaussian distributed, which became the standard way of modeling asset prices for several decades. This suggestion was later replaced by the geometric Brownian motion model, where it is logarithmic price changes that are Gaussian distributed. This is the model Black and Scholes used for asset prices when deriving their famous formula for the pricing of options and derivatives [19]. Discrepancies between empirical data and the Black and Scholes formula due to its underlying assumptions have been observed, and it is known that the geometric Brownian motion model for stock prices is not correct [17, pp. 104 - 105]. A main point is the assumption of log-normal distributed returns, causing an underestimation of the actual probability of rare events as compared to

empirical data [15]. However, both the geometric Brownian motion model and the Black and Scholes formula are still widely used. This follows, as they are easy to work with and in general give useful approximations.

A clear contrast between the approach of econophysicists and economists lies in the description of many empirically observed financial distributions. Physicists have found that many phenomena can be more accurately described using scaling laws[1]. Scaling laws describe the empirically observed distributions exhibiting skewness and leptokurtosis, something the Gaussian distributions suggested by Bachelier in 1900 do not [4, pp. 227]. The contrasting approach using scaling laws was first performed by Pareto in 1897 [14, pp. 119], where it was used to study income distributions. Mandelbrot [9] applied a similar approach to study the fluctuations of cotton prices in 1963, and observed more and larger fluctuations than expected from a Gaussian process.

Other problems studied by econophysicists are the distributions of income and wealth, inverse statistics including gain-loss asymmetry (which will be discussed in section 3.8.2) and modeling of highly volatile and seasonal markets such as the electricity markets, economic shocks and growth rate variations, company sizes and growth rates, scientific discoveries, etc [7]. Much of this work has been published in several journals from the physical ones such as the *Journal of Modern Physics C*, *Physica A*, *Physical Review E* and *European Physical Journal B* to more general scientific journals as *Nature* and multidisciplinary journals such as the *Quantitative Finance*. Some work is also published jointly with economists in economic journals [7].

## 3.2 Thèorie de la Spèculation

Louis Bachelier defended his doctoral thesis *"Thèorie de la Spèculation"*[2] in 1900. The thesis was also published in *Annuales Scientifiques de l'Ecole Normale Supèrieure*, and was a pioneering analysis on the dynamics of the stock and option markets [13, 18]. The aim of the thesis was to derive an expression for the probability of a price fluctuation of a commodity or a market some time in the future, given the current price [13]. As briefly mentioned in section 3.1, Bachelier made the first formulation of a theory of a random walk process, five years before Einstein independently developed the theory of Brownian motion when studying the physics of Brownian particles [10].

[1] Law that states that two quantities are proportional and known to be valid at certain orders of magnitude.

[2] Or in english, "Theory of Speculation".

After considering several financial instruments including futures, options and combinations of these, Bachelier formulated a series of assumptions. Of significant importance is the postulate: *"at a given instant, the market believes neither in a rise nor in a fall of the true price"* [13]. This is equivalent to stating that at any given time, the market is neither *"bullish"* nor *"bearish"* even though individual traders on the other hand may have their opinion on the direction of the market movement. The important hypothesis stated by Bachelier are [16, pp. 29]:

- The successive price movements are statistically independent.
- All information available from the past to the present time is completely accounted for by the present price in a perfect market.
- The same hypothesis is made in an efficient market, but small irregularities are allowed as long as they are smaller than the transaction costs, thus leaving no arbitrage[3] possibilities.
- In a complete market there are both buyers and sellers at any quoted price, having opposite opinions about future price movements. Thus on average, the market does not believe in a net movement.

What is remarkable, is that these assumptions are still among the standard assumptions in modern theory of financial markets.

### 3.2.1 Government bonds with contangoes and their futures

Bachelier basic ideas are formulated on a future-like instrument based on a French government bond with nominal value $S(0) = \text{Fr } 100$ and a 3% interest rate[4]. Every three months, a coupon of $Z = \text{Fr } 0.75$ is detached from the bond, representing the interest on the money invested if the bond is bought. The expiry date of the future is the last trading date in the month, but it can be extended beyond its maturity by paying a contango[5]. The coupon of Fr 0.75 per quarter equals Fr 0.25 monthly, and the contango is normally less than this value. The difference favors the buyer, suggesting the idea of buying futures to carry them forward indefinitely [13].

[3] By arbitrage, one means the possibility of making riskless profit [6]

[4] Fr is shorthand for Franc, the currency in France at the time of Bachelier.

[5] To conserve the position until the next maturity, the buyer of the future can pay the seller an indemnity denoted as a contango [13].

Characteristics of the future instrument are:

- The instrument does not include the obligation of delivering the bond at maturity, unlike modern bond futures. Instead, the price difference of the underlying bond is settled in cash.
- The expiration date is set on the last trading day of the month.
- The instrument can be extended beyond its maturity to the end of the following month by paying a contango K.
- The long position (the buyer of the future) receives the coupons from the instrument.

To consider the regular part of the price of Bachelier's futures, it is assumed that the the market is free of fluctuations. This causes the price of the future $F(t)$ to be completely governed by the price movements of the underlying security $S(t)$. With a 3% interest and nominal value Fr100, the price of the bond will increase linearly in time due to the accumulation of interest until its maturity every three months. At maturity, the price of the bond decreases instantaneously by the amount $Z$ before it starts to grow linearly as the interest again is accumulated.

The price of the future will have more dramatic price changes. It is of zero value immediately after each maturity, but during the three months between the maturities the value increases by an amount $Z$ due to the accumulation of interest. The slope of the future-price is determined by the size of the contango. In the special case where the contango is $Z/3$, the slope is zero and the price increases with an amount equal to the contango every month. The deterministic price of the future, assuming no fluctuations, is illustrated in figure 3.1 for three different contangoes.

Let $t_i - t_{i-1} = 3$ months, where $i$ is an integer. As the coupon is detached from the bond at times $t_i$, the price of the future $F(t)$ and its underlying bond $S(t)$ (assuming $K = 0$) is described by [16, pp. 30]

$$S(t) = S(t_i) + Z\frac{t - t_i}{t_{i+1} - t_i}, \qquad F(t) = S(t) - S(t_i) = Z\frac{t - t_i}{t_{i+1} - t_i}, \tag{3.1}$$

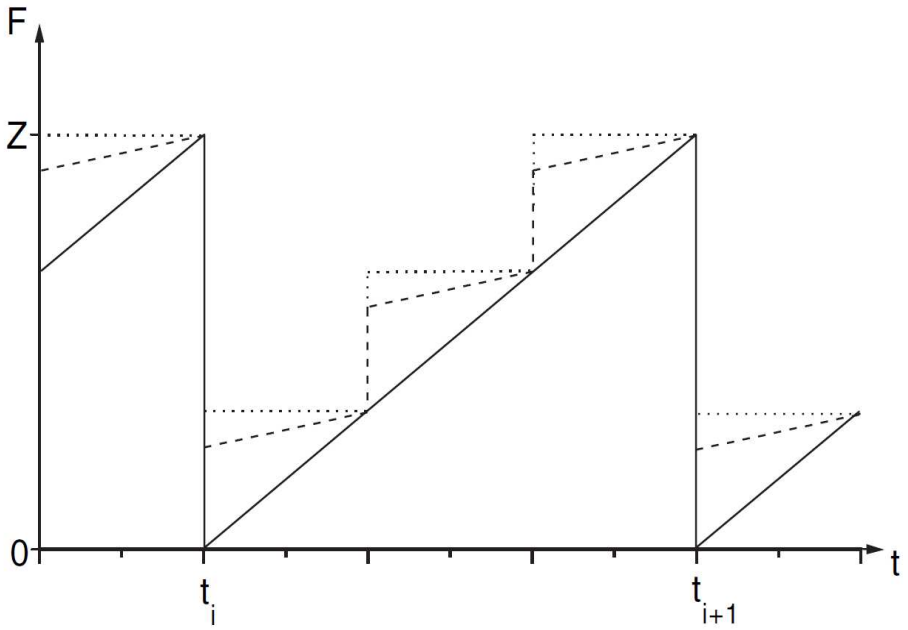



Figure 3.1: Deterministic price of the future, $F(t)$, based on the French governmental bond. This is shown for three different contangoes: $K = 0$ (solid line), $K = Z/3$ (dotted line) and $0 < K < Z/3$ (dashed line). Figure adapted from [16, pp. 31].

assuming no maturities during $t$. As the price movements of the future decrease when the contango increase, the drift is removed in the special case of a contango equal to $K = Z/3$. This is also the case examined by Bachelier [13]. With no drift, the gap between the contango and the monthly coupon favoring the buyer is zero. An important observation stated by Bachelier is that all prices on the different lines of figure 3.1 are equivalent, as the return an investor gets from buying the future at any given time is equivalent. This follows, as the slope is independent of time in the case of a deterministic price evolution. Taking fluctuations into account causes the gap between the future and the bond price to not behave deterministic, and the lines of figure 3.1 are not necessarily straight anymore. It follows that in reality, the return from investing in a future or bond is not the same at all times due to the unpredictable price fluctuations [13].

With a contango $K < Z/3$, the drift of the price of the future between two maturity dates is given by [16, pp. 31]

$$\frac{dF(t)}{dt} = \frac{dS(t)}{dt} - \frac{3K}{t_{i+1} - t_i}. \tag{3.2}$$

As there are no maturities during $t$, the true price of the futures a period $t + T$ from now is described by [16, pp. 31]

$$\tilde{F}(t + T) = F(t) + \frac{dF(t)}{dt}T. \tag{3.3}$$

Allowing fluctuations changes the price slightly, and it follows that there is no guarantee that the quoted price at time $t + T$ will correspond to the predicted price $\tilde{F}(t + T)$.

### 3.2.2 Probabilities in transactions in the stock market

In his work, Bachelier distinguishes between two kinds of probabilities [13]:

- A mathematical probability that can be determined *a priori*, corresponding to the probability studied in games of chance.
- A speculative probability depending on future events (perhaps more appropriately termed as an expectation), and as a consequence is impossible to predict mathematically.

The second probability is what the speculator[6] wants to predict. As the market neither should believe in a rise nor in a fall of the true price, the set of speculators must not believe that the price will drift upwards or downwards. It follows that there must be an equal number of buyers and sellers, and that this probability is a subjective opinion as the speculators counterpart must have the opposite opinion to ensure a complete market. However, drifts have been observed in most markets, and generate net positive expectations for future movements (see figure 1.1). It follows that Bachelier's basic hypothesis that the market is neither bullish nor bearish at any given time must be slightly modified. The modified statement has been formulated *"up to the drift, the market does not expect a net change of the true, or fundamental, prices"* [16, pp. 32].

Including fluctuations and drifts leads to deviations from the deterministic price. If deviations of amplitude $y(t)$ occur with probability $p(y)$, the expected profit from the investment is given by [16, pp. 32]

$$E(y) = \int_{-\infty}^{\infty} y\, p(y) dy. \tag{3.4}$$

[6]Speculator is just a synonym for the buyer or investor.

This quantity is non-zero and positive if the following conditions are fulfilled:

- The drift is non-zero: $\frac{dS}{dt}, \frac{dF}{dt} > 0$
- The contango is set to $K < \frac{Z}{3}$

In that case, the investment is not a fair game. This follows, as the expectation value of the profit is larger than zero, while a fair game of chance fulfills the condition

$$E(y) = 0. \tag{3.5}$$

Bachelier examines the special case with a contango set to the specific value $Z/3$, as discussed in section 3.2.1. This eliminates the drift, such that the expectations of the buyer and seller are zero [13]. In this case, the investment indeed is a fair game of chance. If the contango is not set to this particular value, the underlying drift must be accounted for. This can be done by rewriting the price of the future and its underlying bond by subtracting the drift,

$$x(t) = y(t) - \frac{dS}{dt}t, \qquad x(t) = y(t) - \frac{dF}{dt}t. \tag{3.6}$$

Here $x(t)$ corresponds to the drift corrected price of the instrument, $y(t)$ is the price with no drift correction and the last term corresponds to the drift. This transformation leads to zero expectation value of the profit, and the investment fulfills the condition of a fair game of chance. In other words, the excess profit for the speculator vanishes. As the price $x(t)$ describes a drift free time series, it is denoted as a martingale stochastic process. In discrete time, this corresponds to

$$E(x_{t+1} - x_t | x_t, x_{t-1}, ..., x_0) = 0, \tag{3.7}$$

or in other words that the expectation value formed with the conditional probability conditioned on the earlier observations equals zero.

### 3.2.3 The probability law

Bachelier used three different techniques to determine the distribution of probabilities of price changes, and all three techniques led him to the same distribution. He assumed that the price itself followed a martingale process, or in other words that drifts were accounted for. In this case, the probability distribution of price changes must be symmetric with respect to $x = 0$. It must also decrease sufficiently fast to zero to avoid negative prices of the assets and to make the distribution normalizable. This is a weakness of his

model, in principle allowing negative prices, and one of the reasons for the introduction of the geometric Brownian motion model.

The first method used to derive the probability distribution led Bachelier to the Chapman-Kolmogorov-Smoluchowski equation. Bachelier solved the equation, but never recognized that his solution was not unique. The second method led him to the first formulation of a theory of random walk, five years prior to Einstein. The last method led Bachelier to the diffusion equation, again giving the same result when using specific boundary conditions. The three following derivations of the probability distribution of price changes are based on his doctoral thesis [13].

### The Chapman-Kolmogorov-Smoluchowski equation

Assume that drifts are accounted for in the price of an asset $S(t)$. Denote by $p(x_1, t_1)dx_1$ the probability of a price change $x_1 \leq x \leq x_1+dx_1$ at time $t_1$ and $p(x_2-x_1, t_2)dx_2$ as the probability of a price change $x_2-x_1 \leq x \leq x_2-x_1+dx_2$ at time $t_1 + t_2$.

The joint probability of having a price change to $x_1$ at $t_1$ and to $x_2$ at $t_1 + t_2$ is then $p(x_1, t_1)p(x_2 - x_1, t_2)dx_1dx_2$. The probability of having a change to $x_2$ at time $t_1 + t_2$, independent on the intermediate value $x_1$, is given by

$$p(x_2, t_1 + t_2)dx_2 = \left[\int_{-\infty}^{\infty} p(x_1, t_1)p(x_2 - x_1, t_2)dx_1\right] dx_2. \tag{3.8}$$

During the derivation of this equation, it is implicitly assumed that the price process is memoryless. This follows, as price changes over the interval $(t_1, t_2)$ are independent of price changes during the interval $(0, t_1)$. Several decades after Bacelier, the equation now known in physics and mathematics as the Chapman-Kolmogorov-Smoluchowski (CKS) equation was rederived as a convolution equation for Markov processes[7] [16]. When solving the CKS equation, Bachelier ignored the issue of uniqueness of solutions. The equation has several solutions, a fact not recognized by Bachelier at that time, approaching the problem using a Gaussian distribution,

$$p(x, t) = p_0(t)e^{-\pi p_0^2(t)x^2}, \tag{3.9}$$

where $p_0(t)$ is the probability of the currently quoted price. Inserting equation (3.9) into equation (3.8) yields a relationship fulfilled by $p_0(t_1 + t_2)$,

$$p_0^2(t_1 + t_2) = \frac{p_0^2(t_1)p_0^2(t_2)}{p_0^2(t_1) + p_0^2(t_2)}. \tag{3.10}$$

[7]Statistically independent random processes.

From this relationship, it is easy to show that the time evolution of $p_0(t)$ is given by

$$p_0(t) = \frac{C}{\sqrt{t}}, \tag{3.11}$$

where $C$ is a constant. Performing the substitution $\sigma^2(t) = t/(2\pi C^2)$ leads to

$$p(x,t) = \frac{1}{\sqrt{2\pi\sigma^2(t)}} e^{-\frac{x^2}{2\sigma^2(t)}}. \tag{3.12}$$

This is a Gaussian distribution of mean $\mu$ equal to zero and a time dependent standard deviation $\sigma(t)$, as presented in figure 3.2.

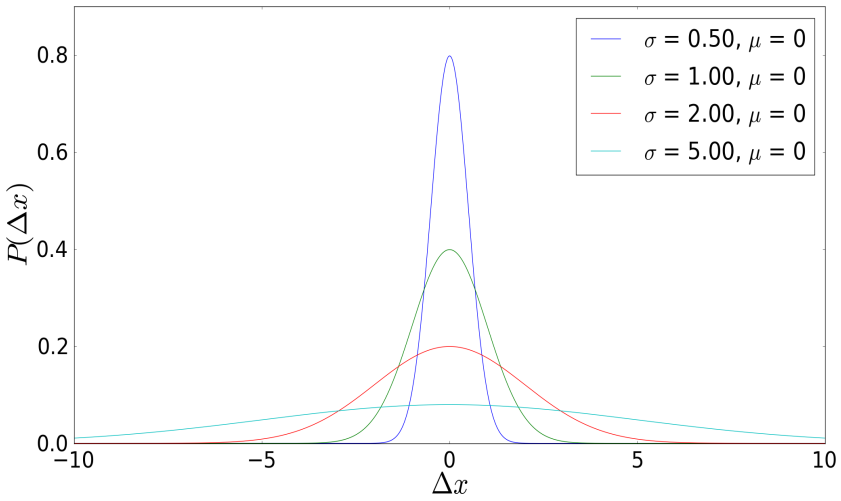



Figure 3.2: The Gaussian distribution for four different standard deviations $\sigma$, and mean $\mu = 0$. As $\sigma(t) \propto \sqrt{t}$, this corresponds to four different times. The distribution at $t = 0$ and thus $\sigma = 0$ corresponds to a delta function in $\Delta x = 0$, as the price of the asset is known at $t = 0$. This is not shown.

Note that the martingale property of the process is reflected in the fact that the mean of the distribution does not move with time. It is also observed that the standard deviation scales as $\sigma(t) \propto \sqrt{t}$, leading the distribution to broaden slowly. It follows that large price movements are extremely rare.

### Random walk

The same probability law as stated in equation 3.12 can be obtained also by considering a discrete model of asset price changes. Consider that there are only two possible events, $U$ and $D$, occurring with probabilities $p$ and $q = 1 - p$ respectively. If these events are considered as price changes by an amount $\pm\Delta x$ in one time step, the probability of observing $\alpha$ realizations of $U$ and $m - \alpha$ events of $D$ in a total of $m$ events will be given by the binomial distribution,

$$p_{U,D}(\alpha, m-\alpha) = \frac{m!}{\alpha!(m-\alpha)!} p^{\alpha}(1-p)^{m-\alpha}. \tag{3.13}$$

This is maximized for $\alpha = mp$ and $m - \alpha = mq$, and corresponds to the most probable price change in a string of $m$ events. For a fair game, $p = q = 1/2$, hence $p - q$ is zero. If $p \neq q$, the market has a drift. Bachelier considers the positive mathematical expectation of the *spread* $h$. The probability of a spread $h$ is the term in the expansion of $(p+q)^m$ where the exponent of $p$ is $mp + h$ and the exponent of $q$ is $mq - h$,

$$p_{U,D}(mp+h, mq-h) = \frac{m!}{(mp+h)!(mq-h)!} p^{mp+h}(1-p)^{mq-h}. \tag{3.14}$$

The spread $h$ can only take integer values. To obtain its mathematical expectation value, it is useful to rewrite $h$ as

$$h = q(mp+h) - p(mq-h). \tag{3.15}$$

The expectation value of $h$ is given by equation (3.4), where the integral must be transformed into a sum in the discrete case. The operation of multiplying the probability $p \propto q^{\mu}p^{\nu}$ by $h = \nu q - \mu p = pq\,(\nu/p - \mu/q)$ is done using the Leibniz rule, and is equivalent to taking the derivative with respect to $p$, subtract the derivative with respect to $q$ and finally multiply the result by $pq$. The positive mathematical expectation of $h$ is then obtained by taking the terms in the binomial expansion of $(p+q)^m$ where $h$ is positive,

$$p^m + mp^{m-1}q + \frac{m(m-1)}{1\cdot 2}p^{m-2}q^2 + ... + \frac{m!}{mp!mq!}p^{mp}q^{mq}, \tag{3.16}$$

and multiply them by $h$. After performing the operations as described above, an interesting observation is made. The derivative of the second term with respect to $q$ is equal to the derivative of the first term with respect to $p$. This also holds for the other terms, cancelling one another two by two. The only term left is the derivative of the last term with respect to $p$, such that the positive expectation of $h$ is given by

$$\frac{m!}{mp!\ mq!}p^{mp}q^{mq}mpq. \tag{3.17}$$

For large $m$, the expression can be simplified further using the asymptotic formula of Sterling,

$$n! = e^{-n}n^n\sqrt{2\pi n}, \tag{3.18}$$

resulting in the following equation for the positive mathematical expectation of $h$,

$$\frac{\sqrt{mpq}}{\sqrt{2\pi}}. \tag{3.19}$$

This leads to the distribution function of price changes,

$$p(h) = \frac{1}{\sqrt{2\pi mpq}}e^{-\frac{h^2}{2mpq}}, \tag{3.20}$$

in the limit where $m \to \infty$, $\alpha \to \infty$ and $h = \alpha - mp$ finite. Bachelier now assumes that $p = q = 1/2$, $h \to x$ and $t = m\Delta t$, where the time is divided into small time intervals $\Delta t$ where the price only varies a little. Notice that Bachelier no longer restricts the spread $h$ to be an integer anymore. The result is again the same Gaussian distribution as found in section 3.2.3,

$$p(x,t) = \frac{1}{\sqrt{\pi\sigma(t)}}e^{-\frac{x^2}{\sigma^2(t)}}, \tag{3.21}$$

where the substitution $\Delta t = t/(2\sigma^2)$ has been used. When published in 1900, this was the first description of a Gaussian random walk process.

### The diffusion law

Bachelier also derived the probability distribution via the diffusion equation. Assume that prices are discrete, $S_n$, and that they are realized at times $t$ in the future with probabilities $p_n$. Also assume that the prices in the small time interval $\Delta t$ can only change by a fixed amount $\pm\Delta S$. The probability of obtaining the price $S_n$ after one step, $p^*$, will in this case be given by $p^* = p_{n+1}/2 + p_{n-1}/2$, as it can only be reached either via an upward move from the price $S_{n-1}$ or by a downward move from the price $S_{n+1}$. The change of probability during the a step $\Delta t$ is therefore given by

$$\Delta p_n = p_n^* - p_n = \frac{p_{n+1} + p_{n-1} - 2p_n}{2}. \tag{3.22}$$

In the limit of continuous prices and time, this converges to[8]

$$\Delta p_n \to \frac{1}{2}\frac{\partial^2 p(S,t)}{\partial S^2}(\Delta S)^2. \tag{3.23}$$

In the same limit of continuous prices and time, the change of probability can also be written

$$\Delta p_n \to \frac{\partial p(S,t)}{\partial t}\Delta t. \tag{3.24}$$

The result is the diffusion equation,

$$D\frac{\partial^2 p(S,t)}{\partial S^2} - \frac{\partial p(S,t)}{\partial t} = 0. \tag{3.25}$$

The diffusion equation can also be derived from Fourier's Law. Fourier's Law states that the heat flow is proportional to the gradient of the temperature, and similarly, Bachelier shows that the probability flow is proportional to the gradient of the probability. Using as initial condition that the price at time $t = 0$ is known, the result is again the Gaussian distribution [13].

## 3.3 Einstein's theory of Brownian motion

Five years after Bachelier defended his doctoral thesis, Einstein published his first and famous article on Brownian motions [10]. As starting point, Einstein used $z$ gram of molecules of a non-electrolyte dissolved in a volume $V^*$. This volume $V^*$ formed a small part of a quantity of liquid of total volume $V$. When the volume $V^*$ is separated from the pure solvent by a partition permeable to the solvent but impermeable for the solute, an osmotic pressure $p$ is exerted on the partition. If the dissolved substance is replaced by small particles also unable to pass through the partition permeable to the solvent, the classical theory of thermodynamics and the molecular-kinetic theory of heat do not agree in the question whether there should be an osmotic pressure. The classical theory of thermodynamics suggests that there will be no force and thus no osmotic pressure acting on the partition[9]. The molecular-kinetic theory of heat on the other side suggests the opposite, according to this theory the dissolved molecule is differentiated from a suspended body solely by its dimensions. As Einstein himself states it, *"it is not apparent why a number of suspended particles should not produce the same osmotic pressure as the same number of molecules"* [10].

[8] $\Delta p_n = p_n^* - p_n = \frac{1}{2}(p_{n+1} + p_{n-1} - 2p_n) = \frac{1}{2}\left((p_{n+1} - p_n) - (p_n - p_{n-1})\right) = \frac{1}{2}(p'_{n+1}\Delta S - p'_n\Delta S) = \frac{1}{2}(p''_n)(\Delta S)^2$

[9] Einstein here neglects the forces of gravity.

Einstein draws the assumption that the suspended particles must perform an irregular movement in the liquid on account of the molecular movement of the liquid, and this way exert a pressure on the partition. Following, Einstein makes a series of assumptions of the properties of the irregular movements, and many are similar to those of Bachelier discussed in section 3.2:

- The movements of single particles are independent of one another to a sufficient degree of approximation.
- The movements of same particle after different intervals of time must be considered as mutually independent processes, as long as the time intervals are small.
- The movements executed by a particle in two consecutive intervals of time $\Delta t$ can be considered as mutually independent phenomena.
- Within the time interval $\Delta t$, a particle moves from $x_i \to x_i + \Delta x$, where $\Delta x$ is random, fulfilling the condition $p(\Delta x) = p(-\Delta x)$.

Based on these assumptions, Einstein derives the diffusion equation (equation 3.25), which is solved using the same boundary-condition as Bachelier: The position at time $t = 0$ is known, $f(x, t = 0) = \delta(x)$. The result is the Gaussian distribution [10],

$$f(x,t) = \frac{n}{\sqrt{4\pi Dt}} e^{-\frac{x^2}{4Dt}}, \tag{3.26}$$

where $n = cN$ is the number of suspended particles and $D$ the diffusion constant.

## 3.4 Probability theory

As the price process of stocks is approximated by stochastic processes, a short introduction to probability theory follows in this section. A probability distribution describes the probability for a continuous random variable of falling within a particular interval, or the probability for a discrete variable to obtain a particular value [23, pp. 298]. In this thesis, the most relevant distributions are considered, but information about other distributions can be found in textbooks such as [17] or [22].

### 3.4.1 Probability distributions

The probability distribution of a continuous variable $X$ cannot be stated in tabular form. It is rather represented as a function $P(x)$, which is a function of the value of the continuous variable $X$, giving the probability that $X$ lies in the small interval $dx$ around $X = x$. This function is denoted as the probability density function[10] and represents the probability density for a variable $x$ of appearing as the argument of the function [22].

The probability of obtaining $x$ between $a$ and $b$ is given by

$$\mathcal{P}(a \leq x \leq b) = \int_a^b P(x)dx. \tag{3.27}$$

A probability density function is always non-negative, and it is normalized such that an integral over the range of $X$ is equal to 1,

$$\int_{x_{min}}^{x_{max}} P(x)dx \equiv 1. \tag{3.28}$$

The limits $x_{min}$ and $x_{max}$ are the lower and upper bounds for where $P(x)$ is defined.

#### Mean-value, variance and standard deviations

The expected value of $X$, the mean value $\mu$, is defined

$$\mu = \langle x \rangle = \int_{x_{min}}^{x_{max}} x\ P(x)dx. \tag{3.29}$$

The variance $\sigma^2$ is the squared deviation from the mean, and given by

$$\sigma^2 = \langle (x - \langle x \rangle)^2 \rangle = \int_{x_{min}}^{x_{max}} (x - \langle x \rangle)^2 P(x)dx. \tag{3.30}$$

The positive square root of the variance is called the standard deviation $\sigma$ of $X$, and is a measure for the variation from the mean value of $X$. In general, higher order moments of the distribution $P(x)$ are defined as

$$m_n = \langle (x - \langle x \rangle)^n \rangle = \int_{x_{min}}^{x_{max}} (x - \langle x \rangle)^n P(x)dx. \tag{3.31}$$

For these moments to exist, the probability density function must decay faster than $1/x^{n+1}$ for $|x| \to \infty$ [17, pp. 7]. If not, the integral diverges and the moments are infinite. For some extreme cases, the variance and standard deviation do not exist. In the most extreme cases, even the mean diverges. This is the case for some Lèvy distributions that will be discussed later.

[10] Or in shorthand just PDF.

**Skewness and kurtosis**

The third and fourth moments are denoted as the skewness $\zeta$ and as the kurtosis $\kappa$ respectively [17, pp. 7],

$$\zeta = \frac{\langle (x-\mu)^3 \rangle}{\sigma^3}, \qquad \kappa = \frac{\langle (x-\mu)^4 \rangle}{\sigma^4} \tag{3.32}$$

The kurtosis describes the degree of peakedness of the distribution, and as a reference the Gaussian distribution has zero kurtosis. The skewness is a measure of asymmetry. An example is when one of the tails is more pronounced than the other, where a distribution having a most pronounced left tail will have a negative skewness. The Gaussian distribution has zero skewness.

### 3.4.2 Gaussian distribution

The most important continuous probability distribution is probably the Gaussian distribution, which is also denoted as the normal distribution. The distribution is bell-shaped, and presented in figure 3.2 for four different standard deviations. An important property of Gaussian random variables is that their sum is also a Gaussian random variable. Hence, the Gaussian is a stable distribution under addition. The shape of the distribution is preserved, up to a scale- and shift-term.

The distribution is ubiquitous, and phenomena from the number of heads in a sequence of coin tosses to the height of a randomly selected person are described approximately by this distribution [17, pp. 7]. According to the Central Limit Theorem, phenomena resulting from a large number of independent events where the distribution is not necessarily described by a Gaussian converges into a Gaussian when the number of events is large enough [22]. It is also the distribution Bachelier assumed controlled the fluctuations of assets [13].

Its general shape is given by [17, pp. 8]

$$P_G(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{3.33}$$

where $\mu$ is the mean and $\sigma$ the standard deviation of the distribution.

### 3.4.3 Log-normal distribution

The log-normal distribution also has a wide range of applications. It is important in finance, as it according to the geometric Brownian motion model

for stock-prices is assumed that it is the logarithmic returns that are independent random variables. Hence it is the logarithm of $X$ that is normally distributed, causing the shape of the distribution to be given by [17]

$$P_{LN}(x;\mu,\sigma) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\ln(\frac{x}{x_0})-\mu)^2}, \tag{3.34}$$

where $\mu$ is the mean and $\sigma$ the standard deviation of the distribution. The distribution is illustrated in figure 3.3 for different combinations of $\sigma$ and $\mu$.

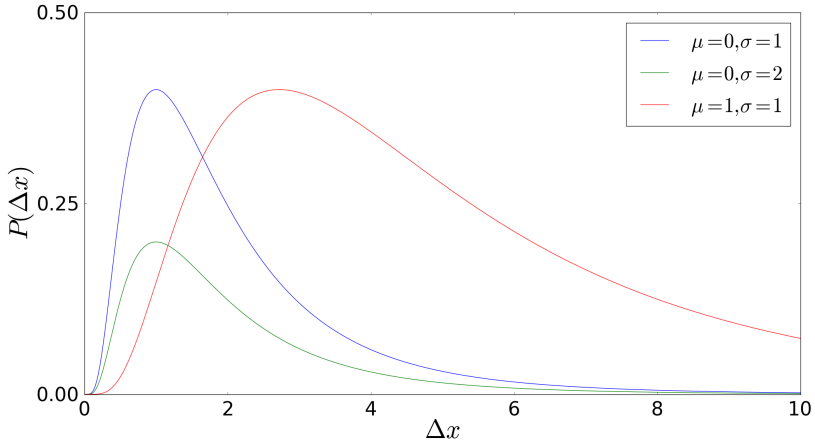



Figure 3.3: Plot of the log-normal distribution for different combinations of $\mu$ and $\sigma$.

### 3.4.4 Lèvy distribution

Lèvy distributions are similar to the Gaussian distributions stable under addition, but have fatter tails [17, pp. 10]. Due to the fat tails, there is an increased probability of events including large price changes. This was the reason for why Mandelbrot introduced such distributions to describe personal income and price changes of some financial assets such as e.g. cotton [9]. The important property of these distributions is their power-law behavior for large arguments, also known as the characteristic Pareto tail [17, pp. 10],

$$L_\mu(x) = \frac{\mu A_\pm^\mu}{|x|^{1+\mu}}, \qquad x \to \pm\infty. \tag{3.35}$$

Note that $\mu$ is a certain exponent and $A_\pm^\mu$ constants known as tail amplitudes (or scale parameters), giving the order of magnitude of the large fluctuations

of $x$. The parameter $\mu$ is limited to the range $0 < \mu \leq 2$ for Lèvy distributions, as $\mu > 2$ does not represent a stable probability function. In the case of $\mu = 2$, the Lèvy distribution reduces to the Gaussian distribution. A consequence of equation (3.35) with $\mu < 2$ is that the standard deviation of the distribution diverges to infinity [17, pp. 11]. This follows, as the probability density does not decay fast enough for the variance integral to converge, as discussed in the beginning of section 3.4.1. For $\mu < 1$ also the mean diverges.

Lèvy distributions are characterized by an asymmetry parameter that measures the relative weight of the positive and negative tail, $\beta = (A_+^\mu - A_-^\mu)/(A_+^\mu + A_-^\mu)$. In the symmetric case where $\beta = 0$, the characteristic function[11] of a symmetric Lèvy distribution is given by [17, pp. 11]

$$\tilde{L}_\mu(x) = e^{-a_\mu |x|^\mu}, \tag{3.36}$$

where $a_\mu$ is a constant proportional to the tail parameter $A_\pm^\mu$. The distribution is presented in figure 3.4. Notice that for decreasing $\mu$, the distribution becomes more and more peaked around the origin and exhibits fatter tails.

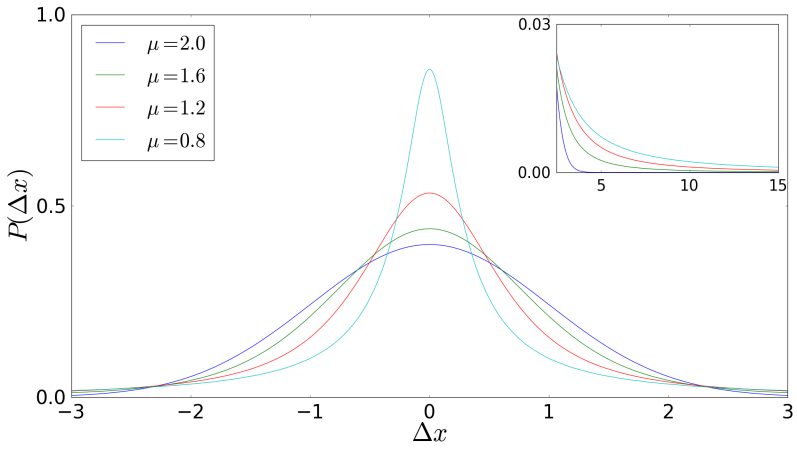



Figure 3.4: The symmetric Lèvy distribution with exponent $\mu = 0.8$, 1.2, 1.6 and 2.0. The case of $\mu = 2.0$ corresponds to a Gaussian. It can be observed that as $\mu$ decreases, the distribution gets more peaked with fatter tails. The inset emphasizes that tails are fatter for decreasing $\mu$.

[11]The characteristic function of a random variable X is the Fourier transform of its probability density, and oppositely the probability density is the inverse Fourier transform of the characteristic function.

### 3.4.5 Student's t-distribution

The distribution is often just denoted as the Student distribution, and given by [17, pp. 14]

$$P(x; a, \mu) = \frac{1}{\sqrt{\pi}} \frac{\Gamma((1+\mu)/2)}{\Gamma(\mu/2)} \frac{a^{\mu}}{(a^2 + x^2)^{(1+\mu)/2}}, \tag{3.37}$$

where the parameter $\mu$ describes the degrees of freedom of the distribution, tending towards a Gaussian distribution for $\mu \to \infty$ (provided that $a^2$ scales as $\mu$). Like Lèvy distributions, also Student distributions have power-tails. The Student distribution for different $\mu$ is presented in figure 3.5.

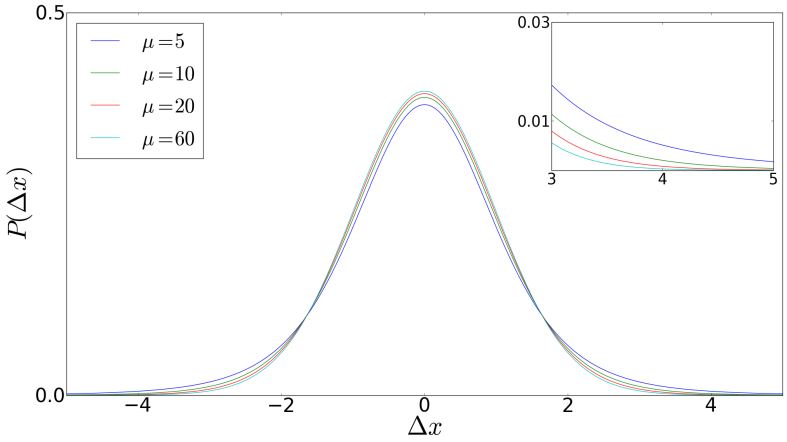



Figure 3.5: The Student's t-distribution for $\mu = 5, 10, 20$ and 60, and $a = \sqrt{\mu}$. From the inset, it is observed that the distribution falls off faster with increasing $\mu$. This is expected, as for $\mu \to \infty$, the Student distribution converges towards a Gaussian distribution.

## 3.5 Stock price processes

The idea of Gaussian distributed price changes developed by Bachelier [13] was later modified. This follows, as the model with stock-price changes controlled by a Gaussian distribution fails to capture a key aspect of stock-prices: The expected percentage return required by investors from a stock is independent from the price of the stock. As Bachelier's model includes a constant drift, it must be modified such that it is the expected return[12]

[12]Or equivalently the expected drift divided by the price of the stock.

that is constant. In addition, Bachelier's model led to a non-zero probability for negative prices. The solution to these problems was to model stock-price processes as a geometric Brownian motion [20, pp. 266], which still is the classical assumption made in theoretical finance [38]. After a short introduction to the general properties of a random walk process, this section will consider geometric Brownian motion and its implications when modeling stock-prices.

### 3.5.1 Markov processes

A Markov process is a stochastic process[13] where only the present value of a variable is relevant for predicting the future. The past history of the variable and the way the present has emerged from the past is not relevant for what will happen in the future. Therefore, the process is said to be memoryless. Stock-prices are assumed to follow a Markov process, as predictions for the future of the price should be unaffected by earlier prices.

### 3.5.2 Wiener processes

Wiener processes are continuous-time stochastic processes, and are often denoted as Brownian motion. Consider a variable following a Markov process, such that the only relevant information is the current value of the variable. If the variable $z$ fulfills the two following properties, it is said to be a Wiener process [20, pp. 261]:

- The change $\Delta z$ during a small period of time $\Delta t$ is $\Delta z = \epsilon\sqrt{\Delta t}$, where $\epsilon$ is standard Gaussian distributed with $\sigma^2 = 1$ and $\mu = 0$.
- The values of $\Delta z$ for two different short intervals of time, $\Delta t$, are independent.

The normal distribution is stable under addition. The change in $z$ during a time interval $T$ is given by

$$\Delta z(T) = z(T) - z(0) = \sum_{i=1}^{N} \epsilon_i \sqrt{\Delta t}, \tag{3.38}$$

and as the $\epsilon_i$'s are normally distributed and independent of each other, also the variable $\Delta z(T)$ must be normally distributed with mean $\mu_{\Delta z}(T) = 0$ and standard deviation $\sigma_{\Delta z}(T) = \sqrt{N\Delta t} = \sqrt{T}$. In the limit of continuous time, the time interval $\Delta t \to dt$ and thus the basic Wiener process $\Delta z \to dz$. The

[13] When the changes of a variable over time occur in an uncertain way, the variable is said to follow a stochastic process.

mean change per unit time for a stochastic process is known as the drift rate, and it follows that the basic Wiener process has no drift rate. As its variance equals $1 \cdot T$, the process has a variance rate of 1. It is useful to define a generalized Wiener process, including both a drift rate and a variance rate [20, pp. 263],

$$dx = adt + bdz. \tag{3.39}$$

It is observed that the first term corresponds to an expected drift of $a$ per unit time, while the second term represents the strength of the fluctuations or noise around this expected drift. It follows that the constants $a$ and $b$ correspond to the drift and the variance rate respectively. As the basic Wiener process itself has unit variance rate, the constant $b$ represents the variance of the fluctuations of the process. It follows that the change $\Delta x$ in the value of the variable $x$ during the time interval $\Delta t$ is given by

$$\Delta x = a\Delta t + b\epsilon\sqrt{\Delta t}, \tag{3.40}$$

with mean $\mu_{\Delta x} = a\Delta t$ and standard deviation $\sigma_{\Delta x} = b\sqrt{\Delta t}$. This model corresponds to the model of Bachelier's, where $\Delta x$ is the price change of a stock during the time interval $\Delta t$, and is by many considered as the origin of mathematical finance [13]. Note that Bachelier considered a case with no drift, such that $a = 0$.

### 3.5.3 Geometric Brownian motion of stock prices

To ensure a constant expected return of a stock of price $S(t)$, the drift rate must be proportional to the stock price $S(t)$ up to some constant parameter $\mu$ corresponding to the expected rate of return of the stock. During the small time interval $\Delta t$, the expected increase in price is $S\mu\Delta t$. In the case of zero fluctuations, the price increment in a small interval $\Delta t$ is given by [20, pp. 266]

$$\Delta S = \mu S\Delta t. \tag{3.41}$$

In the limit of continuous time and price, $\Delta t \to 0$, this is equal to

$$dS = \mu S dt. \tag{3.42}$$

An integration from 0 to $T$ results in an expression for the stock-price at times $T$,

$$S_T = S_0 e^{\mu T}, \tag{3.43}$$

where $S_0$ and $S_T$ are the stock price at times 0 and $T$. Hence, in the special case with no fluctuations, the stock-price is deterministic and grows at a continuously compounded rate of $\mu$ per unit of time. This also justifies the allegation in the introduction, that stocks are assumed to have a close to exponential drift. The case of zero fluctuations obviously is an artificial one. In real financial markets, stocks do have a non-zero volatility. However, it is fair to assume that an investor is just as uncertain of the percentage return, independently of the stock-price at the specific time. In other words, it is reasonable to assume that the variability of percentage returns is constant during a small time interval $\Delta t$, independently of the stock-price. It follows that the standard deviation of price fluctuations during the small time interval should be proportional to the stock-price [20, pp. 266],

$$\mathrm{d}S = \mu S \mathrm{d}t + \sigma S \mathrm{d}z. \tag{3.44}$$

The two terms correspond to the drift and the random part of the price changes respectively. The equation can easily be rewritten by dividing both sides by $S$, resulting in

$$\frac{\mathrm{d}S}{S} = \mu \mathrm{d}t + \sigma \mathrm{d}z. \tag{3.45}$$

This is known as geometric Brownian motion, and it seen that the left side (which is equal to the return) is normally distributed with standard deviation $\sigma$ and an expected rate of return $\mu$. In theoretical finance, one of the classical assumptions for asset prices is that they follow such a geometric Brownian motion [38]. Today, it is known that the model is not correct due to its underestimation of large fluctuations, but it is heavily used as it is easy to work with and gives good approximations when used. Figure 3.6 presents a model-stock based on geometric Brownian motion (GBM) along with the Oslo stock exchange listed company Statoil (STL)[14], during the period March 2009 to March 2011. Notice how similar the two cruves are.

[14]Data collected from Yahoo! Finance [21].

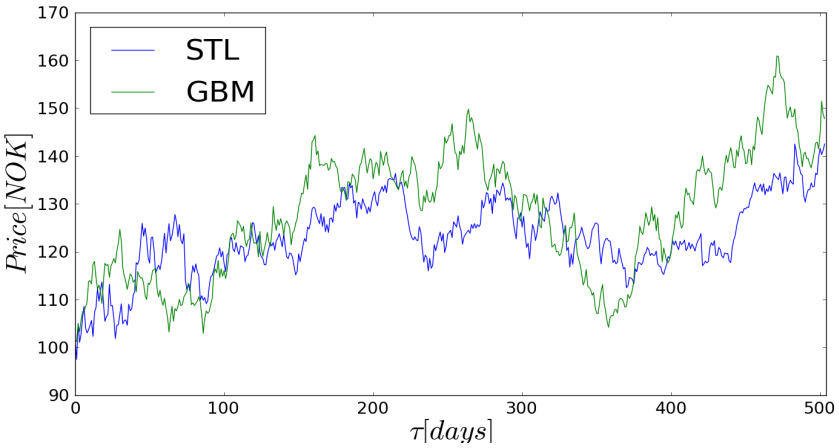



Figure 3.6: Blue line: Stock-price for the company Statoil during the period March 2nd, 2009, to March 1th, 2011. Green line: A model stock-price process, following a geometric Brownian motion (GBM) with drift coefficient $\mu = 0.067\%$ and $\sigma = 1.75\%$ per trading day (similar to that of STL). This equals the values of STL.

## 3.6 General matrix theory

In this section, a short introduction to matrices, covariance and correlation matrices, eigenvalues and eigenvectors is presented. This background is necessary for the understanding of the theory of random matrices introduced in the following section.

### 3.6.1 Eigenvalues and eigenvectors

An eigenvalue $\lambda$ and an eigenvector $\mathbf{v}$ of a matrix $\mathcal{A}$ fulfill the condition [30]

$$\mathcal{A}\mathbf{v} = \lambda\mathbf{v}, \tag{3.46}$$

where $\mathcal{A}$ is an $N \times N$ square matrix. The eigenvector must be nonzero, as the case of an eigenvector equal to the null-vector causes all eigenvalues to be valid. It is common to specify that the eigenvector $\mathbf{v}$ is associated with the eigenvalue $\lambda$, and eigenvalues are also denoted as proper values or characteristic values [30]. The characteristic equation for calculating the eigenvalues for a square matrix can be derived from equation (3.46), leading to

$$(\mathcal{A} - \lambda \mathbf{I})\mathbf{v} = \mathbf{0}. \tag{3.47}$$

This system has a nontrivial solution $\mathbf{v} \neq \mathbf{0}$ if and only if the determinant of its coefficient matrix is zero [30],

$$\det(\mathcal{A} - \lambda \mathbf{I}) = \mathbf{0}. \tag{3.48}$$

This equation is denoted as the characteristic equation, and is an $N$th degree polynomial for an $N \times N$ square matrix. Such an equation has $N$ solutions, where some of them may be complex. The solutions are not necessarily distinct, and eigenvalues from multiple roots are denoted as degenerate.

### 3.6.2 Eigenvalues and eigenvectors of covariance matrices

As an important aspect of risk management in finance is to estimate correlations between price movements of different assets, it is useful to define the covariance matrix[15]. Let the two series $\{X_i\}$ and $\{Y_i\}$ be of length $T$ and contain identically and independently distributed (iid) random variables. The covariance of the two series is given by

$$\sigma_{XY} = \frac{1}{T}\sum_{i=1}^{T}(x_i - \mu_x)(y_i - \mu_y), \tag{3.49}$$

where $\mu_x$ and $\mu_y$ are the mean values of the two series. This quantity is a measure of how much the two series change together, and in the special case where the series are identical, the covariance equals the variance. If the variables are standardized into having zero mean and unit variance, the covariance is given by the simplified expression

$$\sigma_{XY} = \frac{1}{T}\sum_{i=1}^{T} x_i y_i. \tag{3.50}$$

In this case, the covariance is also known as the correlation, as $\sigma_{XY} \in [-1, 1]$ for standardized variables. In finance, series of returns $\eta_i$ from each stock in a portfolio over a certain period $\tau$ are analyzed and used as a measure of risk. The returns $\eta_i$ are defined as

$$\eta_i = \frac{\delta x_i}{x_i} = \frac{x_{i+1} - x_i}{x_i}, \tag{3.51}$$

[15]Covariance matrices are among the cornerstones in Markowitz's theory of optimal portfolios.

where $x_i$ is the price of stock $x$ at time $t_i$. As discussed in chapter 2, the logarithmic return $r_i$ is approximately the same as this. Assume having a portfolio consisting of $N$ stocks and their returns $\eta_i$ over the last $T$ trading days. With one stock per row, this forms an $N \times T$ matrix $\mathbf{M}$. The covariance matrix has elements $C_{i,j}$ that correspond to the covariance between series $i$ and series $j$, given by equation (3.49). To avoid confusion, the equation is rewritten using returns,

$$C_{ij} = \frac{1}{T}\sum_{k=1}^{T}(\eta_k^i - \mu_i)(\eta_k^j - \mu_j) = \langle \eta^i \eta^j \rangle - \mu^i \mu^j, \tag{3.52}$$

where $\eta_k^i$ corresponds to the return from stock $i$ at time $t_k$. In the case of standardized variables, the matrix elements are given by

$$C_{ij} = \frac{1}{T}\sum_{k=1}^{T}\eta_k^i \eta_k^j = \langle \eta^i \eta^j \rangle. \tag{3.53}$$

In this case, the covariance matrix equals the correlation matrix. The covariance matrix $\mathbf{C}$ of the matrix $\mathbf{M}$ containing $N$ series of length $T$ can also be written symbolically as

$$\mathbf{C} = \frac{1}{T}\mathbf{M}\mathbf{M}^T. \tag{3.54}$$

The superscript $T$ denotes transposition, and $\mathbf{M}$ is the matrix containing the relevant time series. From standard matrix multiplication rules, it follows that the correlation matrix must be a square matrix. It is also symmetric, as the covariance between element $i$ and $j$ is the same as that between element $j$ and $i$. The symmetric correlation matrices have only real eigenvalues. In fact, correlation matrices *must* fulfill this property. This follows, as the presence of negative eigenvalues would make it possible to create a portfolio with negative variance. Determining empirical correlation matrices is complicated, and needs a lot of calculations. In fact, $N(N-1)$ entries must be determined from the $N$ time-series of length $T$. If the number of observations $T$ is not very large compared to the number of stocks $N$, the correlation matrix will be more or less dominated by noise. In other words, large parts of the covariance matrix is random [24]. Even though large parts of the matrix consist of noise, it can contain relevant information that must be carefully considered. This is discussed in section 3.7.

The analysis of eigenvalues and eigenvectors from correlation matrices based on financial data such as stock returns is interesting. As all correlation matrices are symmetric, it follows that they can be diagonalized with their eigen-

values on the diagonal[16]. Each eigenvector can be interpreted to describe a portfolio with return completely uncorrelated with the portfolios described by the other eigenvectors. This will be discussed in the following part of this section.

Assume that the eigenvector $\mathbf{v}_a$ corresponds to the eigenvalue $\lambda_a$, where the integer $a \in [1, \mathrm{N}]$. An eigenvector is a linear combination of the different assets $i$, fulfilling $\mathbf{C}\mathbf{v}_a = \lambda_a \mathbf{v}_a$. Hence eigenvector $\mathbf{v}_a$ is a list of weights $v_{a,i}$ of stocks. The variance of the return of a portfolio having a fraction $v_{a,i}$ of stock $i$ is [17, pp. 146]

$$\sigma_a^2 = \left\langle \left( \sum_{i=1}^{N} \frac{v_{a,i}}{x_i} \delta x_i \right)^2 \right\rangle = \sum_{i,j=1}^{N} v_{a,i} v_{a,j} C_{ij} = \mathbf{v}_a \cdot \mathbf{C}\mathbf{v}_a = \lambda_a, \tag{3.55}$$

where it is assumed that the returns $\eta_i = \frac{\delta x_i}{x_i}$ are standardized. The eigenvalue $\lambda_a$ corresponds to the variance of the portfolio constructed from the weights $v_{a,i}$. As eigenvectors always are orthogonal[17] [30], it is observed that the correlation of the return of two portfolios constructed from two different eigenvectors is zero [17],

$$\left\langle \left( \sum_{i=1}^{N} \frac{v_{a,i}}{x_i} \delta x_i \right) \left( \sum_{j=1}^{N} \frac{v_{b,j}}{x_j} \delta x_j \right) \right\rangle = \sum_{i,j=1}^{N} v_{a,i} v_{b,j} C_{ij} = \mathbf{v}_b \cdot \mathbf{C}\mathbf{v}_a = \delta_{a,b} \lambda_a, \tag{3.56}$$

where $\delta_{a,b}$ is the Kronecker delta. The result is a set of uncorrelated random returns $e_a$, corresponding to returns from portfolios constructed from the weights $v_{a,i}$. Using the notation where $\eta_i$ is the return from stock $i$, the resulting return from the portfolio described by $\mathbf{v}_a$ is given by [17, pp. 146]

$$e_a = \sum_{i=1}^{N} v_{a,i} \eta_i, \qquad \langle e_a e_b \rangle = \lambda_a \delta_{a,b}. \tag{3.57}$$

This can also be considered the opposite direction. The initial returns $\eta_k^i$ can also be considered as linear combinations of the uncorrelated factors,

$$\eta_i = \frac{\delta x_i}{x_i} = \sum_{a=1}^{N} v_{a,i} e_a. \tag{3.58}$$

[16]Assume forming matrix $\mathbf{P}$ with rows corresponding to the eigenvectors of the covariance matrix $\mathbf{C}$. In that case, the matrix $\mathbf{P}^{-1}\mathbf{C}\mathbf{P}$ will (due to the orthogonality of the eigenvectors) be a diagonal matrix with entries corresponding to the eigenvalues of $\mathbf{C}$.

[17]The orthogonality of eigenvectors is easy to show, starting from equation (3.46): I: $\mathbf{A}\mathbf{x} = g\mathbf{x}$ II: $\mathbf{A}^T\mathbf{y} = p\mathbf{y}$. First, equation I is multiplied by $\mathbf{y}^T$ from the left. Equation II is transposed and multiplied by $\mathbf{x}$ from the right. Following, equation I is subtracted from equation II, leaving $(p-g)\mathbf{y}^T \cdot \mathbf{x} = 0$. As the eigenvalues $p$ and $g$ are distinct, the eigenvectors $\mathbf{y}$ and $\mathbf{x}$ must be orthogonal.

This holds, as the transformation matrix $v_{a,i}$ from the initial vectors to the eigenvectors is orthogonal. As the correlated fluctuations of a set of random variables is decomposed in terms of the fluctuations of underlying uncorrelated factors, the decomposition is called a *principal component analysis* [17, pp. 146]. Principal components based on financial returns often have an economic interpretation in terms of sectors of activity, such as e.g. aviation or energy.

## 3.7 Random matrix theory

Random matrix theory (RMT) predicts the distribution of eigenvalues of matrices with entries that are identically and independently distributed (iid) random variables drawn from a probability distribution. The start of RMT has been traced back to the work of Wishart in 1928 [46], but the real start of the field is attributed to Wigner in 1955 [47]. Wigner was motivated by its applications in nuclear physics, and his idea was to describe energy levels of atomic nuclei and their fluctuations in position in terms of statistical properties of very large symmetric matrices with iid entries. Today, RMT is applied in many different fields from mathematical finance and statistics, nuclear physics and communication to biology [17, 24, 46, 48, 49].

This section provides a short introduction to RMT, where the density of eigenvalues is presented at the end of this section. As will be seen, the theory is developed on the assumption of infinitely large matrices. No real-world matrices are infinitely large, and hence a small numerical experiment is presented where it is shown that RMT also describes the density of eigenvalues for smaller matrices very well.

### 3.7.1 The density of eigenvalues

Let $\mathbf{H}$ be a square matrix of size $N \times N$, filled with iid random variables. As the correlation matrices considered in this thesis are symmetric, it is also assumed that $\mathbf{H}$ is symmetric, having elements $H_{ij} = H_{ji}$. In the limit of very large matrices where $N \to \infty$, the distribution of eigenvalues is to a large extent independent of the elements of the matrix [17, pp. 161]. The following derivation is based on the introduction in [17, pp. 161 - 163].

Let the density of eigenvalues be given by

$$\rho(\lambda) = \frac{1}{N} \sum_{a=1}^{N} \delta(\lambda - \lambda_a), \tag{3.59}$$

where $\delta$ is the Dirac delta-function. Let the resolvent[18] $\mathbf{G}(\lambda)$ of the matrix $\mathbf{H}$ be defined as

$$G_{ij}(\lambda) = \left(\frac{1}{\lambda\mathbf{I} - \mathbf{H}}\right)_{ij}, \tag{3.60}$$

where $\mathbf{I}$ is the identity matrix. The trace of $\mathbf{G}(\lambda)$ expressed using eigenvalues of $\mathbf{H}$ is

$$\mathrm{Tr}\mathbf{G}(\lambda) = \sum_{\mathrm{a}=1}^{\mathrm{N}} \left(\frac{1}{\lambda - \lambda_{\mathrm{a}}}\right). \tag{3.61}$$

To calculate $\rho(\lambda)$ for large $N$, it is useful to represent the Dirac delta-function by the identity

$$\frac{1}{x - i\epsilon} = \mathrm{PP}\frac{1}{x} + i\pi\delta(x) \qquad\qquad \epsilon \to 0, \tag{3.62}$$

where PP means the principal part. By replacing the delta-function in equation (3.59), knowing that the eigenvalues of a symmetric matrix always is real, the density of eigenvalues can be written

$$\rho(\lambda) = \frac{1}{N\pi}\Im\left(\mathrm{Tr}\mathbf{G}(\lambda - i\epsilon)\right). \tag{3.63}$$

It follows that an expression for the resolvent must be found. This can be done by establishing a recursion relation, allowing to compute $\mathbf{G}(\lambda)$ for a matrix $\mathbf{H}$ with one extra row and one extra column. The resolvent element $G_{00}^{N+1}(\lambda)$ is calculated using the standard formula for matrix inversion,

$$G_{00}^{N+1}(\lambda) = \frac{\mathrm{minor}(\lambda\mathbf{I} - \mathbf{H})_{00}}{\det(\lambda\mathbf{I} - \mathbf{H})}, \tag{3.64}$$

where the superscript stands for the size of the matrix $\mathbf{H}$ and the subscript implies that row and column 0 are excluded from the matrix $(\lambda\mathbf{I} - \mathbf{H})$ appearing in the minor. The determinant appearing in the denominator can be expanded along its first row, using the relation [31],

$$\det(\mathbf{A}) = \sum_{j=1}^{N}(-1)^{i+j}\ a_{ij}\ \det\mathbf{A}_{i,j}, \tag{3.65}$$

[18]The resolvent is introduced as a tool for finding the density of eigenvalues. To understand why this is helpful, it is best to consider an example: Consider the problem $(\mathbf{A} - \lambda\mathbf{I})\mathbf{u} = \mathbf{b}$. That is, given $\mathbf{A}$, $\mathbf{b}$ and $\lambda$, find $\mathbf{u}$. This can be done by solving the alternate problem $(\mathbf{A} - \lambda\mathbf{I})\mathbf{G}(\lambda) = \mathbf{I}$, where $\mathbf{G}_\lambda = (\mathbf{A} - \lambda\mathbf{I})^{-1}$. $\mathbf{G}(\lambda)$ is a matrix known as the resolvent of the matrix $\mathbf{A}$. After solving for $\mathbf{G}(\lambda)$, the vector $\mathbf{u}$ is found by solving $\mathbf{u} = \mathbf{G}(\lambda)\mathbf{b} = (\mathbf{A} - \lambda\mathbf{I})^{-1}\mathbf{b}$.

where $a_{ij}$ is element $(i, j)$ of matrix $\mathbf{A}$, and the determinant is expanded along the $i$th row. After expanding along the first row of the matrix $\mathbf{X} = (\lambda\mathbf{I} - \mathbf{H})$, the denominator of equation (3.64) (in shorthand D) equals

$$D = \sum_{j=0}^{N} (-1)^j \; x_{0j} \; \mathrm{det}\mathbf{X}_{0j}. \tag{3.66}$$

The term for which $j = 0$ can be placed outside the sum, as the matrix $X_{00}$ has neither row or column 0. Following, the determinant $\mathrm{det}X_{0j}$ can be expanded similarly to what was done along the first row, only now about the first column. The result is

$$D = x_{00} \; \mathrm{det}\mathbf{X}_{00} + \sum_{j=1}^{N} (-1)^j \; x_{0j} \; \sum_{i=1}^{N} (-1)^{i+1} \; x_{i0} \; \mathrm{det}\mathbf{X}_{ij}, \tag{3.67}$$

where the extra sign enters as the columns now are numbered from 1 to $N$, not from 0 to $N$. The final result for the denominator is

$$D = x_{00} \; \mathrm{det}\mathbf{X}_{00} - \sum_{j=1,i=1}^{N} (-1)^{i+j} \; x_{0j} \; x_{i0} \; \mathrm{det}\mathbf{X}_{ij}. \tag{3.68}$$

Inserting this into the inverted equation (3.64) results in an expression for the resolvent $G_{00}^{N+1}(\lambda)$,

$$\frac{1}{G_{00}^{N+1}(\lambda)} = \lambda - h_{00} - \sum_{j,i=1}^{N} \; h_{0j} \; h_{i0} \; G_{ij}^{N}(\lambda), \tag{3.69}$$

where $h_{ij}$ is element $(i, j)$ of the matrix $\mathbf{H}$. Note that the sign-term $(-1)^{i+j}$ is drawn into $G_{ij}^{N}(\lambda)$. By assuming that $h_{ij}$ are iid random variables of zero mean and a variance[19] $\langle h_{ij}^2 \rangle = \sigma^2/N$, it can be shown that $G_{ij} \propto 1/\sqrt{N}$ for $i \neq j$, while $G_{ij}$ remains finite in the case $i = j$ [17, pp. 162]. Hence, terms with $i \neq j$ can be discarded. The term $h_{00} \propto 1/\sqrt{N}$ is small compared to $\lambda$, and can be neglected. This leaves a simplified recursion relation only valid when $N \to \infty$,

$$\frac{1}{G_{00}^{N+1}(\lambda)} \approx \lambda - \sum_{i=1}^{N} h_{0i}^2 \; G_{ii}^{N}(\lambda), \tag{3.70}$$

[19]The variance is assumed to be equal to $\langle h_{ij} \rangle = \sigma^2/N$, as this ensures the finiteness of the vector resulting from the matrix $\mathbf{H}$ acting on another vector. Each component of the vector will in this case be a sum of $N$ variables, and for $N \to \infty$ the only way to ensure finiteness is by letting the variance scale as $\langle h_{ij}^2 \rangle = \sigma^2/N$.

According to the Central Limit Theorem discussed in section 3.4.2, this sum converges towards $\sigma^2/N \sum_{i=1}^{N} G_{ii}^{N}(\lambda)$ for $N \to \infty$. Hence $G_{00}^{N+1}(\lambda)$ converges into a well defined limit for large $N$,

$$\frac{1}{G_\infty(\lambda)} = \lambda - \sigma^2 G_\infty(\lambda). \tag{3.71}$$

Such a second order equation is easily solved, giving

$$G_\infty(\lambda) = \frac{1}{2\sigma^2}(\lambda - \sqrt{\lambda^2 - 4\sigma^2}). \tag{3.72}$$

For this equation to have a non-zero imaginary part when adding a small term $i\epsilon$ to $\lambda$, the square root itself must be imaginary. Hence, according to equation (3.63), the final result for the density of eigenvalues of the symmetric square matrix is $\mathbf{H}$ is

$$\rho(\lambda) = \frac{1}{2\pi\sigma^2}\sqrt{4\sigma^2 - \lambda^2}, \qquad |\lambda| \leq 2\sigma. \tag{3.73}$$

This equation is known as *Wigner's semi-circle law* for the density of states. According to equation (3.54), the correlation matrix can be written as $\mathbf{C} = \mathbf{H}\,\mathbf{H}^T$, up to some constant. The matrix $\mathbf{H}$ is not restricted to be a square matrix, and can in fact be any rectangular matrix. In financial data, rows often correspond to different stocks while columns correspond to the observations. Hence, the number of columns is often far lager than that of the rows. In the specific case where $\mathbf{H}$ is a square matrix, the eigenvalues of the resulting correlation matrix $\mathbf{C}$ can be obtained by squaring the eigenvalues of matrix $\mathbf{H}$. This is easily seen by manipulating equations (3.46) and (3.54), knowing that the eigenvalues of a square matrix and its transpose are equal[20] [17, pp. 163],

$$\lambda_C = \lambda_H^2. \tag{3.74}$$

[20] $\det(\mathbf{A} - \lambda\mathbf{I}) = \det(\mathbf{A} - \lambda\mathbf{I})^T = \det(\mathbf{A}^T - \lambda\mathbf{I})$.

Assuming that all elements of $\mathbf{H}$ are iid random variables, the density of eigenvalues of the correlation matrix $\mathbf{C}$ can therefore be obtained using the following relation,

$$\rho(\lambda_C)d\lambda_C = 2\rho(\lambda_H)d\lambda_H, \qquad \lambda_H > 0. \tag{3.75}$$

The factor of two arises from the two solutions $\lambda_H = \pm\sqrt{\lambda_C}$. When $\mathbf{H}$ is a square matrix, inserting equation (3.73) into equation (3.75) results in the following distribution of eigenvalues for the correlation matrix,

$$\rho_C(\lambda) = \frac{1}{2\pi\sigma^2}\sqrt{\frac{4\sigma^2 - \lambda_C}{\lambda_C}}, \qquad 0 \leq \lambda_C \leq 4\sigma^2, \tag{3.76}$$

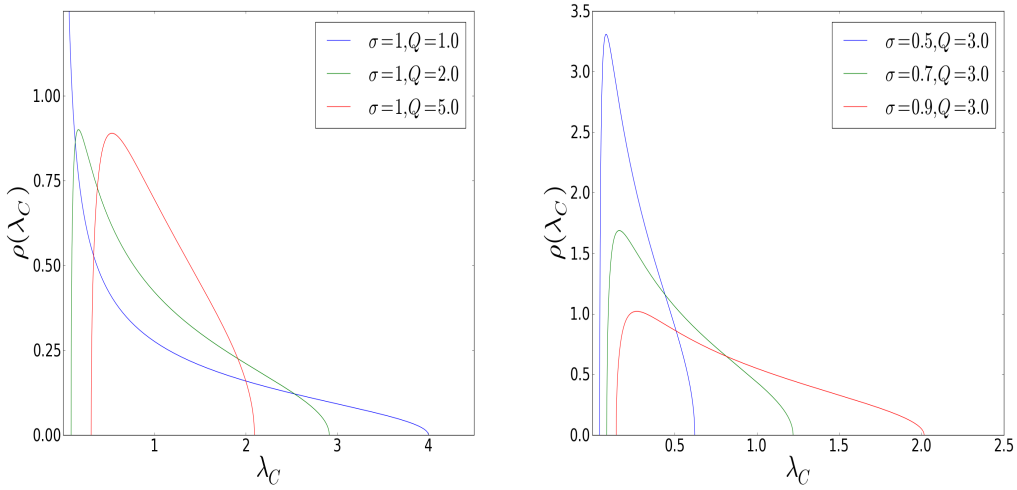
where $\sigma^2/N$ is the variance of the elements in $\mathbf{H}$ and $\sigma^2$ in other words is the average eigenvalue of $\mathbf{C}$. A similar formula exists for rectangular $\mathbf{H}$. Assuming that $N, T \to \infty$ and that their ratio $Q = T/N \geq 1$ is fixed, the density of eigenvalues of the correlation matrix $\mathbf{C}$ is given by [17, pp. 163]

$$\rho(\lambda_C) = \frac{Q}{2\pi\sigma^2}\frac{\sqrt{(\lambda_{max} - \lambda_C)(\lambda_C - \lambda_{min})}}{\lambda_C}, \qquad \lambda_{min} \leq \lambda_C \leq \lambda_{max}. \tag{3.77}$$

The constants $\lambda_{max}$ and $\lambda_{min}$ are given by

$$\lambda_{min}^{max} = \sigma^2\left(1 + \frac{1}{Q} \pm 2\sqrt{\frac{1}{Q}}\right). \tag{3.78}$$

Curves resulting from equation (3.77) are presented in figure 3.7 for different standard deviations $\sigma$ and ratios $Q$. An inspection of the figure reveals that except for the particular case where $\mathbf{H}$ is a square matrix and $Q = 1$, the lower bound of eigenvalues of $\mathbf{C}$ is positive. Hence, there are no eigenvalues between 0 and $\lambda_{min}$. Near this edge, the density exhibits a sharp maximum and then decays until the upper edge given by $\lambda_{max}$ is reached. Again there are no eigenvalues between $\lambda_{max}$ and $\infty$. For the particular case where $\mathbf{H}$ is a square matrix, there is no lower edge. The density diverges as $\propto 1/\sqrt{\lambda}$ before reaching $\lambda_{max}$ where it dies off. It is also noted that a decreasing standard deviation $\sigma$ causes the density to be more peaked [17, pp. 163].




(a) (b)

Figure 3.7: Density of eigenvalues for random matrices as predicted from equation (3.77). (a) Illustrated for three different ratios $Q$. (b) Illustrated for three different standard deviations $\sigma$.

As no real world matrices are infinitely large, noise will always be present in empirical correlation matrices. This is always the case for financial data. It follows that a large part of the empirical correlation matrix is random and must be considered carefully when information is extracted from it. As the smallest eigenvalues and their corresponding eigenvectors determine the least risky portfolio[21], it is extremely important being able to distinguish true information from random noise. This is where the theory of random matrices is useful. A *null hypothesis* matrix[22] is compared to the empirical correlation matrix, and deviations from the random matrix case can possibly reflect true information. Lalox et al. [24] compared the empirical distribution of eigenvalues from a correlation matrix based on $N = 406$ stocks, each with $T = 1309$ observations, to the distributions of eigenvalues of a completely random matrix as given by equation (3.77). An excellent fit to theory was obtained, with the exception of several eigenvalues observed above the theoretical upper limit $\lambda_{max}$. Their result is presented in figure 3.8. The highest eigenvalue was observed to be 25 times as large as $\lambda_{max}$, and its associated eigenvector is believed to correspond to the market itself, as it has roughly

[21]Small eigenvalues corresponds to eigenvectors with most entries zero, in other words to single assets or smaller groups of assets. Thus the noise hides information about the correlations between single assets or small groups of assets.

[22]A purely random matrix obtained from finite time series of independent assets.

equal components of all stocks. Also other eigenvalues were found to be significantly larger than $\lambda_{max}$. These eigenvalues are believed to correspond to different sectors of the economy, as briefly mentioned in section 3.6.2.

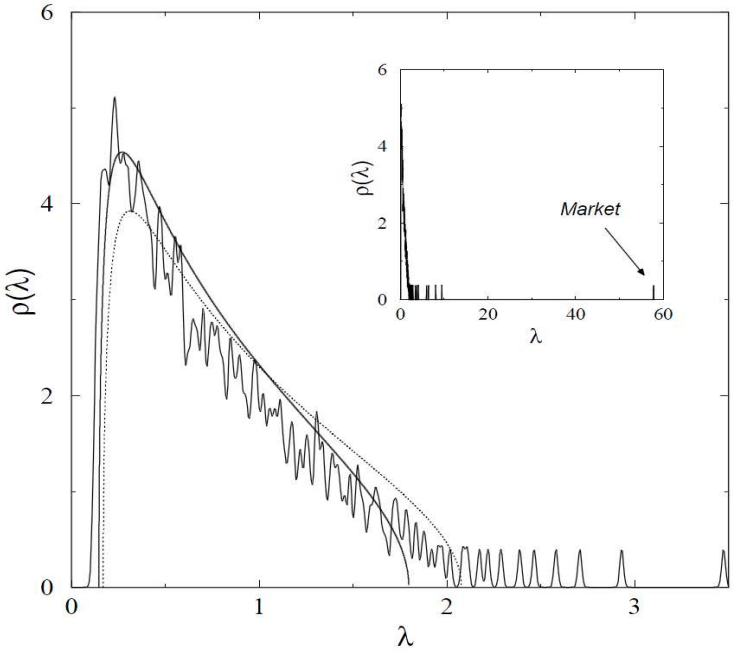



Figure 3.8: Smoothed density of the eigenvalues of $\mathbf{C}$, based on historical data from 406 stocks of the S&P500 during the years 1991 - 1996. The solid and dotted lines are fits using equation (3.77). The solid line corresponds to $\sigma^2 = 0.85$, the theoretical value obtained after subtracting the variance corresponding to its highest eigenvalue. The dotted line corresponds to $\sigma^2 = 0.74$, where also the smaller (but larger than predicted) eigenvalues corresponding variances have been subtracted. The inset presents the same plot, but also the largest eigenvalue is included. This is observed to be nearly 25 times as large as the predicted upper limit. Figure adapted from [24].

### 3.7.2 A numerical experiment on finite sized matrices

Random matrix theory (RMT) is based on the assumption of infinitely large matrices. All real-world matrices are finite. Hence it is important to check whether RMT can be used approximately for finite matrices. The numerical experiment is done by modeling $N$ stocks over a period of $T$ observations, based on the model of geometric Brownian motion discussed in section 3.5.3. The density of eigenvalues is calculated from the resulting correlation matrix, and compared to the theoretical prediction of a purely random matrix as stated in equation (3.77).

In general, an excellent agreement is found between model-data and theoretical predictions from the theory of random matrices. The agreement is excellent for large matrices having about 1500 rows and more, but for smaller matrices it is somewhat noisy and must be averaged over. This averaging can be done by doing an ensemble average or by averaging over nearby datapoints. The averaging procedure causes also the density of smaller matrices to fit very well to the theoretical prediction. The experiment is divided into two parts, first relatively large matrices with more than 1500 rows. Following, matrices with only 500 rows are considered and averaged over using both methods as described above.

#### Large matrices

Matrices of $N = 1500$ rows up to $N = 8000$ rows are considered. In principle it is interesting to consider as large matrices as possible, but the time it takes to calculate the correlation matrix as well as the eigenvalues (and the corresponding eigenvectors) scales approximately as the third power of the correlation matrix size $N$ [53]. Hence, the size will be limited to $N = 8000$ in this experiment. As will be seen, the density of eigenvalues from the matrix containing 8000 model-stocks indeed fits the theory from the random matrices excellently such that there is no need to consider larger matrices.

The matrices considered are of the following sizes ($N \times T$):

- A $8000 \times 8000$ matrix.
- A $6000 \times 6000$ matrix.
- A $4000 \times 12000$ matrix.
- A $1500 \times 4500$ matrix.

The number of columns is varied, but it is the number of rows that is the important factor. This follows, as it is from the correlation matrix the density of eigenvalues is calculated. More columns would only improve the statistics for the correlations between the series, while the number of stocks determines the dimensions of the correlation matrix. Figure 3.9 shows the density of eigenvalues along with the theoretically predicted density from the theory of random matrices. Notice that no averaging is done in any of these cases. It is observed that the model-data fits the predicted density of eigenvalues very well, but that more noise is introduced for the smallest $N$.

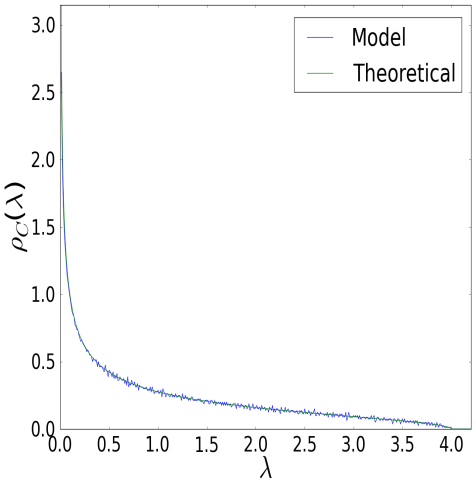
### Smaller matrices

In this thesis, the matrices concerned will be rather small with about 500 rows and 1500 columns. Therefore, it is also of interest to consider smaller matrices. The matrices considered are the following:

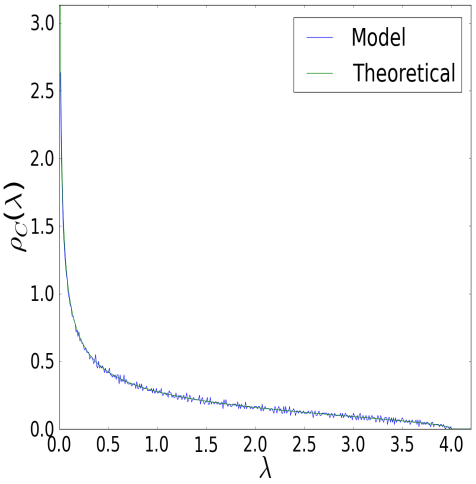
- A $500 \times 1500$ matrix.
- An ensemble average of 10 matrices of size $500 \times 1500$ matrices.
- An ensemble average of 100 matrices of size $500 \times 1500$ matrices.
- A $500 \times 1500$ averaged over the two nearest data-points.

The resulting plots of theoretical densities of eigenvalues, as well as the density based on the model-stocks, are shown in figure 3.10. It is seen that the $500 \times 1500$ matrix itself leads to a noisy density of eigenvalues, but an ensemble average or an average over nearby data-points leads to a smoothened density that fits the predicted density very well. The conclusion is that matrices with only 500 rows fit very well to theory when averaged correctly. As only eigenvalues deviating considerably from the predications will be considered in this thesis, better fits are not needed.

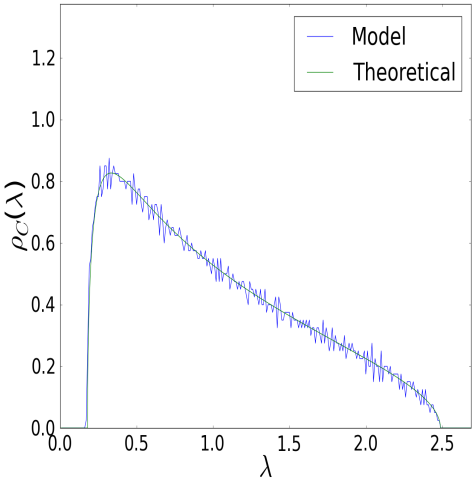
The procedure followed to smoothen the empirical eigenvalue densities that are calculated in this work is an average over the nearest data-point. This smoothens the densities but does not affect the result. This follows, as it is the largest eigenvalue itself that is considered and not the density.



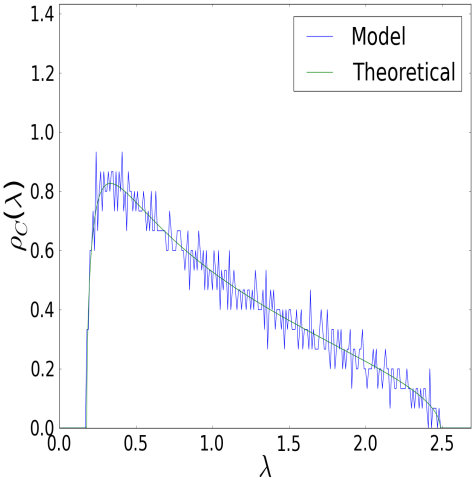

(a) $Q = 1$, $N = 8000$ and $T = 8000$.




(b) $Q = 1$, $N = 6000$ and $T = 6000$.




(c) $Q = 3$, $N = 4000$ and $T = 12000$.




(d) $Q = 3$, $N = 1500$ and $T = 4500$.

Figure 3.9: The density of eigenvalues as predicted from the random matrix theory (equation (3.77)) and the empirical densities from model-stocks following geometric Brownian motion for different combinations of $Q$, $N$ and $T$. Observe that the model-data fits very well to the theory for the largest matrices, but that more noise is introduced for the smallest values of $N$.

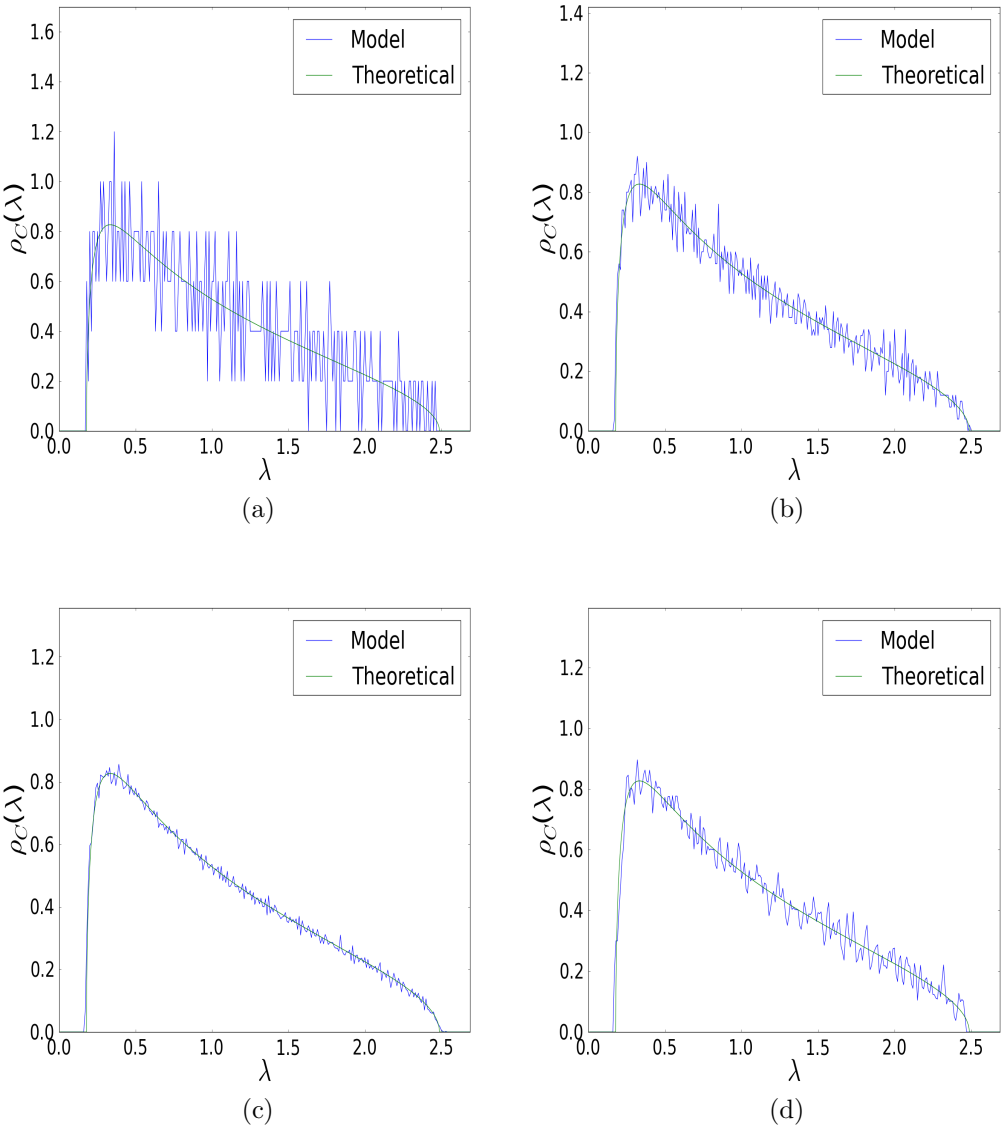



Figure 3.10: The density of eigenvalues for (a) a $500 \times 1500$ matrix. Figures (b) and (c) are ensemble averages over 10 and 100 matrices of size $500 \times 1500$. Figure (d) is the density from one matrix of size $500 \times 1500$, but where an average over the two nearest points has been performed to smoothen the distribution. The figures also contain a plot of the density of eigenvalues as predicted from RMT (equation (3.77)). Observe that most of the noise vanish doing either an ensemble average or an averaging over nearby points.

## 3.8 Inverse statistics

Time-series analysis is a central topic in many disciplines within physics, economy and finance. An example where time-series analysis is important in physics is turbulence, where water molecules move in what seems to be a random fashion with violent changes of direction [5, p. 88]. In finance, enormous amounts of financial data are available. Stock indexes and individual stock quotes from most major stock exchanges have been recorded and stored the last century, in addition to interest rates and exchange rates. The classical analysis of a financial time-series has been to choose a typical timescale, and to find the typical return over this period based on historical data. Such distributions of returns are often used as proxies for the performance of stocks and markets over a certain period [40]. Ahlgren et al. [34] denoted this type of analysis as *forward statistics*, as the properties of the probability distribution of returns are calculated from historic data over a fixed time window.

In the inverse statistics approach, one "inverts" the question and asks *"What is the typical time span needed to generate a fluctuation or a movement (in the price) of a given size"* [36]? Thus the distribution of waiting times needed to reach a fixed level of return for the first time is studied in the inverse statistics approach.

### 3.8.1 The inverse statistics distribution

To calculate the inverse statistics, it is useful to introduce the logarithmic return $r_{\Delta t}(t)$ at a fixed time $t$ calculated over a time window $\Delta t$ [36],

$$r_{\Delta t}(t) = s(t + \Delta t) - s(t). \tag{3.79}$$

The quantity $S(t)$ is the asset price and $s(t) = \ln S(t)$ and is simply the logarithm of the price. In other words, the log-return is simply the log-price change of the asset. For small changes in price, this is approximately equal to the normal return $\eta_i = \delta S_i / S_i$, as discussed in chapter 2. In the inverse statistics approach, the distribution of waiting times needed to reach a fixed level of return is studied. If the investment is made at time $t$, the investment horizon is defined as the time $\tau_\rho(t) = \Delta t$ it takes to satisfy either $r_{\Delta t}(t) \geq |\rho|$ (gain) or $r_{\Delta t}(t) \leq -|\rho|$ (loss) for the first time. In particular, one searches for the shortest waiting time $\tau_{\pm|\rho|}(t)$ one must wait before reaching the fixed return level $\pm|\rho|$ for the *first* time [36].

As discussed in section 3.5.3, a classical assumption in finance is that the stock-price follows a geometric Brownian motion. Hence, the log-return follows an ordinary Brownian motion. Under this assumption, inverse statistics correspond to the *first passage probability* of a random walker [36], and is known analytically [35, pp. 363]. The distribution of waiting times for a random walker to cross a barrier for the first time at a distance $\rho$ from its starting point is given by [35, pp. 363]

$$f(\tau_\rho) = \frac{\rho}{\sqrt{4\pi D\tau_\rho^3}} e^{-\frac{\rho^2}{4D\tau_\rho}}, \tag{3.80}$$

where the diffusion constant of the random walker is $D = \frac{\sigma^2}{2\Delta_t}$ and $\sigma$ is the variance of the random walker's step size distribution. The discrete time-interval $\Delta t$ between two consecutive steps is set to 1 for simplicity. This distribution of waiting times is referred to as the *first passage probability density* and is a member of the *inverse Gamma* and *inverse Gaussian* distribution families [34]. As the empirical stock-price process is known not to follow a geometric Brownian motion, Simonsen et al. [36] suggested to use the generalized Gamma distribution as basis to fit the empirical investment horizon distributions,

$$p(t) = \frac{\nu}{\Gamma(\frac{\alpha}{\nu})} \frac{\beta^{2\alpha}}{(t+t_0)^{\alpha+1}} \exp\left(-\left(\frac{\beta^2}{t+t_0}\right)^\nu\right). \tag{3.81}$$

This distribution was observed to parametrize the data excellently, as will be further discussed in the following section. Notice that the distribution reduces to the first passage probability in the limit of $\alpha = 1/2$, $\beta = \sqrt{\rho^2/4D}$, $\nu = 1$ and $t_0 = 0$.

For long waiting times, equation (3.80) decays as a pure power-law decay $p(\tau_\rho) \propto \tau^{-3/2}$ for all $\rho$. The special case with $\rho = 0$ leads to a pure power-law decay, and the case of $\rho \neq 0$ leads to a maximum before asymptotically reaching the power-law regime with the same exponent. It is easy to show from equation (3.80) that the maximum is located at

$$\tau_\rho^* = \frac{\rho^2}{6D} = \frac{\Delta t}{3}\frac{\rho^2}{\sigma^2}. \tag{3.82}$$

This implies that the maximum of inverse statistics distributions scales with the return-level $\rho$, following the power-law $\tau_\rho^* \propto \rho^\gamma$ with $\gamma = 2$. The first study of inverse statistics within finance was conducted by Simonsen et al. [36], using daily closure data from Dow Jones Industrial Average (DJIA) from 1896 to 2001. The maximum of the inverse statistics distribution was coined *the optimal investment horizon* [36], and found to deviate from theoretical

predictions based on the geometric Brownian motion of stock-prices. This will be discussed in the following section. Note that it is not the return-level alone, but rather the return-level scaled by the standard deviation of the log-returns, $\rho/\sigma$, that enters expressions (3.80) and (3.82). As the quantity appears squared, the position of the maximum is independent of the sign of the fixed return-level. Hence for a random walker, the maximum is theoretically predicted to be invariant under change of sign of the return-level $\rho$. For financial data, this is not observed. Empirical results based on data from the DJIA showed that the loss inverse statistics curve is shifted towards shorter waiting times relative to the gain curve, now known as *the gain-loss asymmetry* [38].

### 3.8.2 The gain-loss asymmetry in financial markets

Based on a pure geometric Brownian motion model for stock-prices, the inverse statistics distributions should be invariant under the change of sign of return-level $\rho$. However, there is an apparent asymmetry between the empirical investment horizons for positive and negative return-levels $\rho$. This is observed from figure 3.11, presenting the resulting inverse statistics distributions calculated by Jensen et al. [38] based on detrended DJIA daily closure prices. It is also revealed that the empirical data fit the generalized Gamma distribution (3.81) excellently for both positive and negative return-levels. As was pointed out in [38], in addition to the asymmetry in the positions of the two peaks, there also is a higher probability of finding short investment horizons for negative return-levels compared to for positive return-levels. This supports the phrase often heard in financial contexts, *"Draw-downs are faster than draw-ups"* or *"It takes time driving up prices, compared to driving them down"* [38]. It is also observed that the optimal investment horizon[23] is consistently found to occur first for negative returns in liquid and mature western markets, while the opposite has been been reported in some cases for emerging markets [34].

According to equation (3.82), the investment horizon $\tau_\rho$ is proportional to the return-level $\rho$ to the power of an exponent $\gamma$, fulfilling $\tau_\rho^* \propto \rho^\gamma$. As discussed in section 3.8.1, this exponent is theoretically predicted to be $\gamma = 2$ for stock prices assumed to follow a geometric Brownian motion. The results of Jensen et al. [38] are presented in figure 3.12, and it is clear that their observation is not what one would expect from theory. For the lowest return-levels, the dependence seems to be unknown. For larger return-levels, however, the unknown dependence is observed to be less pronounced. In particular, the

[23]The peaks of the inverse statistics distributions were coined *optimal investment horizons* by Simonsen et al. [36].

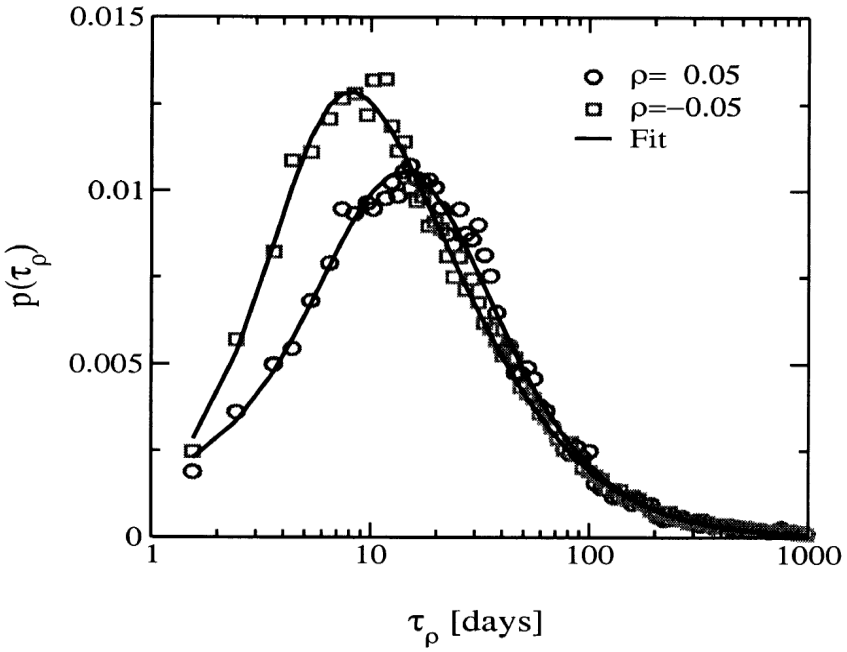



Figure 3.11: Inverse statistics distributions for the fixed return-level $\rho = \pm 0.05$ for the DJIA. Open symbols correspond to the empirical distribution and solid lines are fits to the generalized Gamma distribution (equation 3.81). Figure adapted from [38].

asymmetry starts to emerge at roughly $|\rho| \sim 10^{-2}$, and to saturate with a magnitude of about 200 days as compared to a few days for smaller returns. Correspondingly, the exponent $\gamma$ was observed to increase from nearly zero to about 1.8, where it seemed to saturate [38]. This implies that if a scaling regime does exist, data seems to favor $\gamma < 2$. Zhou et al. [39] reported that the exponent $\gamma$ seemed to depend on the specific market examined. They found that $\gamma$ obtained smaller values in emerging markets as compared to the more mature and liquid western markets, and speculated that the exponent is a measure for the maturity of the market in some sense.

Another interesting aspect of the gain-loss asymmetry is that it has been observed in several major stock indexes such as the DJIA and S&P 500, but not in the individual stocks the indexes are composed of [37]. However, later research has indicated that a weak asymmetry appears also for individual stocks, but only for sufficiently large return-levels [40, 50].

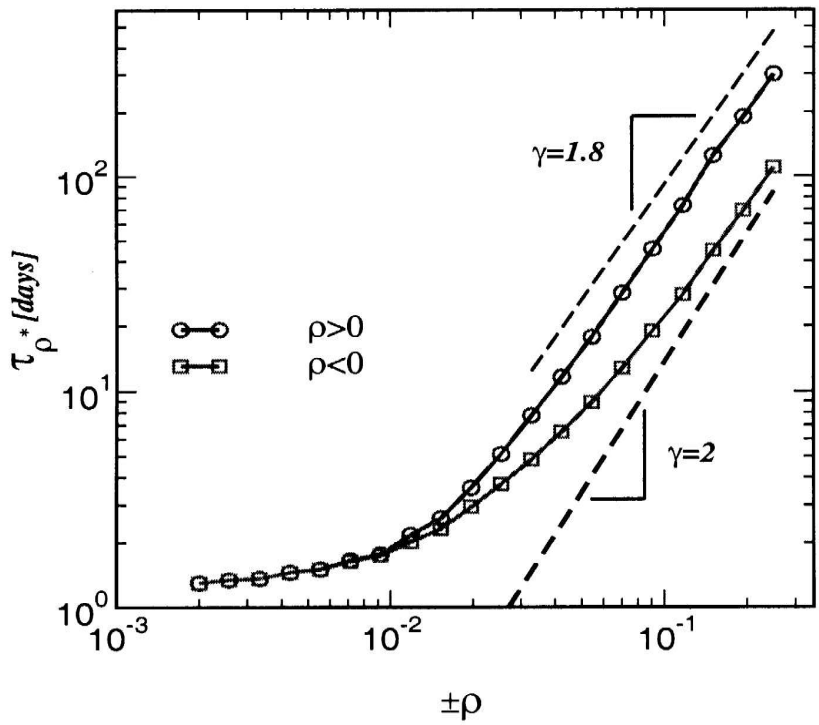



Figure 3.12: The optimal investment horizon $\tau_\rho^*$ for positive (open circles) and negative (open squares) return-levels $\pm\rho$, based on detrended historical data from the DJIA. The data has been detrended using wavelets. It is observed that the exponent $\gamma$ increases with increasing $|\rho|$ and that the asymmetry saturates at significantly large return-levels. The upper dashed line is the empirical found $\gamma \simeq 1.8$ for large returns, and the lower dashed line is the theoretically predicted $\gamma = 2$. Figure adapted from [38].

Reasons for this asymmetry have been speculated by several authors, but is still debated in literature [34, 38, 40, 41, 42, 44]. In the first publications reporting the gain-loss asymmetry, it was speculated that the reason of the phenomena was market dynamics such as collective effects between the individual stocks [37, 38]. In particular, it was speculated that negative signals could synchronize the price drops of individual stocks, causing the index to exhibit more dramatic drops. The financial crisis starting late 2008 is an example of such a collective effect. Nearly all stocks dropped simultaneously. The *fear factor model* was developed to investigate the gain-loss asymmetry [42], based on the key idea of enhanced stock-stock correlations during periods of falling markets. This model was constructed to explain the paradox of indexes showing gain-loss asymmetry and individual stocks

showing no well pronounced asymmetry. Shortly explained, the model introduced a collective *fear factor*, able to initiate the stocks of the model to all move downward while they at other times could move independently of each other. The model introduced the synchronization via simultaneous down-movements of all stocks at some time-steps, with a frequency given by the fear factor parameter $p$. Hence, at all time-steps there is a probability $p$ that all stocks move down synchronously. Similarly, with a probability $1 - p$ all stocks move independently, making random adjustments to their logarithmic price. To ensure that the individual stocks follow a geometric Brownian motion with no drift, the probability for a stock to move upwards, $q$, is slightly larger than the probability $1 - q$ for the stock to move downwards. This causes the abrupt down-movements to be compensated by an upward drift in the calmer periods between each of the synchronized draw-downs. Hence every day there are no synchronized movements of the stocks, the probability $q$ of an individual stock-price moving upward is slightly larger than the probability $1-q$ for it to move downward. Such a model will by construction not introduce any asymmetry into the individual stocks [42]. The model was also generalized by Siven et al. [43] to allow the market to remain in the distressed mode where the moves of the stocks are highly correlated for a longer period.

The results of the fear factor model are presented in figure 3.13. It is observed a gain-loss asymmetry similar to that observed for the DJIA, except from minor differences for small waiting times and the height of the optimal investment horizons. It was speculated that this was due to the lack of an *optimism factor*, decreasing the height of the maximum, widening the distribution towards smaller waiting times. However, Donangelo et al. [42] concluded that the gain-loss asymmetry indeed could be introduced by single stock synchronization. It has also been questioned if the leverage[24] effect could be of same origin as the gain-loss asymmetry. To answer this, Ahlgren et al. [44] introduced the *frustration governed market model*. The model was found to reproduce the empirical found gain-loss asymmetry excellently, but there were special cases where the model produced leverage but no gain-loss asymmetry. Their conclusion was that the two phenomena not necessarily were of same origin. Siven et al. [50] on the other hand concluded that there was a temporal dependence structure present in stocks, closely related to that giving rise to the leverage effect. This was based on a modification of the retarded model of Bouchaud et al. [51], where the absolute amplitude of the price changes does not follow the price level instantaneously (as is assumed with geometric Brownian motion) but rather was concluded to follow

[24]The correlation between future volatility and past return, with time lag $\tau$ [44, 51].

a moving average of the price over the past few months. However, further research has supported the fear factor model, as empirical studies indicated that the correlations between stocks are stronger during market drops than during market rises [40]. The strength of correlations between stocks during falling and rising markets is to be further studied in this thesis, making use of random matrix theory as explained in section 3.7 to extract the significant correlations between stocks during different modes of the market.

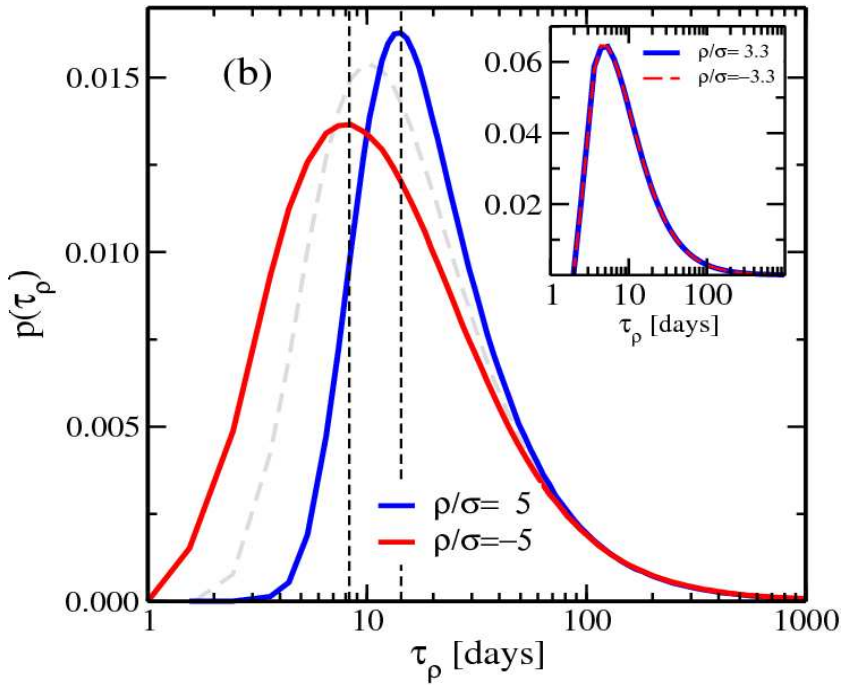



Figure 3.13: Results for the inverse statistics obtained from the fear factor model for a return-level $|\rho| = 0.05$. Comparing to figure 2.1, it is observed that the result is similar except from lower probabilities for short waiting times and the height of the peaks. The inset shows the gain and loss distributions for the individual stocks in the model. Figure adapted from [42].

# Chapter 4

# Dataset

This chapter contains information about the high-frequency dataset that is analyzed in this work, along with an introduction to stock indexes and their construction. A gain-loss asymmetry as discussed in section 3.8.2 is observed in the index, and presented for several return-levels. The gain-loss asymmetry is also investigated for individual stocks.

## 4.1 The dataset

The data have been obtained from Nykredit Bank in Denmark, but were originally recorded by Bloomberg. The data consist of individual stock quotes from 492 of the largest companies in Europe, starting December 1th 2010 lasting until 19th of April 2011. The stock quotes are updated every minute, which is why the dataset is called a high-frequency dataset. The ticker of the companies as well as the country they are based and their GICS[1] sector name are presented in appendix A.1, table A.1. Raw-data obtained from Nykredit Bank only contained values for minutes where there had been trading activity. Hence, price-series of liquid stocks were longer than those of the illiquid ones. Therefore, the data was modified into series of equal length, where the stock price was set to equal the last traded price for minutes without any trading activity. The resulting set contains 50568 price quotes for each stock.

## 4.2 Indexes

A stock index measures the performance of a specific section of a market. There exist several different indexes measuring the performance of a given category of stocks, typically defined in terms of geography (country, con-

[1]GICS is shorthand for Global Industry Classification Standard [56].

tinent, world), market capitalization (small cap, mid cap and large cap) or economic sector (e.g. technology, transportation, banking). Indexes are computed and published either by stock exchanges such as NYSE or Euronext, by publishing companies such as Dow Jones, McGraw Hill (S&P) or Financial Times, or by investment banks such as Morgan Stanley [17, pp. 72]. As the indexes measure the performance of specific sectors, they are used as benchmarks for the performance of individual stocks or portfolios.

Examples of indexes are the global Dow Jones Global Indexes, national indexes such as the Dow Jones U.S Indexes, and more sector specific indexes such as Dow Jones Sector Indexes including the Dow Jones U.S. Select Aerospace & Defense Index and the Dow Jones U.S. Select Regional Banks Index [54].

An index consists of several companies selected from the markets and sectors taken into consideration, weighted in a specific way. There are two main ways of weighting an index: (i) price weighting and (ii) capitalization weighting. In a price-weighted index, the price of each stock is the only factor determining the index value. This way of weighting an index completely ignores the market cap of the component companies, causing large fluctuations in small companies to heavily influence the index value. A capitalization-weighted index takes into account the market cap of its component companies, hence smaller companies are weighted less and do not influence the index value as much. Capitalization-weighting of indexes is most common, but some well-known indexes such as the Dow Jones Industrial Average (DJIA) are price-weighted[2].

In this work, a price-weighted index is constructed based on the Bloomberg data. The fact that the 492 companies belong to different countries in Europe and hence are listed in different currencies must be taken into consideration. Most of the companies are listed in euros, but some are listed in pounds sterling or even different Nordic currencies. This may lead to bias in the index, as the exchange rates between the currencies fluctuate and therefore can cause increased stock-price fluctuations and correlations[3]. It is reasonable to

[2]The reason why the DJIA is price weighted is purely historical. When the index was created, computations were done manually. Summing the price of its constituent stocks was the easiest way of calculating the index value, which is why it was made as a price-weighted index [17, pp. 73].

[3]Assume two companies listed in different currencies are strongly anti-correlated. If the respective currencies also anti-correlate and this is not taken into consideration, it might actually cause the companies to appear as correlated depending on the strength of the original anti-correlations.

assume that e.g. a weakening of the pound with respect to other currencies can lead to an increase in the stock-price of companies listed in pounds. The mechanisms behind this are complex, but one of the mechanisms is that many companies have earnings in other currencies than the currency they are listed in.

An additional feature that must be considered when constructing an index is that stock-prices listed in euros or pounds normally are of lower absolute value compared to stocks listed in Nordic currencies. This is reasonable, as the rate between a Norwegian krone and the euro (pound) is approximately 8 (9) to 1. This imbalance causes stocks of higher absolute value to influence an index more than stocks of lower absolute value. It follows that stocks listed in "weak"[4] currencies will be weighted heavier than they should. This problem would vanish if high-frequency currency exchange rates were available, such that all stock-price series could be converted into the same currency. As this kind of currency exchange data could not be obtained for the work in this thesis, all 492 stock-price series were instead divided by their median[5] value, causing the stock-price series to start from a more similar starting point. This is believed to be a good approximation, as it is reasonable that currency fluctuations are not very large within the less than 6 month window covered by the high-frequency data used in the current work.

A price-weighted index is constructed according to [17, pp. 73],

$$I(t) = \frac{1}{d(t)} \sum_{i=1}^{N} \tilde{S}_i(t), \tag{4.1}$$

where $d(t)$ denotes the divisor[6] of the index at time $t$ and $\tilde{S}(t)$ is the stock-price divided by its median. In price-weighted indexes such as the DJIA, this divisor is updated periodically to offset the effect of any changes in the component stocks such as splits, dividend payouts and other factors such as bonus issues etc. This is done to ensure that the value of the index with new (or modified) constituents equals that of the index before the modification. If this is not done, the index value is not kept constant and it is not a proper measure of performance. As the dataset used in this work covers a period of about 6 months, it is reasonable to fix this value to a constant, $C$. This follows, as it can be assumed not to be too much changes in the component

[4]Weak is here not describing the properties of the currency itself, but the currency conversion rate into other currencies.

[5]The median is not as affected by potential errors or spikes in the data.

[6]This divisor is chosen such that the index equals a certain reference value at a certain date. In this work, it has been set such that the constructed index starts at 100 when the stock exchange opens at 9AM the 1st of December, 2010.

stocks. Hence, the divisor is equal to this constant for all times, $d(t) = C$. The constant $C$ is chosen such that the index starts at a value of 100. Figure 4.1 presents the resulting index.

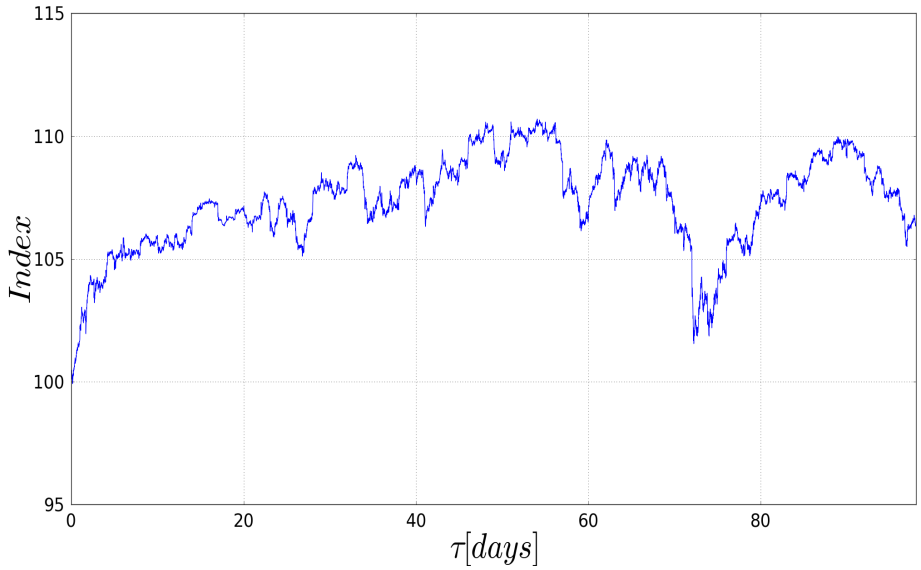



Figure 4.1: Price-weighted index based on the high-frequency data, constructed following equation 4.1. The index is normalized into starting from 100. One trading day is between 09:00 and 17:35, lasting 515 minutes.

Real markets have imperfections that are reflected in the index. The most pronounced of these effects is the overnight effect, arising from the fact that stock exchanges are only open a limited period of the day. News released when the market is closed cannot affect the market price immediately, but are rather accumulated and lead to a gap between the previous closing price of the exchange compared to todays opening price. This effect has been reported to be significant in terms of volatility, as the overnight contribution to the volatility has been found to equal nearly a quarter of the total daily volatility [17, p. 85]. In the case of high-frequency data, it is observed that the effect is strong, as price changes from minute to minute naturally are much less than intraday changes. There are also other imperfections such as variable trading activity, depending on the time of the day with a minimum around lunchtime and peaks around opening and closing time, leading the volatility to follow a $U$-shaped daily pattern [57, pp. 245].

## 4.3 Gain-loss asymmetry

Earlier work on inverse statistics and the gain-loss asymmetry have considered daily closure prices from different indexes [34, pp. 253]. However, an analysis of high-frequency data with observations every minute is to the best of our knowledge never performed before. Therefore, the inverse statistics distributions for both the index itself, as well as some of its constituent stocks are presented and discussed in this section.

### 4.3.1 Gain-loss asymmetry in the index

As stock markets have a drift due to the overall growth in the world economy, the index should be detrended before calculating any inverse statistics distributions. If this is not done, the presence of drifts can lead to shifts in the positions of the optimal investment horizons. A positive drift will cause the optimal investment horizon for positive returns to shift towards shorter waiting times, as there in this case will be an increased frequency of positive returns. Oppositely, the optimal investment horizon for negative returns is shifted towards longer waiting times [40]. Earlier analysis of inverse statistics distributions have used long time-series of daily closure prices, as this is easily (and freely) available online. One example is the DJIA data used by Jensen al. [38], covering 11 decades from 1897 to 2001. To remove the trend, they used a technique known as wavelet filtering. This technique is useful, having the advantages of being non-parametric and not dependent on any economic assumptions.

The index used in this work consists of 1 minute data over a period of less than 6 months, and it is reasonable to assume that drift is not an important component as compared to its importance in the daily data. Therefore, no detrending is done on the data in this work. Based on the index, inverse statistics distributions are calculated for both positive and negative return-levels $|\rho|$, where the return-level is expressed in terms of the minutely standard deviation $\sigma$ of the index log-returns. Expressing the return-level this way is more correct, as standard deviations typically are higher for log-returns of individual stocks than for indexes. This will become more clear in section 4.3.2.

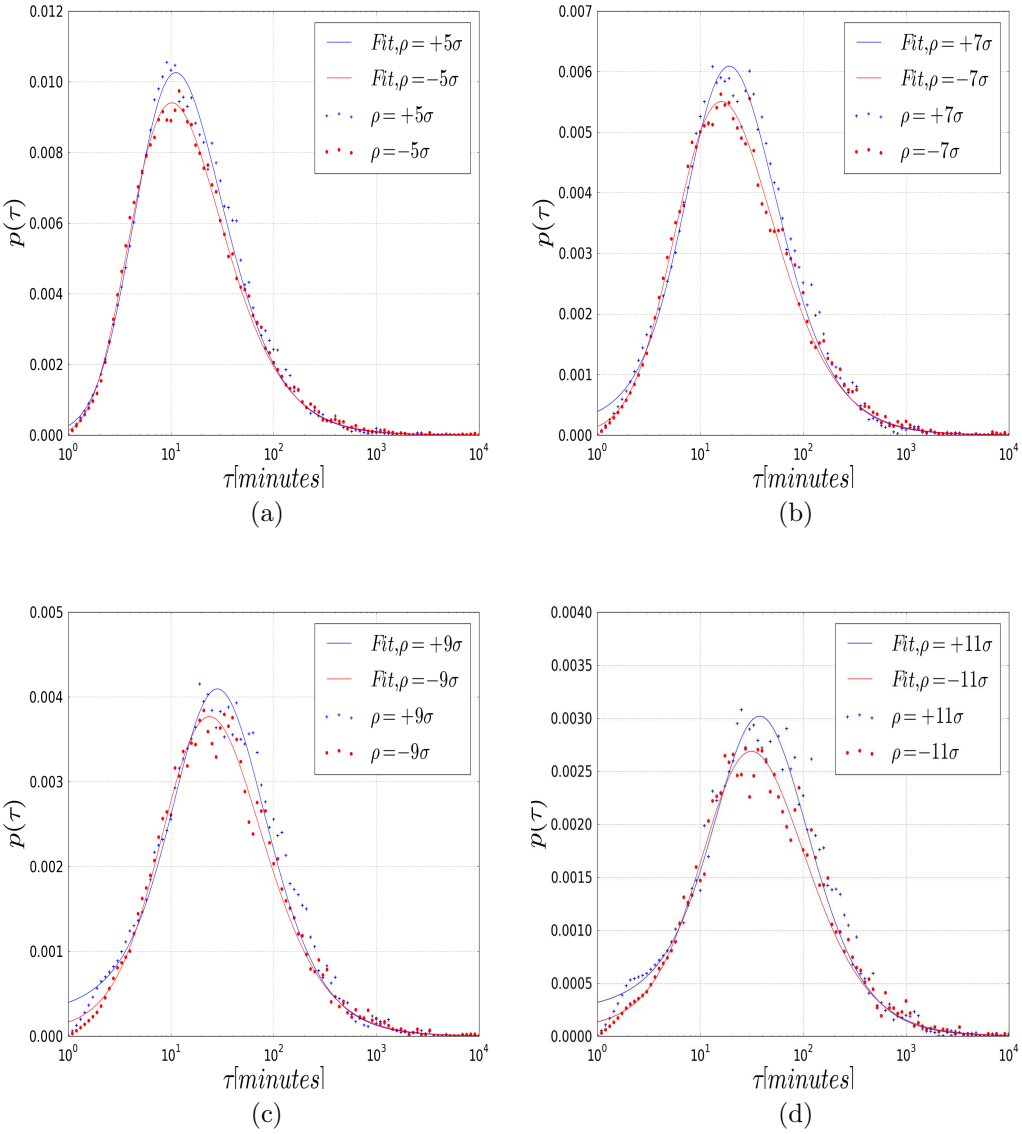



Figure 4.2: Inverse statistics distributions for the constructed index (red and blue dots). The return-levels are (a) $|\rho| = 5\sigma$, (b) $|\rho| = 7\sigma$, (c) $|\rho| = 9\sigma$ and (d) $|\rho| = 11\sigma$, where $\sigma \approx 0.027\%$ is the minutely standard deviation of the index log-returns. Solid lines represent the least squares fit of equation (3.81) to the empirical data, with $\tau^*_{\pm|\rho|}$ and parameters $\nu, \alpha, \beta$ and $t_0$ presented in appendix B.1, table B.1. All distributions are normalized.

The resulting inverse statistics distributions are fit to the generalized Gamma distribution, which is found to parameterize the data very well. Results for four different return-levels are presented in figure 4.2. As observed in other indexes, all distributions of figure 4.2 show a rather well defined and pronounced maximum, followed by fat tails for long waiting times. Fat tails indicate a nonzero probability of large waiting times, and are believed to reflect periods where the market is either relatively calm or going slowly downwards (upwards) before rising (dropping) again. The short waiting times around the maximum reflect more volatile periods with stronger fluctuations. Due to the higher probability of these events, such modes with an increased volatility seem to be more common [36]. None of the distributions are invariant under change of sign of the return-level $|\rho|$, and hence a gain-loss asymmetry is present for all four return-levels. Note that the tail exponent $\alpha + 1$ (discussed in section 3.8.1) is found to be indistinguishable from the "random walk" value 3/2 for all considered return-levels.

Figure 4.2 suggests that the optimal investment horizon for a gain is longer than the optimal investment horizon for a loss. In other words, it is faster to loose money than to earn money. The asymmetry is not significant for the smallest return-level $|\rho| = 5\sigma$, where it is observed to equal 1 minute. However, it increases with the return-level and is observed to equal roughly 6 minutes for the return-level $|\rho| = 11\sigma$. For the largest return-level considered in this work, $|\rho| = 16\sigma$, the asymmetry was found to be approximately 15 minutes. Larger return-levels than this are not considered, as the statistics in this case becomes too poor. The fact that the asymmetry is not significant for the smallest return-levels, and that it increases with the return-level, is in accordance with earlier observations by Jensen et al. [38] based on DJIA daily closure prices. In this section, the inverse statistics distributions are only presented for four distinct return-levels. However, they were calculated for $|\rho|/\sigma \in [1, 16]$. For each return-level, the positions of the optimal investment horizons $\tau^*_{\pm|\rho|}$ were recorded. These are presented in figure 4.3 as a function of $|\rho|/\sigma$; the return-level scaled by the minutely standard deviation of the index log-returns. As also observed by Jensen et al. [38] and discussed in section 3.8.1, small or no asymmetry is apparent for the smallest return-levels. For larger levels of return, the asymmetry emerges. The specific dependence between the asymmetry and the return-level is unknown, but for larger return-levels it seems to saturate, following the power law $\tau^*_{\pm|\rho|} \propto |\rho|^\gamma$ with $\gamma \approx 1.65$ for levels of return larger than about $|\rho| = 6\sigma$. However, the statistics is too poor to draw any conclusion on a power-law behavior, but it is noted that also Jensen et al. [38] observed $\gamma < 2$, based on DJIA daily closing prices.

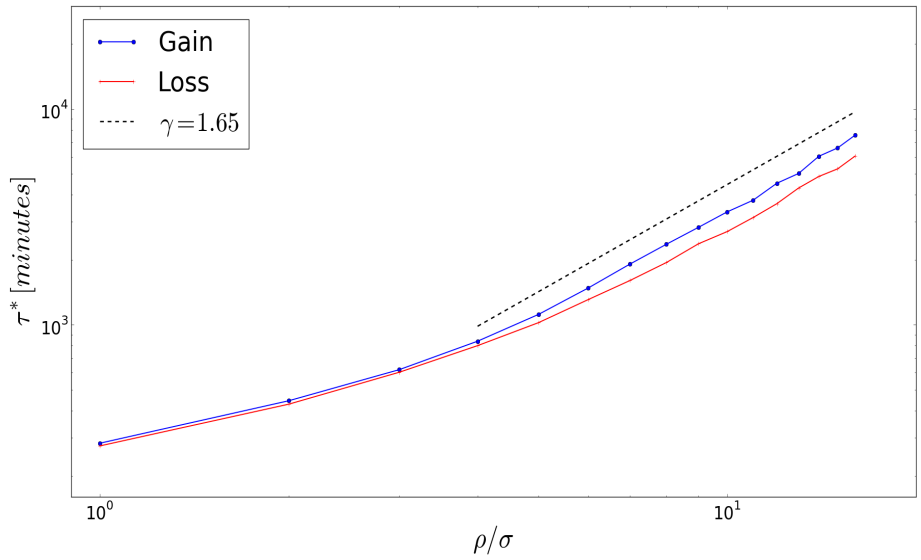



Figure 4.3: The dependence of the optimal investment horizons $\tau^*_{\pm|\rho|}$ on the return-level $|\rho|$, where the return-level is scaled by the minutely standard deviation of the index log-returns. Blue line with circles (red line with crosses) corresponds to positive (negative) levels of return. If a pure geometric Brownian price process is assumed, $\tau^*_{\pm|\rho|} \propto |\rho|^\gamma$ with $\gamma = 2.0$ for all values of $\rho$. Empirically, it is found that the scaling behavior is not $\gamma = 2.0$, but $\gamma \approx 1.65$, and only for large levels of return. This scaling behavior is indicated by the black dotted line.

As already discussed, Jensen et al. [38] considered daily closure prices of the DJIA, and found that the time it took to achieve a loss was shorter than the time it took to achieve a gain of the same magnitude. The result from the high-frequency data is the same, the expected time for a gain is longer than that of a loss. However, Jensen et al. [38] found the probability of a loss to be higher than that of a gain. The high-frequency data, however, seem to suggest the opposite, as the probability of a gain is higher than the probability of a loss for all four cases presented in figure 4.2. The same is also observed for the other return-levels considered. Intuitively, one could speculate that the observed higher probability of the gain optimal investment horizon arises from the normalization procedure. The fits are performed on normalized empirical distributions, and it is therefore important to be sure that these contain all relevant information. Now, assume that the gain distribution

has a fatter tail[7] than the loss distribution. If parts of the tails are ignored when calculating the empirical distributions, the normalization in this case will increase the ratio between the gain and the loss optimal investment horizons. This can lead to poor statistics when performing the fits. Whether the empirical tails actually are significantly different can be calculated by solving the following integral,

$$\int_{10^4}^{\tau_{\max}} \left[ p_{\text{loss}}^{\text{empirical}}(\tau) - p_{\text{gain}}^{\text{empirical}}(\tau) \right] \, \mathrm{d}\tau, \tag{4.2}$$

where the lower limit is set to $10^4$ for simplicity. The upper limit $\tau_{\max}$ equals the maximal length of waiting times, and is limited only by the length of the time-series used. The result is that the difference in probability accounted for by the two tails is less than 0.05% for the four cases in figure 4.2, not nearly enough to cause any significant bias when normalizing the empirical distributions. However, to ensure the empirical distributions to contain all information, they are calculated for all possible waiting times $\tau \in [1, \tau_{\max}]$ in this work. Note that in figure 4.2, the distributions are only presented for $\tau \in [1, 10^4]$, as it is the first part of the distributions that is most interesting. The conclusion is that the normalization procedure cannot be the origin of the observed higher gain probability. It can be discussed whether trends in the index can be a factor causing the observed higher probability for the gain optimal investment horizon. However, the time covered by the data used in this work is less than 6 months, such that it is hard to determine whether there are any clear trends present at all. Hence, neither potential trends can be the cause of the observed higher probability of gains. This leads to the conclusion that the observed higher probability for gains must be a property of the high-frequency nature of the data.

### 4.3.2 Gain-loss asymmetry for individual stocks

As discussed in section 3.8.2, the gain-loss asymmetry has been difficult to observe for individual stocks. One of the first publications on the gain-loss asymmetry actually reported it to be present in indexes, but *not* in individual stocks [37]. However, this was found not to be true, as Siven et al. [50] observed the asymmetry also in individual stocks. The reason why Johansen et al. [37] did not observe any asymmetry in individual stocks was that they expressed the return-level as a percentage, instead of expressing it in terms of the standard deviation of stock log-returns. As standard deviations of log-returns typically are larger for individual stocks than for indexes, it follows that expressing the return-level as a fixed percentage leads to a discrepancy

[7] Note that both tails fall as a power-low with $\tau^{-3/2}$. However, the tails also have an amplitude determined by three other parameters ($\nu$, $\beta$ and $t_0$) that can affect them.

between indexes and individual stocks when calculating the inverse statistics distributions for the same return-level. To reproduce the results of Johansen et al. [37], the inverse statistics distributions for two companies in the index have been calculated. Note that the return-level used is the same return-level as was used to produce the distributions of figure 4.2b. In other words, the return-level of the stocks is set to $|\rho| = 7\sigma_I$, where $\sigma_I$ is the minutely standard deviation of the index log-returns. The companies are XTA (Xstrata PLC) and ABBN (ABB Ltd), and the results are presented in figure 4.4.

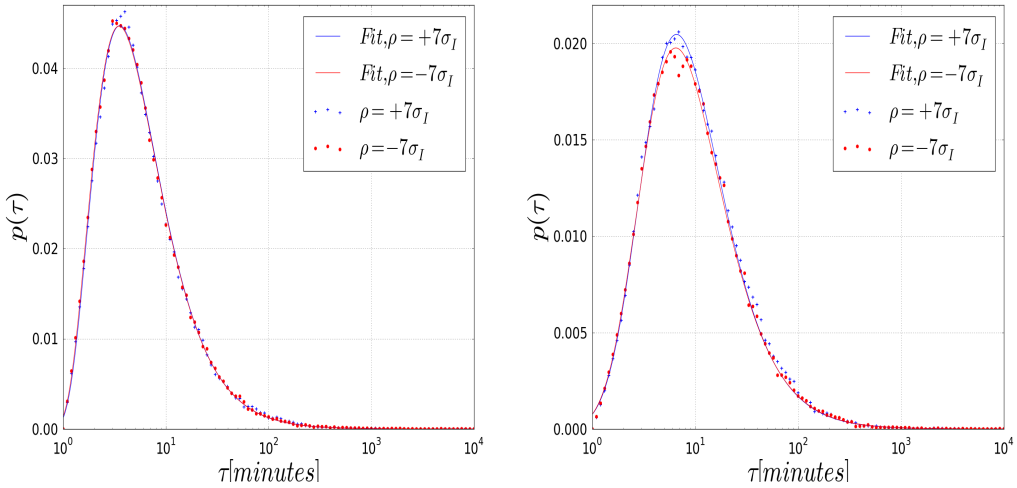



Figure 4.4: Inverse statistics distributions (red and blue dots) for the individual stocks (a) XTA and (b) ABBN. The return-level is the same for both stocks, $|\rho| = 7\sigma_I$, where $\sigma_I \approx 0.027\%$ is the standard deviation of the minutely index log-returns. Solid lines represent the least squares fit of equation (3.81) to the empirical data, with $\tau^*_{\pm|\rho|}$ and parameters $\nu, \alpha, \beta$ and $t_0$ presented in appendix B.1, table B.3. All distributions are normalized.

It is clear that the stocks show no or small asymmetry, in accordance with observations of Johansen et al. [37] for daily closure prices. Figure 4.5 presents the inverse statistics distributions for the same companies with same return-level as in figure 4.2b, only that the level is expressed in terms of the standard deviation of the log-returns of the stocks themselves instead of the index. This is also how Siven et al. [50] observed the asymmetry in individual stocks. It is clear that both stocks in figure 4.5 now exhibit a weak gain-loss asymmetry. This is exactly what one would expect, as individual stocks normally fluctuate more than the index. The asymmetry is also

found to increase when the return-level is further increased. The conclusion is that in accordance with earlier observations [50], the gain-loss asymmetry is apparent also in individual stocks.

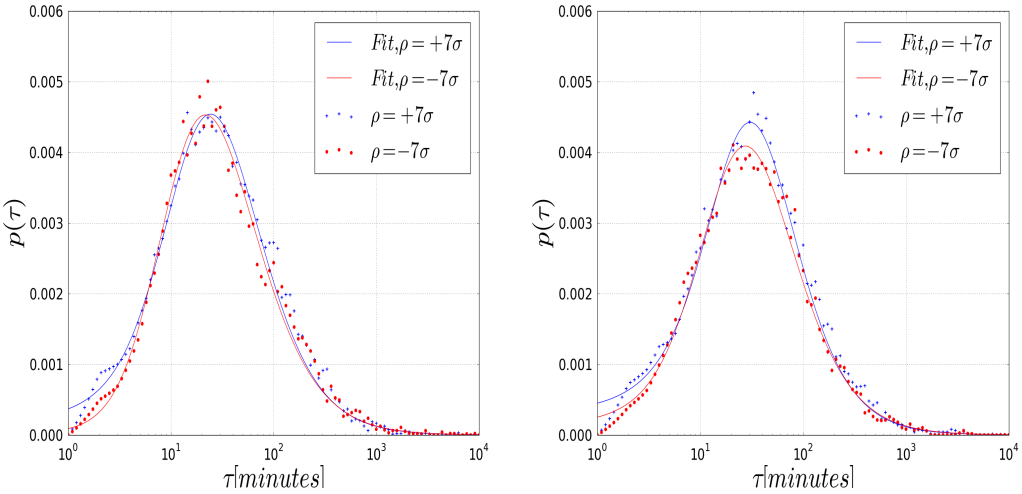



Figure 4.5: Inverse statistics distributions (red and blue dots) for the individual stocks (a) XTA and (b) ABBN. The return-level for both stocks is $|\rho| = 7\sigma$, where $\sigma$ is the minutely standard deviation of the log-returns of the respective stocks: $\sigma_{\text{ABBN}} \approx 0.067\%$ and $\sigma_{\text{XTA}} \approx 0.11\%$. Solid lines represent the least squares fit of equation (3.81) to the empirical data, with $\tau^*_{\pm|\rho|}$ and parameters $\nu, \alpha, \beta$ and $t_0$ presented in appendix B.1, table B.3. All distributions are normalized.

However, the asymmetry is harder to observe in individual stocks than in indexes. It follows that the asymmetry in the index cannot be caused by the asymmetry observed in individual stocks, as this is so much weaker for individual stocks. As discussed in section 3.8.2, the gain-loss asymmetry in the index is believed to arise from a collective effect where the stock-prices are changing in a much more correlated manner when the index is falling [37, 38]. This is also what was indicated by Balogh et al. [40], conducting a set of statistical tests on the DJIA index and its constituent stocks. This will be investigated further in this work, using principal component analysis and random matrix theory to monitor the strength of collective trends in the market.

# Chapter 5

# Method

This chapter contains a description of the procedure followed when analyzing the high-frequency data introduced in section 4.1. In short, an eigenvalue decomposition of a set of correlation matrices is performed to obtain a set of uncorrelated basis vectors. As will be discussed in chapter 6, one of these eigenvectors is found to describe the market excellently. This is the eigenvector corresponding to the largest eigenvalue, also known as the *first principal component*. In chapter 6, it is observed that the largest eigenvalue can be used as an index describing the strength of correlations in the market, and it is therefore considered in detail in chapter 7.

## 5.1 Calculating the density of eigenvalues

The density of eigenvalues $\rho(\lambda_C)$ of the correlation matrix is computed based on time-windows containing 1548 observations from each of the 492 stocks. The number 1548 is chosen to correspond roughly to that of Laloux et al. [24], using 406 stocks and 1306 observations, and corresponds to exactly 3 trading days. The data are originally prices of 492 stocks updated every minute. In order to analyze the fluctuations of the stocks, their log-returns $r_k$ are calculated according to equation (3.79), which is rewritten here for simplicity,

$$r_i(t_k) = s_i(t_k) - s_i(t_{k-1}), \tag{5.1}$$

where $s_i(t_k) = \ln S_i(t_k)$, and $S_i(t_k)$ is the price of stock $i$ at time $t_k$. Following, every log-return is standardized according to

$$\tilde{r_i}(t_k) = \frac{r_i(t_k) - \bar{r_i}}{\sigma_i}, \tag{5.2}$$

where $r_i(t_k)$ is the log-return of stock $i$ at time $t_k$, $\bar{r}_i$ is the average log-return for stock $i$ and $\sigma_i$ is the standard deviation of the log-returns of stock $i$ (i.e. the volatility). This procedure leads to $N = 492$ time-series of standardized log-returns having zero mean and unit variance. The procedure for analyzing the data is as follows:

- From the data, a set of time-windows of length 1548 minutes is created. This is done by first creating a time-window containing the first 1548 minutes. This is slid through the data, using discrete steps $X$. It follows that the first time-window consists of the first 1548 minutes, the second of the minutes in the interval $[X, X + 1548]$, etc. For reference, let the time-window be denoted as $\mathbf{M}_k$, where the index $k$ denotes how many steps that has been taken, starting from 0.

- Correlation matrices $\mathbf{C}$ are calculated from each of the time-windows $\mathbf{M}_k$. These are correlation matrices, not covariance matrices, as the standardization procedure described in equation (5.2) causes their elements to all lie in the interval $C_{ij} \in [-1, 1]$. A value $C_{ij} = 1$ corresponds to a perfect correlation, while the value $C_{ij} = -1$ corresponds to perfect anti-correlation.

- Eigenvalues and eigenvectors of the correlation matrices $\mathbf{C}$ are calculated and sorted after size, and the densities of eigenvalues $\rho(\lambda_C)$ are calculated.

This procedure of a sliding time-window is graphically illustrated in figure 5.1 for two different step-sizes $X$. Figure 5.1a illustrates the step-size $X = 1548$ minutes, while figure 5.1b illustrates a smaller step-size. In this work, the largest eigenvalue has been calculated using values $X \in [10, 1548]$. For each of the computed correlation matrices, the density of eigenvalues is calculated according to [24]

$$\rho(\lambda_C) = \frac{1}{N}\frac{\mathrm{d}n(\lambda_C)}{\mathrm{d}\lambda_C}, \tag{5.3}$$

where $n(\lambda_C)$ is the number of eigenvalues of $\mathbf{C}$ less than $\lambda_C$. An analytical derivation of the density of eigenvalues on the other hand can be found in section 3.7.1. The resulting figures presented in chapter 7 consist of this empirical density of eigenvalues, as well as the theoretical predicted density for a suitable value for the volatility of the random part, $\sigma^2$. As will be discussed in chapter 7.1, this volatility can actually be treated as an adjustable parameter. This follows, as the trace of the correlation matrix must be kept constant.

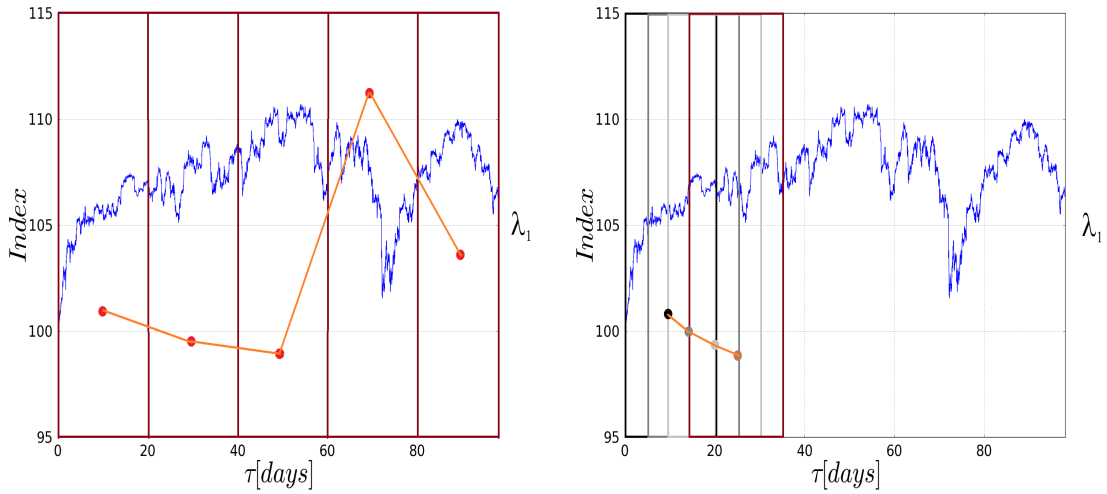



Figure 5.1: (a) Illustration of how the data is divided into time-windows. Red dots correspond to observations of $\lambda_1$ (the largest eigenvalue of the correlation matrix) from each time-window. (b) Illustration of how the time-windows are slid through the data, leading to a higher frequency of the observations of $\lambda_1$. Dots of various colors origin from the observations of $\lambda_1$ from overlapping time-windows drawn in same color. The orange line in both figures is an interpolation between the observations of $\lambda_1$ based on the different time-windows. The illustration of time-windows and observations of $\lambda_1$ is superimposed onto the index, as the eigenvalues are calculated based on the stocks in the index.

After the set of eigenvalue densities is calculated, the densities are compared to theoretical densities based on a *null hypothesis* purely random matrix. As discussed in section 3.7.1, the eigenvector corresponding to the largest eigenvalue has been observed to describe the market excellently. This will be investigated in chapter 6 for the current dataset, but the conclusion is that it does describe the market excellently. It is also observed that the largest eigenvalue can be used as an index, monitoring the strength of collective trends in the market. It follows that the *if* the largest eigenvalue is significantly larger than the predictions from random matrix theory, it implies the existence of true correlations between stocks in the market. This is indeed what is observed. Therefore, by comparing the largest eigenvalue to the index, it is possible to reveal if there are any connections between the two quantities and thus between the strength of collectivity and the index.

## 5.2 Analyzing the composition of the eigenvector corresponding to the largest eigenvalue

By performing an eigenvalue decomposition of the correlation matrix, a new and uncorrelated basis is calculated. This basis consists of the *principal components*, corresponding to the eigenvectors of the correlation matrix. The *first principal component* is the eigenvector corresponding to the largest eigenvalue.

Each eigenvector is a recipe of a specific portfolio, where element $i$ of the eigenvector corresponds to the weight of stock $i$ needed to construct the specific portfolio. A correlation matrix based on 492 stocks leads to 492 orthogonal eigenvectors and thus 492 uncorrelated portfolios. As the first principal component is assumed to correspond to the market, this is investigated in chapter 6. If the first principal component describes the market, it should not contain any negative elements. This follows, as the market cannot contain short-positions[1], such that all elements should be strictly positive. It should also have sectors weighted equally to those of the market, which can be investigated by sorting the weights after GICS sectors.

It is also discussed how the size of the eigenvalues are connected to the amount of information they carry. This is important, as the largest eigenvalue will be used as an index describing the strength of correlations in the market in this work.

[1]A short-position is taken when an investor borrows stocks and sells them. Before the stocks are returned to their owner, the stock-price may increase or decrease. The investor earns money if the stock-price falls such that the stocks can be bought back at a lower price before the are delivered back to the owner.

# Chapter 6

# Eigenvalues of the correlation matrix

Eigenvalues of correlation matrices are the corner-stones of the discussion in chapter 7. Hence, it is important to understand their relation to the market and how their size is connected to the strength of collective phenomena. This is investigated and discussed in the current chapter.

## 6.1 Eigenvalues and eigenvectors of a correlation matrix

The change of basis from a possibly correlated basis into a new and uncorrelated basis is called a *principal component analysis*, and is exactly what is done in this work. The original and possibly correlated basis is changed into a new and uncorrelated basis, where the new basis is the orthogonal eigenvectors of the correlation matrix. These are also known as *principal components* and, as discussed in section 3.6.1, describe uncorrelated portfolios based on possibly correlated stocks.

It was also discussed that eigenvectors corresponding to the largest eigenvalues are associated with the most correlating stocks in empirical data. Such clusters of correlated stocks form economic sectors, as stocks within a single economic sector intuitively exhibit significant correlations compared to stocks from different sectors. As some factors also influence all stocks, the market is represented by one of the eigenvectors. This has been observed to be that corresponding to the largest eigenvalue [24], or in other words the *first principal component*.

As a start, it is interesting to discuss how a correlation matrix behaves for the two extreme cases of $N$ perfectly correlated and $N$ completely uncorrelated time-series [58]:

- For $N$ perfectly correlated time-series: An $N \times N$ matrix of ones (a matrix where all elements are equal to one).
- For $N$ perfectly uncorrelated time-series: An $N \times N$ identity matrix (having elements along the diagonal equal to one and all other elements equal to zero).

Notice that this requires that the two series are infinitely long, such that the noise vanishes. Solving the characteristic equation, equation (3.48), yields the eigenvalues of the two correlation matrices:

- The $N \times N$ matrix of ones: Only one non-zero eigenvalue equal to the dimension of the matrix, $\lambda = N$.
- The $N \times N$ identity matrix: $N$ degenerate eigenvalues, $\lambda_n = 1$.

The correlation matrix based on $N$ completely uncorrelated and infinitely long time-series would indeed be the identity matrix. However, as no real time-series are infinitely long, noise will be present in the matrix. Such random correlations are caused by random stocks correlating at random times, similar to what is observed in financial markets. Noise will also be reflected in the eigenvalues of the correlation matrix, causing them to be distributed around 1 rather than being a degenerate set of identical eigenvalues all equal to 1. This is probably close to what would be observed if examining stock-price series from stocks belonging to different economic sectors.

Many of the stocks considered in this work belong to the same sector. Hence, correlations between stocks within the same sector, as well as other correlations randomly arising form time to time are present. However, it is always true that a correlation matrix has all diagonal elements equal to 1, as all series are perfectly correlated with themselves. As the trace of such a matrix is invariant[1], the sum of eigenvalues must always be equal to the dimension $N$ of the matrix, causing what we know as eigenvalue repulsion. Correlations within the time-series cause some eigenvalues to be larger than others [58], and the only way to keep the sum of eigenvalues constant is if other eigenvalues decrease. This leads to a significantly increased distance between large

[1]The trace is given by $\mathrm{Tr}(\mathbf{C}) = \sum_{\mathrm{i}=1}^{\mathrm{N}} \lambda_{\mathrm{i}}$, and is invariant when the basis of the correlation matrix is changed. As the sum of the diagonal elements of a correlation matrix always equals N, the trace is equal to N.

eigenvalues and the bulk distribution of smaller eigenvalues. Müller et al. [58] investigated effects of correlations within a subset of time-series, and observed an interesting effect. The introduction of correlations to a subset $K$ of a total of $M$ time-series caused the largest eigenvalue to increase while the $K-1$ lowest eigenvalues decreased. In other words, $K$ eigenvalues react to the correlations between the $K$ time-series. Another observation was that the strength of the induced correlations as well as the dimensions of the correlated subsystem were the factors determining the amount by which the lower and upper eigenvalues would change.

It follows that eigenvalue sizes are strongly connected to correlations between a system's components, here the 492 stocks. Collective phenomena can be understood as the correlated behavior of a system's components, and these phenomena are believed to be the origin of the gain-loss asymmetry many indexes exhibit [34, 38]. It is therefore interesting to examine this further, approaching the problem somewhat differently than what was done by Balogh et al. [40]. In section 6.2, it is concluded that the eigenvector corresponding to the largest eigenvalue describes the market excellently. It follows that the largest eigenvalue is strongly connected to the strength of collectivity in the market. Therefore, the problem whether collective trends are stronger during falling than rising markets is approached using the largest eigenvalue as an index describing the strength of collectivity.

## 6.2 Composition of the largest eigenvectors

This section presents a discussion of the eigenvector corresponding to the largest eigenvalue, and how it is related to the market. Figure 6.1 presents the distribution of components of the eigenvectors corresponding to the three largest eigenvalues, $\lambda_1$, $\lambda_2$ and $\lambda_3$. Notice that all eigenvalues and eigenvectors considered in this section are those obtained from the first time-window, consisting of the first 1548 minutes of the data. In the following discussion, an eigenvector arising from eigenvalue $\lambda_i$ (where $i = 1$ corresponds to the largest eigenvalue) is denoted as $\mathbf{v}_i$.

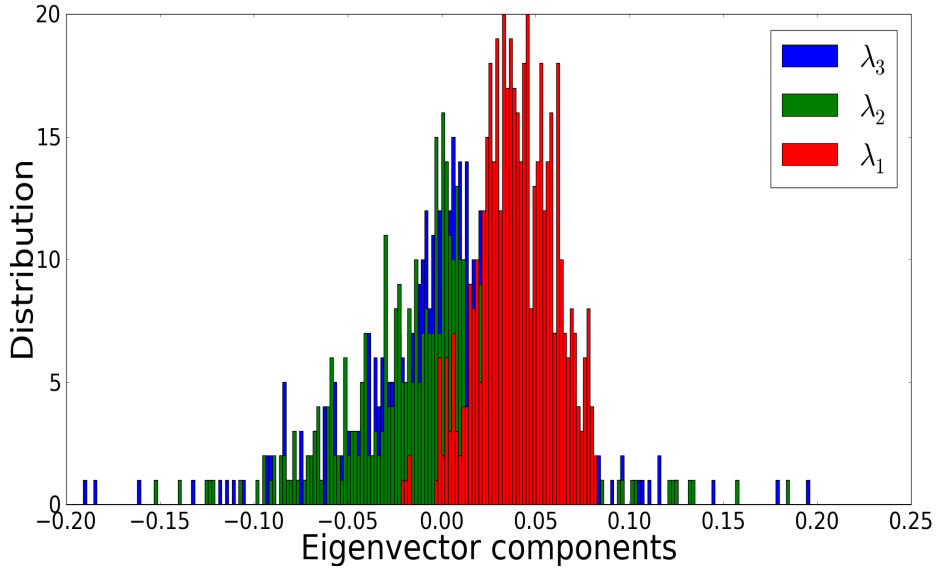
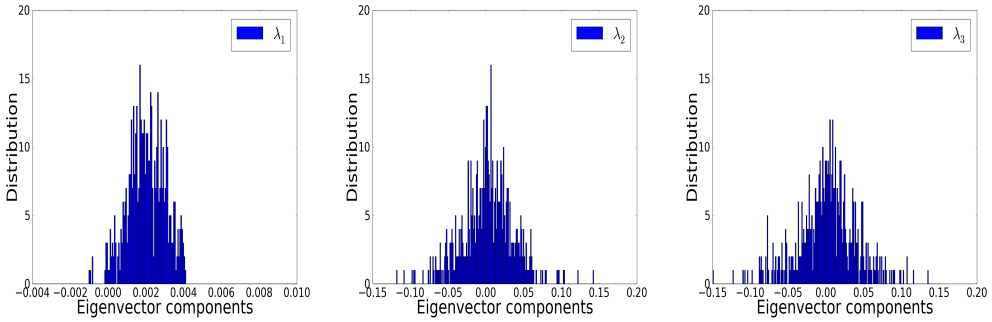



Figure 6.1: The histograms of non-normalized eigenvector components corresponding to the three largest eigenvectors of the correlation matrix calculated from the first time-window.

An inspection of figure 6.1 leads to the observation that components from $\mathbf{v}_1$ are *differently* distributed than components of $\mathbf{v}_2$ and $\mathbf{v}_3$. Figure 6.2 presents the same histograms in individual plots to underline this difference.




(a) (b) (c)

Figure 6.2: Individual histograms of components of the eigenvectors corresponding to the three largest eigenvalues.

It is clear that components of $\mathbf{v}_1$ are distributed with a positive mean, while components of $\mathbf{v}_2$ and $\mathbf{v}_3$ have approximately zero mean. This is consistent with the assumption that $\mathbf{v}_1$ describes the market, as a portfolio reproducing the market cannot hold any short positions. Figure 6.3 presents the portfolio described by the eigenvector $\mathbf{v}_1$, where weights (elements of the eigenvector) are sorted after GICS sector. Notice that the portfolio is normalized, having total value of 1 euro, as each weight is multiplied by the market capitalization (in euros) of its corresponding company before the whole portfolio is normalized.

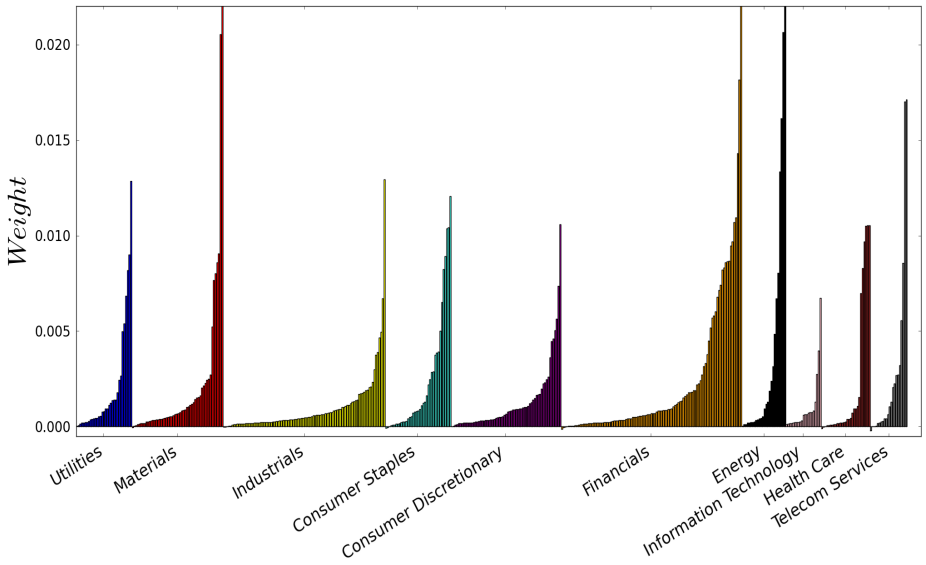
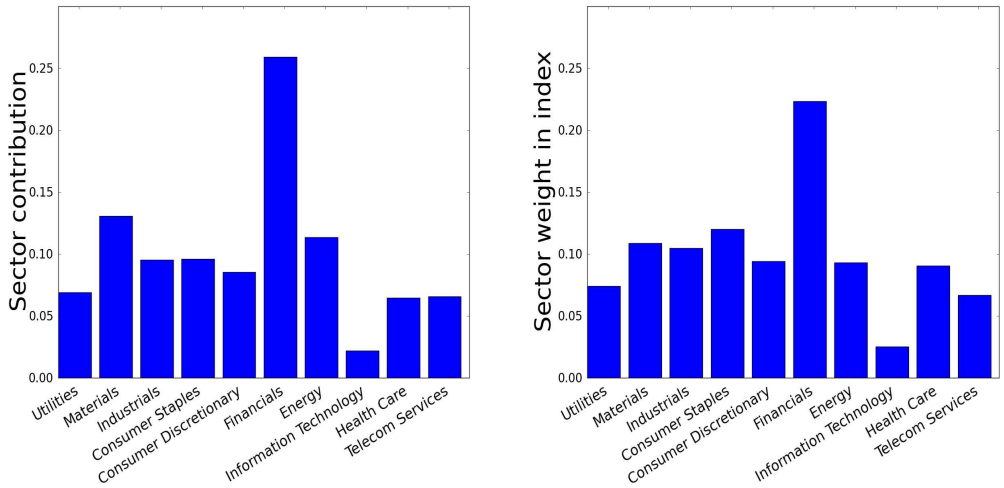



Figure 6.3: Illustration of the portfolio described by the largest (and normalized) eigenvector $\mathbf{v}_1$, where the market capitalization of the companies has been taken into account. Each column corresponds to the value of an individual stock in the portfolio, and the portfolio is of total value 1 euro. The portfolio is divided into sectors, as it is not possible to mark each of the 492 columns with its corresponding ticker.

As clearly observed from figure 6.3, the values of the components in the portfolio (the components of $\mathbf{v}_1$) are strictly positive, except from about 2% that are slightly negative. This is believed to arise due to currency fluctuations that are not taken into account when calculating the correlation matrices. As these negative components of the portfolio are very small compared to the other components, it seems reasonable that $\mathbf{v}_1$ represents the market. Also note that the elements of the largest eigenvector, before they are multiplied by the market capitalization of their corresponding stocks, are roughly

of same size. The final test is to check whether the different sectors are weighted similarly to their weighting in the market. This has been calculated, and is illustrated in figure 6.4. An inspection reveals that sectors in the portfolio described by $\mathbf{v}_1$ are weighted almost identically to how they are weighted in the market itself. Another observation is that some sectors are weighted more than others. As an example, the financial sector is weighted almost 10 times as much as the health care sector. This follows naturally from the fact that there are more financial companies in the current dataset, and that they often are of relatively high market capitalization. The conclusion is that the portfolio described by eigenvector $\mathbf{v}_1$, corresponding to the largest eigenvalue $\lambda_1$, indeed can be considered as the market itself. This has also been concluded also by other authors [24, 26, 27, 28]. It follows that $\lambda_1$ is strongly connected to the strength of collective trends in the market.




(a) Weighting of sectors in the portfolio represented by $\mathbf{v}_1$.

(b) Weighting of sectors in the market.

Figure 6.4: Figures (a) and (b) illustrates the weighting of the 10 GICS sectors in the portfolio described by the eigenvector $\mathbf{v}_1$ and the market, respectively.

# Chapter 7

# Results and discussion

This chapter is aimed at quantifying the strength of collective trends in the market, as this is believed to be the origin of the gain-loss asymmetry [37]. Balogh et al. [40] have found evidence for the existence of stronger collective trends during falling than rising markets, but their work contained several averaging procedures that in principle *can* affect their results. The approach in this chapter does not contain any averaging, and can therefore provide support to their results. The fact that a gain-loss asymmetry is observed also for an index consisting of high-frequency data is itself remarkable (see section 4.3.1), as one intuitively would believe that this effect would take some time to fully set in. This suggests that the market reacts fast to news and other updates, and that the time it takes for collective effects to set in is on the scale of minutes. It has been speculated that programmed trading based on relatively simple algorithms could be the reason behind the observed asymmetry, as this reacts almost instantaneously to stock-price changes. However, Balogh et al. [40] conclude that this produces symmetric correlations and therefore cannot be the origin of the asymmetry.

The chapter presents several empirical eigenvalue densities, calculated from periods where the market is either calm or exhibiting sudden drops or rises. These densities are discussed and compared, and it is observed several eigenvalues that deviate from the noise band[1] described by RMT. The largest of these is, as discussed in chapter 6, considered to describe the strength of collective trends in the market, and denoted as $\lambda_1$ in the following discussion. As the magnitude of $\lambda_1$ reflects the strength of collective trends in the market, its temporal dependence is compared to that of the index to see whether the two quantities are connected in some way.

[1]The bulk distribution of the eigenvalues arises from the random part of the correlation matrix, and can therefore be considered as noise.

## 7.1 The empirical density of eigenvalues

Eigenvalue densities are calculated according to the procedure described in section 5.1: A time-window of length $T = 1548$ minutes (3 trading days) is slid through the data, using discrete steps of $X$ minutes. The density of eigenvalues is calculated from each time-window[2], and compared to theoretical predictions from RMT. As will be seen, the result is an excellent fit to theory, except from some eigenvalues significantly larger than the theoretical predictions. As all densities seem to exhibit the same features, it is not necessary (nor possible, as several hundred densities are calculated) to present all of them here. However, eight plots representing different modes of the market are presented and discussed in this section.

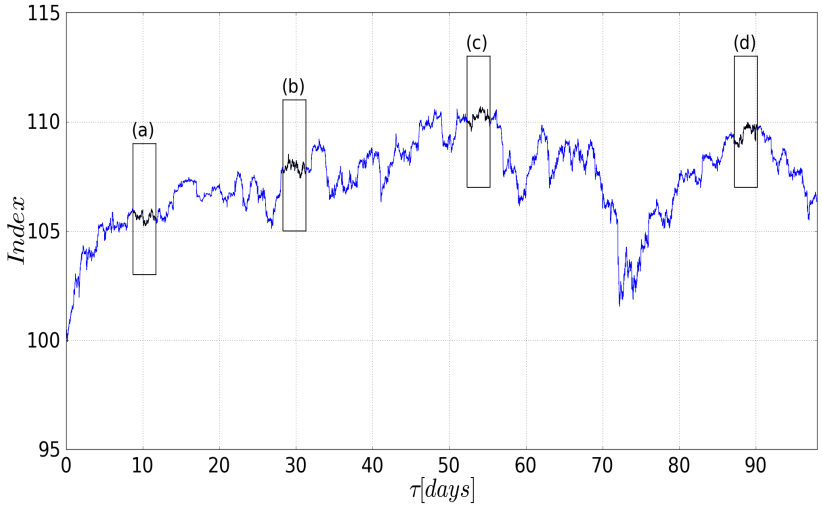



Figure 7.1: Blue line: The index. Black boxes: Time-windows where the index exhibits calm periods of low volatility, having fluctuations of varying sign. Empirical densities of eigenvalues based on these time-windows are presented in figure 7.3. The indices above the time-windows are included as a reference, and correspond to subfigure indices of figure 7.3.

[2]Note that the empirical eigenvalue densities are smoothened according to the procedure described in section 3.7.2. This does not affect the results, as it is the eigenvalues themselves that are considered when discussing the strength of collective trends in the market.

Consider the index constructed in section 4.2 and presented in figure 4.1. This is clearly observed to exhibit several periods where the index is rather calm, exhibiting a low volatility at least on short time-scales of a few trading days. Four such calm periods are marked in figure 7.1. Their corresponding eigenvalue densities are presented in figure 7.3, with subfigures indexed according to the indices in figure 7.1.

The index is also observed to exhibit several periods of high volatility, where the consecutive changes are more or less of same sign. This leads to larger drops or rises of the index. Figure 7.2 presents four such periods, with corresponding empirical eigenvalue densities presented in figure 7.4. Similarly to figure 7.3, subfigures of figure 7.4 are indexed according to the indices in figure 7.2.

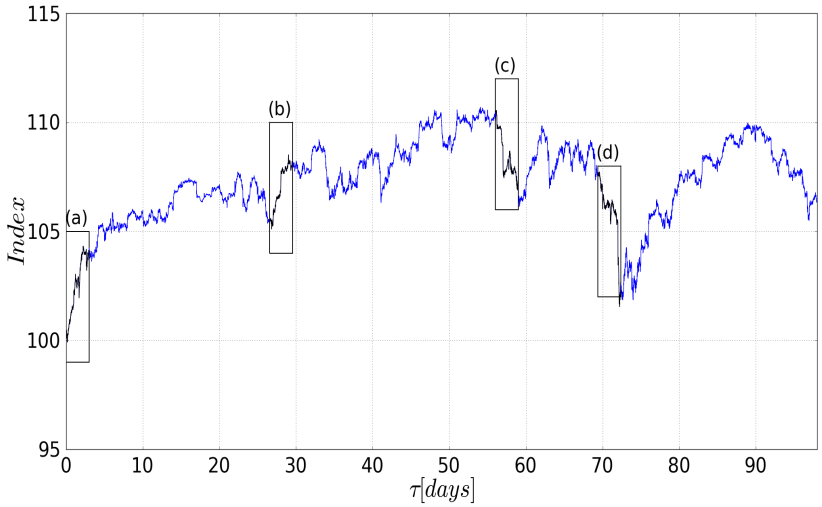
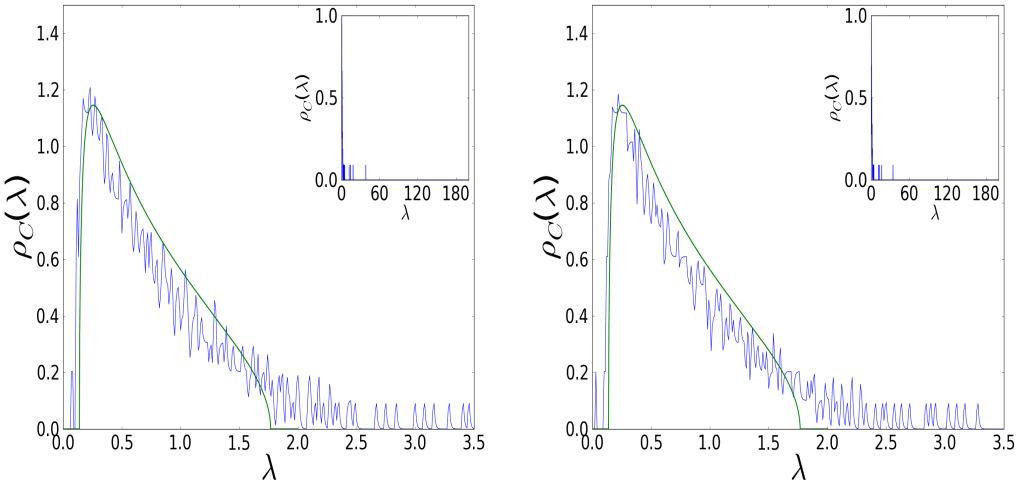


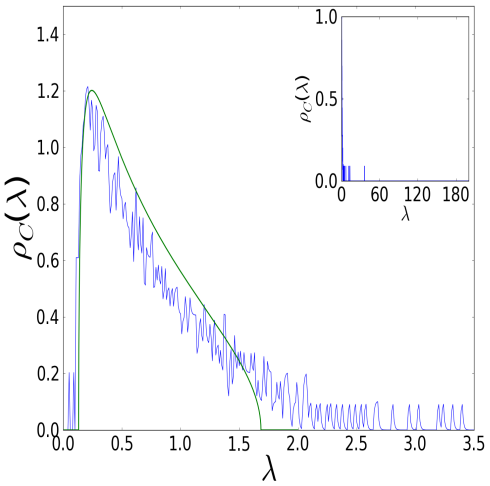

Figure 7.2: Blue line: The index. Black boxes: Time-windows where the index exhibits periods of high volatility, where fluctuations are more or less of the same sign. Empirical densities of eigenvalues based on these time-windows are presented in figure 7.4. The indices above the time-windows are included as a reference, and correspond to subfigure indices of figure 7.4.



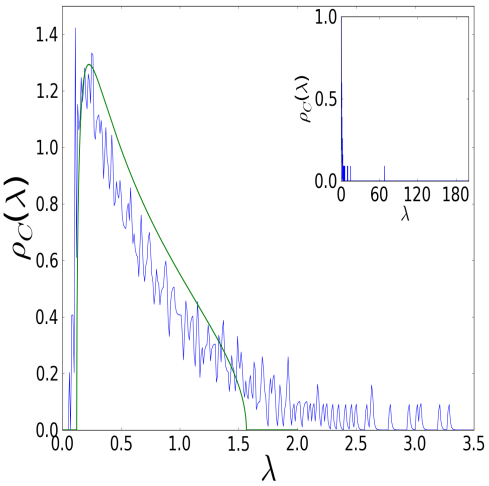

(a) $\sigma^2 = 0.72$




(b) $\sigma^2 = 0.72$

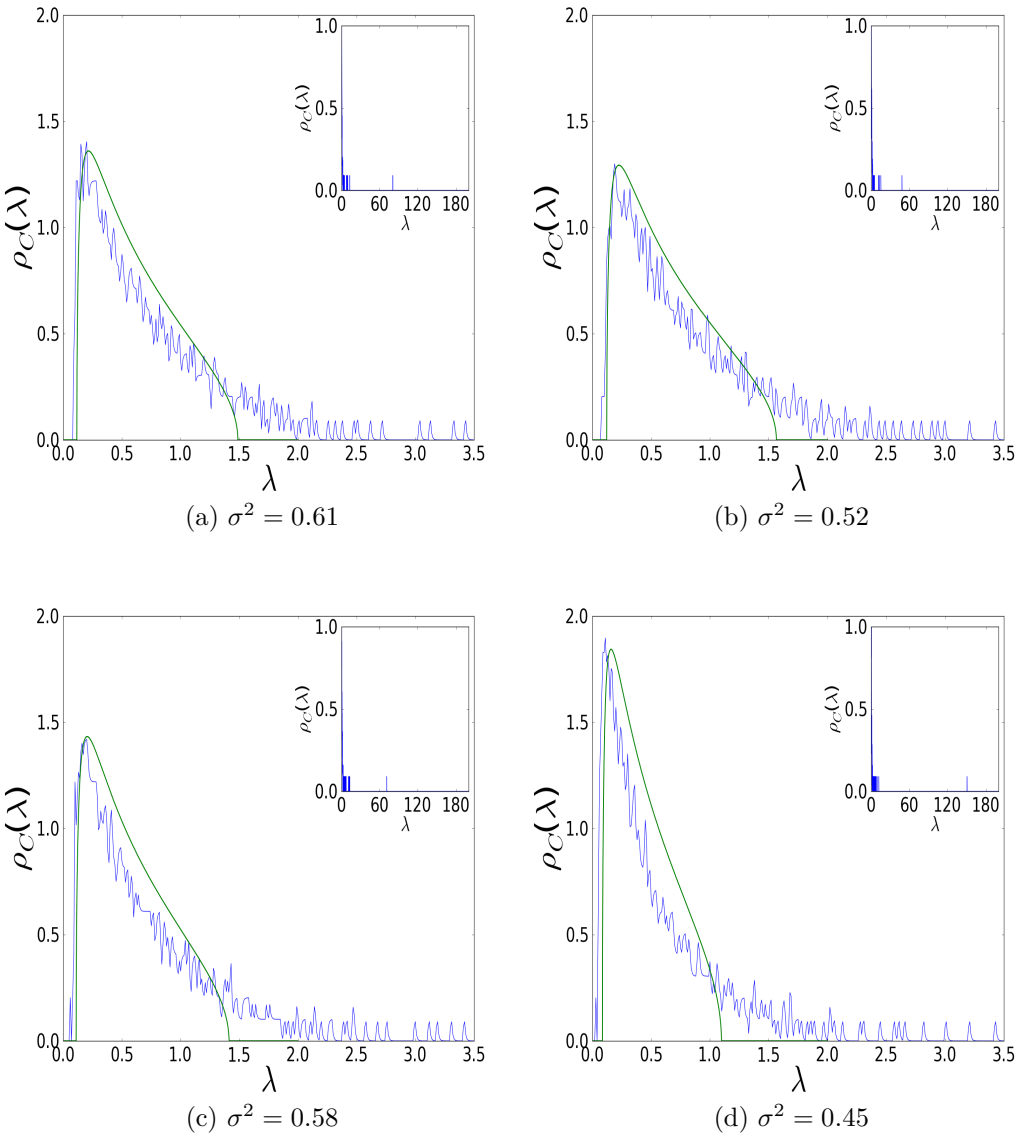



(c) $\sigma^2 = 0.69$




(d) $\sigma^2 = 0.64$

Figure 7.3: Blue line: Smoothed density of eigenvalues of the correlation matrix $\mathbf{C}$, extracted from $N = 492$ of the largest European assets for time-windows of length $T = 1548$ minutes. Time-windows the specific densities correspond to are illustrated in figure 7.1. Green line: Theoretically predicted density of eigenvalues, based on a purely random matrix of same size with $\sigma^2$ as seen below the subplots. Inset: Same plot, but the largest eigenvalue is also included.




(a) $\sigma^2 = 0.61$ (b) $\sigma^2 = 0.52$

(c) $\sigma^2 = 0.58$ (d) $\sigma^2 = 0.45$

Figure 7.4: Blue line: Smoothed density of eigenvalues of the correlation matrix $\mathbf{C}$, extracted from $N = 492$ of the largest European assets for time-windows of length $T = 1548$ minutes. Time-windows the specific densities correspond to are illustrated in figure 7.2. Green line: Theoretically predicted density of eigenvalues, based on a purely random matrix of same size with $\sigma^2$ as seen below the plots. Inset: Same plot, but the largest eigenvalue is also included. Note that figures (a) and (b) correspond to sharp upturns in the index, while (c) and (d) correspond to sharp index drops.

An inspection of figures 7.3 and 7.4 reveals that all eight eigenvalue densities have a bulk density of small eigenvalues distributed approximately around 1. In addition, all densities show several large eigenvalues exceeding the upper limit predicted from RMT. This is similar to the observations of Lalox et al. [24] based on daily closure prices of the DJIA, and implies that parts of the correlation matrices contain true information. This follows, as the large eigenvalues of a correlation matrix reflect the presence of true correlations. Following, theoretical densities fit the empirical densities excellently. The only difference between the two is an apparent lowering of the empirical bulk distribution, in addition to the presence of large eigenvalues. However, the lowering of the bulk distribution is caused by the fact that densities are normalized. Hence, bulk distributions of empirical densities must be lower than those of the theoretical densities to account for the presence of large eigenvalues outside the interval predicted by RMT.

The goal of this analysis is to compare empirical eigenvalue densities to theoretical densities, based on the *null hypothesis* that correlation matrices are completely random. However, the fact that the largest eigenvalues are significantly larger than the predicted upper limit of the eigenvalue densities contradicts this assumption. As discussed in chapter 6, the largest eigenvalue corresponds to the eigenvector that describes the market. Eigenvectors of the other large eigenvalues describe economic sectors. It is fair to assume that eigenvectors orthogonal to eigenvectors corresponding to the largest eigenvalues can be considered as random noise [24], and hence the random part of the correlation matrix does only account for a part of the system's total volatility. It follows that the volatility accounted for by the random part of the correlation matrix can be considered as an adjustable parameter. As discussed in section 3.6.2, the magnitude of the eigenvalue corresponds to the volatility accounted for by its corresponding eigenvector. It follows that the volatility[3] of the random part can be described as

$$\sigma^2 = 1 - \sigma_\lambda^2, \quad \text{where} \quad \sigma_\lambda^2 = \sum_{\mathrm{i}=1}^{\mathrm{M}} \frac{\lambda_\mathrm{i}}{\mathrm{N}}. \tag{7.1}$$

Note that M is the number of eigenvalues significantly larger than the upper limit from RMT, and N the number of stocks used to calculate the correlation matrix. It follows that the quantity $\sigma_\lambda^2$ is the part of the total volatility accounted for by the M largest eigenvalues. If the correlation matrix is completely random, no eigenvalues will be observed outside the interval predicted from RMT[4]. In that case, the random part will account for all the volatility

[3]Remember that the volatility of the log-returns is equal to 1 with the standardization followed in this work.

[4]Note that there will always be some discrepancies, as all real time-series are of finite

of the system, and $\sigma^2$ should be set to one. As discussed in section 6.1, introducing correlations will lead to the presence of larger eigenvalues carrying information about the correlations causing them. Their corresponding eigenvectors will account for (large) parts of the system's total volatility. It follows that the part of the volatility accounted for by the random part of the correlation matrix in this case must decrease, leading to $\sigma^2 < 1$. This explains why $\sigma^2$ was observed to be less than one in figures 7.3 and 7.4.

The first periods considered are those of figure 7.1. These are calm periods characterized by low volatility and fluctuations of varying sign, leading the index to remain nearly unchanged. The corresponding eigenvalue densities are presented in figure 7.3. It is observed that $\sigma^2 \in [0.69, 0.72]$ for time-windows 7.1a-c, while $\sigma^2 = 0.64$ for time-window 7.1d. This implies that the fraction of total volatility accounted for by the largest eigenvalues is approximately the same for the three first time-windows, but slightly larger for the last time-window. This is also confirmed by considering the largest eigenvalue, observed to equal $\lambda_1 \approx 40$ for time-windows 7.3a-c and $\lambda_1 \approx 65$ for time-window 7.1d. This means that the volatility accounted for by the market has increased from roughly 8% for the three first time-windows, to 13% for the last time-window. It follows that stronger collective trends must be present during period 7.1d. To point on any specific event that would cause a stronger collectivity during this period is not straightforward, as the period appears to be similar to the other three periods. However, it is noticed that time-window 7.3d is placed at the end of a long period with a rising index. This leads to the speculation that the transition from a period with a rising index to a calm period where the index remains nearly unchanged can cause increased collectivity in the market. This is inspired by the fact that collective trends get stronger when the index drops, as will be discussed in the following part of this section. In other words, the end of a rising period results in an uncertain market[5]. Uncertainty is often followed by dropping prices or prices that remain nearly unchanged during longer periods, as many investors relocate their money to reduce their risk exposure.

Eigenvalue densities from the four periods of figure 7.2 are presented in figure 7.4. These periods are characterized by high volatility and fluctuations of more or less same sign, leading the index to either rise or drop. The first observation is that the densities seem to be narrower and more peaked than

length. However, this will only cause blurred edges with small deviations, not significant deviations such as those observed in figures 7.3 and 7.4.

[5]This is a 'chicken and egg' problem, as it can just as well be the increased uncertainty in the market that caused the rising period to end.

those originating from calm periods[6]. As discussed in section 3.7.1, this implies that the volatility accounted for by the random part of the correlation matrix has decreased. The same is concluded by considering the parameter $\sigma^2$, taking values $\sigma^2 \in [0.45, 0.61]$ which are all less than those used for the calm periods.

Time-windows 7.2c-d are periods where the index exhibits a sharp drop. Their corresponding largest eigenvalue is observed to equal roughly $\lambda_1 \approx 70$ and $\lambda_1 \approx 150$ respectively, implying that the market accounts for roughly 14% and 30% of the total volatility. Hence, as monitored by $\lambda_1$, the collective trends are stronger during index drops than during calm periods. The fact that the collective trends are stronger during the period with the largest drop (time-window 7.2d) indicates that the size of the index drop is important for the strength of the correlations. It follows that it would be favorable to introduce *individual* fear factors with the ability to synchronize parts of the market such as selected economic sectors to the fear factor model [42], as also was suggested by Simonsen et al. [41]. The next periods to be considered are those where the index exhibits sharp rises, time-window 7.2a-b. Their corresponding largest eigenvalue is observed to equal roughly $\lambda_1 \approx 80$ and $\lambda_1 \approx 50$ respectively, and it follows that the market accounts for about 16% and 10% of the total volatility. The first case seems to correspond to significantly stronger correlations than during the recently discussed calm periods. However, the investigation of an index similar to that constructed in this work reveals that a drop of roughly 4% ended a day prior to the start of the dataset[7]. As was observed from time-windows 7.2c-d, periods where the index drops strongly are characterized by strong collective trends. This leads to the speculation that the increased collectivity observed in time-window 7.2a is a remain from the preceding drop in the index. The latter case corresponds to collective trends just slightly stronger than those of the calm periods 7.1a-c, and also this period takes place after a roughly 2% drop. Again, it is speculated that the slightly increased collectivity is a remain from the preceding drop. If this is the case, it seems as if periods where the index rises not necessarily are connected to stronger correlations between stocks.

The fact that $\lambda_1$ is significantly larger than the upper bound from RMT for all the above presented densities indicates that a certain minimum of collectivity is present in the market at all times. It is interesting to discuss where such correlations originate from, as one would not expect any correlations between stocks to be present in ideal markets where stock-price changes

[6]Note that the eye can be slightly misled, as the y-axis covers a narrower interval in figure 7.3 than what it does in figure 7.4.

[7]This can be seen by considering the S&P Europe 350 [55].

are completely independent. Real markets, however, are composed of sectors consisting of related companies. Therefore, the continuous flow of news into the market causes stock-stock correlations to appear from time to time, especially between companies belonging to same sector. News can also affect the market as a whole and cause correlations to appear between most stocks, an effect often amplified by human psychology. These latter correlations are implicitly accounted for by the fear factor model of Donangelo et al. [42]. In principle, the model introduces collective trends only during periods where the fear factor causes a synchronized index drop. However, as the chance of a stock moving up is slightly larger than the chance of moving down during periods without synchronized movements, this causes a weak collective trend to be present at all times.

So far, it is observed that a certain minimum level of collective trends is present at all times where the index not exhibits any sharp rises or drops. If the index does exhibit a sharp drop, this causes the strength of collective trends in the market to increase by an amount that seems to be connected to the size of the drop. Sudden rises of the index on the other hand are speculated to not be connected to an increased collectivity. However, it is not possible to draw any firm conclusion based on only eight empirical densities of eigenvalues. As the interesting quantity is $\lambda_1$, it is therefore more useful to consider its temporal dependence. A comparison of $\lambda_1$ to the index can reveal any potential connections between the two, and is performed in section 7.2.

As a last point, it is discussed why the values of $\sigma^2$ used in the fits of the above empirical densities are found to be smaller than observations from other authors. The parameter was found to lie in the interval $\sigma^2 \in [0.45, 0.72]$ for all eight densities presented in this chapter, while Lalox et al. [24] observed that $\sigma^2 = 0.74$ gave the best fit to their results[8]. However, their results were based on daily closure prices of the S&P500, covering the period 1991 to 1996. The data used in this work are of higher frequency, and therefore show different features than daily closure prices do. It follows that the difference must arise from the nature of the dataset. This is reasonable, as the autocorrelation function of log-returns is fast decaying and characterized by a correlation time much shorter than a trading day. It follows that (positive) correlations lasting only several minutes exist, but are not reflected in daily closure prices. Based on 1 minute data from the S&P500, the length of these short-range correlations was found to be about 20 minutes [5, p. 55]. It follows that these short-range correlations decrease the randomness of

[8]The results of Lalox et al. [24] are presented in figure 3.8.

correlation matrices based on high-frequency data compared to daily closure prices, explaining why $\sigma^2$ was observed to be smaller than the observations of Lalox et al. [24].

It is also interesting to speculate why $\lambda_1$ suddenly increases during large drops in the index. The index will always fluctuate, and in general follow a trend that has been observed to be close to exponential (see figure 1.1). When the index drops, several events can cause the drop to be amplified. One of these originates from the stop-loss[9] limit followed by many investors: When a stock falls enough to reach the stop-loss limit, investors sell and cause further draw-downs of stock-prices. Such draw-downs can lead to fear among other investors, resulting in a cascade and a synchronized fall also for stocks not necessarily correlated with the stock where the original fall first started. This would cause gain-loss asymmetry to be observed in the index, but also in individual stocks of the index. As the asymmetry is hard to observe in individual stocks [50], this suggests that stop-loss limits are not the origin of the observed asymmetry. There is also an opposite event termed short-squeezes[10]. A short-squeeze is induced when investors buy assets due to a strong rise in the stock price. As borrowed stocks must be delivered back to their owner, investors fear for further increases of the stock-price and hence buy stocks on increasing prices. Other investors can follow the trend and buy the same or other stocks, hoping that prices will continue to rise. The result is again a cascade, leading to a synchronized rise of the index. This will cause the opposite asymmetry, as the sudden rise of the index will decrease (increase) the investment horizons for positive (negative) levels of return. This follows, as the short-squeeze works as an attractor decreasing the probability of long waiting times for positive returns. This will also cause a gain-loss asymmetry in the index (of opposite sign, i.e. with optimal investment horizon shorter for gains than losses), but also in individual stocks. It follows that neither short-squeezes are the origin of the asymmetry. In addition, if the two effects are of same size, they will cancel each other and not cause any asymmetry.

[9] A limit many investors often set as a lower limit, where investors sell their assets if this limit is reached to minimize their loss.

[10] An investor that borrows stocks and sells them has a short-position in the stock. Shorting is a bet on falling stock-prices, but also when prices increase the stocks must be bought back at a time agreed when the stocks were first borrowed. If many investors have shorted stocks in a company that has increased in value, the buy-back of stocks that are to be delivered back to their owner can cause the stock-price to increase rapidly.

## 7.2 The largest eigenvalue $\lambda_1$ and the index

Far more than the eight empirical eigenvalue densities presented in the last section have been calculated, and no firm conclusion could be drawn from only eight empirical densities. As the interesting quantity from each time-window is the largest eigenvalue $\lambda_1$, it is therefore useful to consider $\lambda_1$ itself as a function of time. This follows, as $\lambda_1$ is used to monitor the strength of collective trends. Comparing this to the index can reveal any possible connections between the two. Figure 7.5 presents the movements of $\lambda_1$ based on a step-size $X = 300$ minutes for the sliding time-window. However, note that a plot of higher resolution will be presented and discussed at the end of this section. The reason why the curve describing $\lambda_1$ starts after the index curve is that only one set of eigenvalues is calculated from each time-window. In this work, it has been chosen to place $\lambda_1$ at the end of the time-window to ensure that it is entirely determined from earlier observations. In other words, $\lambda_1(t_k)$ is based on time-window $\mathbf{M}_k$ of length $T = 1548$ minutes (3 trading days), covering a period $t \in [t_k - T, t_k]$, where $t_k = T + n_k X$. Note that $t_k \leq t_{\max}$, where $t_{\max}$ is the length of the stock-price series and $n_k$ an integer starting from zero.

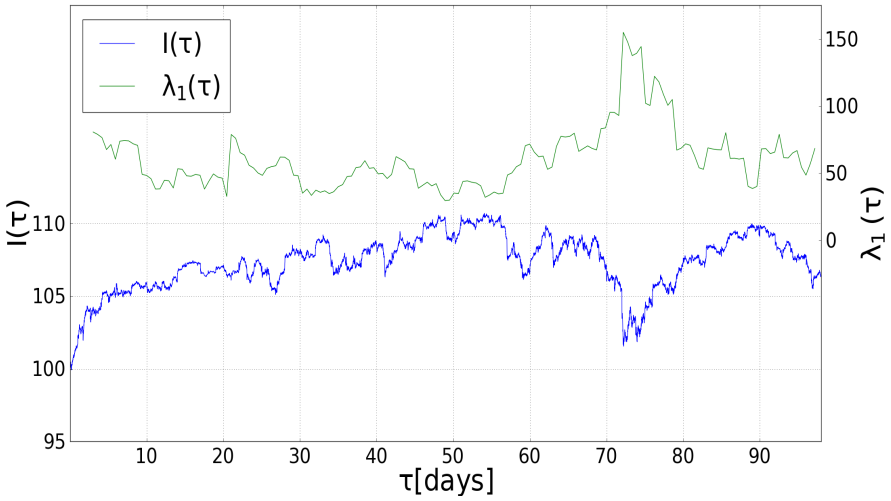



Figure 7.5: Blue line: The index. Green line: The largest eigenvalue $\lambda_1$. The step-size between each time-window of length $T = 1548$ minutes is set to $X = 300$ minutes.

An inspection of figure 7.5 reveals that $\lambda_1$ always is significantly larger than the predictions from RMT, confirming that collective trends of a certain size are present at all times. A closer look at the two curves reveals an apparent anti-correlation. This observation is further supported by the correlation coefficient $\rho_{\mathrm{I},\lambda_1} = -0.61$, based on the total length of the series. The increasing strength of collective trends during index drops is speculated to arise from the psychology of investors that, ignited by some internal or external event, causes investors to sell synchronized. This is exactly the key idea behind the fear factor model of Donangelo et al. [42], where the fear factor triggers all stocks to move downward at the same time. However, whether the decreasing collectivity is due to the rising index, or just due to the fact that the drop has ended, is not possible to determine with certainty. This follows, as the current dataset not contains any periods with a sharply rising index that not follow a recent drop. However, no periods with a sharply rising index are related to an increasing collectivity. This leads to the preliminary conclusion that periods where the index sharply rises *not* are related to an increasing collectivity. Note that this does not exclude the possibility of periods where a rising index is positively correlated to the collectivity as monitored by $\lambda_1$, only that these periods are not observed when the index sharply rises in the vicinity of sharp drops. However, such periods are difficult to observe from figure 7.5. The running correlation between $\lambda_1$ and the index gives relevant information on correlations between the index and $\lambda_1$ on a chosen scale, and thus makes it easier to observe such periods. The reason for why such periods are interesting, is that they indicate the presence of a potential *optimism factor*, forcing stocks to rise synchronously. The running correlation also provides further information on the connection between the index and the collective trends in general.

The running correlation $\rho_{\mathrm{I},\lambda_1}(\tau_k)$ is calculated from time-windows of length $T = 2500$ minutes (roughly one trading week), where $\rho_{\mathrm{I},\lambda_1}(\tau_k)$ is based on a time-window covering the period $\tau \in [\tau_k - T/2, \tau_k + T/2]$. Note that $t_k = T/2 + n_k X$ where $t_k \leq t_{\max}$, $t_{\max}$ is the length of the series containing $\lambda_1$ and $n_k$ an integer starting from zero. To calculate $\lambda_1$, a step-size of $X = 10$ minutes was used. The result is presented in figure 7.6

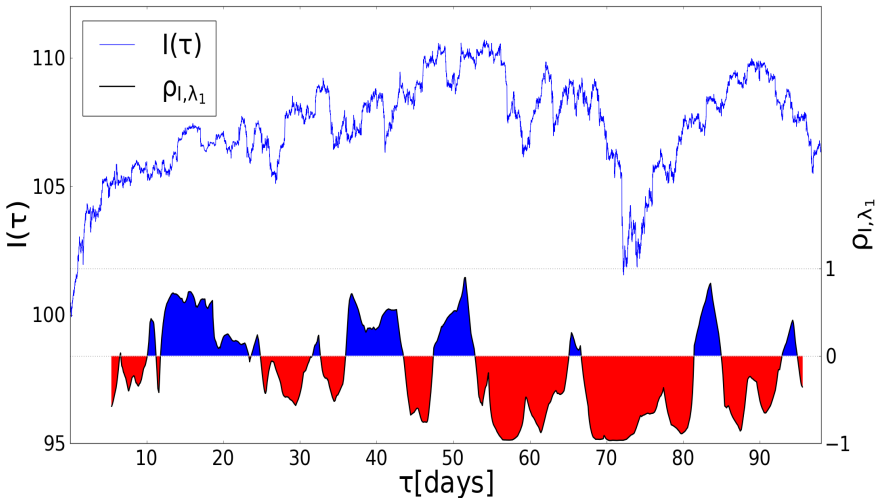



Figure 7.6: Blue line: The index. Black line: The running correlation between the largest eigenvalue $\lambda_1$ and the index, based on a time-window of length $T = 2500$ minutes. The area below the curve is shaded red (blue) for negative (positive) correlations to guide the eye. The step-size between each time-window used for the calculation of $\lambda_1$ is $X = 10$ minutes.

Figure 7.6 reveals that the sign of the correlation between the index and $\lambda_1$ is not constant, but changing several times during the nearly 6 month period covered by the data. However, the general trend seems to be that periods with sharp drops or rises of the index are anti-correlated to $\lambda_1$. This indicates that strong index drops are connected to a strongly increasing collectivity in the market, as monitored by $\lambda_1$. As recently discussed and pointed out in section 7.1, it is not possible to determine whether periods where the index sharply rises are connected to decreasing collective trends or not. Determining this would require a dataset containing periods where the index sharply rises, but not in the vicinity of a preceding drop. However, figure 7.6 reveals that periods where the index rises sharply are anti-correlated to $\lambda_1$. Based on this, it is concluded that the periods where the index rises sharply in general are *not* related to an increasing collectivity.

During periods where the index changes are less dramatic, it is observed both positive and negative correlations between the index and $\lambda_1$. This makes it possible to observe periods where the index rises and is *positively* correlated to $\lambda_1$, indicating the existence of the *optimism factor* that forces stocks to rise synchronously. Figure 7.6 reveals that the most pronounced such period is that lasting roughly from day 12 to day 22. During this period, the in-

dex first rises and then enters an uncertain and volatile period with no clear trend. It is observed that the strength of the correlation is strong during the rising part of the period, but that it decreases towards insignificant levels for the latter volatile part. The fact that the rising index is positively correlated to $\lambda_1$ indicates that the collectivity, as monitored by $\lambda_1$, increases parallely with the index. This suggests the existence of the *optimism factor*. There are also shorter periods showing the same features, such as the periods lasting roughly from day 37 to 43 and day 81 to 83. The periods where the index is positively correlated to $\lambda_1$ also contain corrections, where the falling index is related to a decreasing collectivity. This is speculated to be caused by investors considering the corrections as opportunities to buy 'cheap' stocks, therefore leaving already existing trends.

The results so far are that periods where the index exhibits a sharp drop and then rises coincide with an anti-correlation between the index and the strength of collective trends in the market. It follows that collective trends get stronger during sharp index drops, and that their strength decreases during the following period where the index rises. It can therefore be concluded that the drop relates to an increasing collectivity. Whether the decreasing collectivity is caused by the following period where the index rises, or that it also would happen if the drop was followed by a calm period, is not possible to determine. This follows, as no periods with a significantly rising index occurring out of the vicinity of a preceding drop are present in the current dataset. However, what can be concluded is that periods where the index rises strongly in general *not* cause the collectivity in the market to increase. As periods with a falling and rising index in general not necessarily only occur after each other, but also can be separated by calm periods allowing the market to stabilize, it is strongly suggested that collective trends are stronger during falling than rising markets. This observation provides support to the results of Balogh et al. [40]. However, Balogh et al. [40] used daily closure prices from DJIA, which by nature are very different from the high-frequency data used in this work. To provide further support to their results, it will be useful to consider similar data. Therefore, an identical analysis to that performed on the high-frequency dataset is presented in chapter 8, using the same dataset as that used by Balogh et al. [40]. Among the results is also the observation of periods where the index is positively correlated to the strength of collective trends as monitored by $\lambda_1$. These periods seem to occur when the index does not change as dramatically, and are in addition relatively rare, which is why they did not affect the conclusion that the collective trends appear to be stronger during falling than rising markets. However, the observation of such periods where the index is rising suggests the existence of the *optimism factor* that forces stocks to rise synchronously.

Some of these periods also contain smaller drops or corrections, speculated to reflect investors buying what they believe are 'cheap stocks' during smaller corrections of the index, thus deviating from existing trends to pick the best stocks. In other words, the small drops do not ignite enough fear to cause a sharp index drop, leaving the period characterized by optimism. The analysis has also indicated that there are aspects of the fear factor model of Donangelo et al. [42] (recently modified by Siven et al. [43] to account for longer periods of synchronization) that with an advantage can be introduced. First of all, the *optimism factor* introducing an index draw-up characterized by highly correlated stocks should be added to the model. It is also suggested that the model should have *individual* fear factors for each economic sector in addition to the global fear factor affecting all stocks, as also has been suggested by Simonsen et al. [41].

The last part of this section is an analysis of the movements of $\lambda_1$ with a higher resolution, as discussed in the start of the section. This is done using a smaller step-size for the sliding time-window, and reveals some interesting aspects of $\lambda_1$. Figure 7.7 presents the movements of $\lambda_1$ compared to those of the index, using a step-size $X = 10$ minutes.

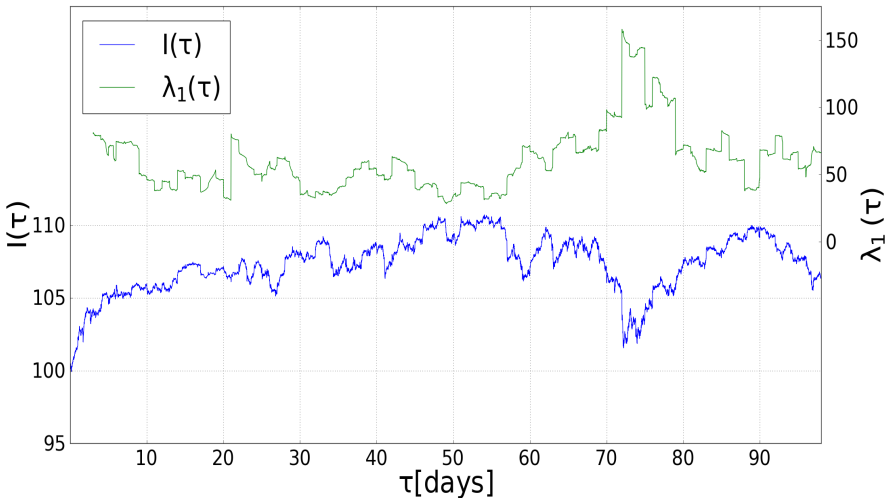
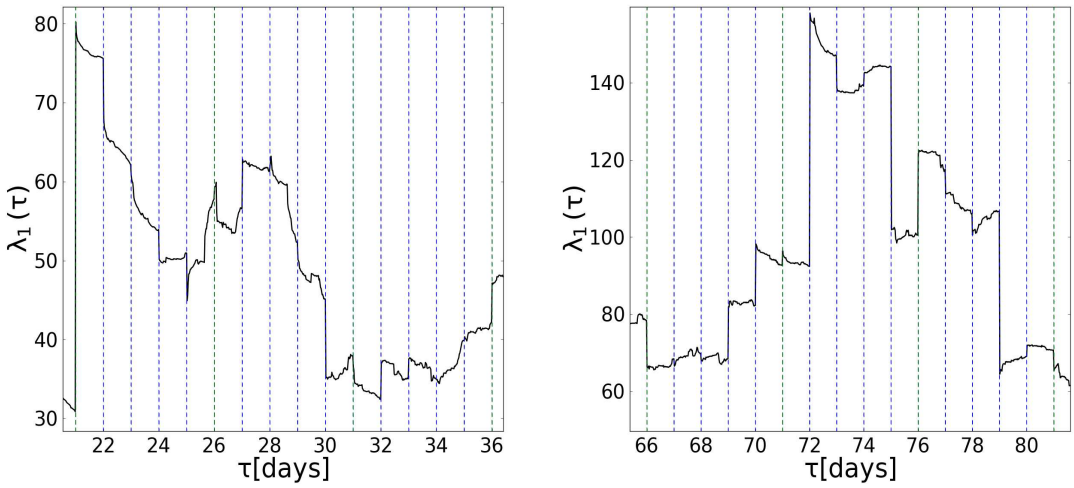



Figure 7.7: Blue line: The index. Green line: The largest eigenvalue $\lambda_1$. The step-size between each time-window of length $T = 1548$ minutes is set to $X = 10$ minutes. Notice that the curve describing $\lambda_1$ exhibits periodical jumps of what appear to be a random sign.




(a) (b)

Figure 7.8: Black line: The largest eigenvalue $\lambda_1$ as a function of time. Vertical lines: Blue dotted lines separate each trading week, green dotted lines separate each trading day.

The first observation from figure 7.7 is that $\lambda_1$ seems to exhibit strong jumps of varying sign. Figure 7.8 presents two zooms into figure 7.7, with vertical lines separating each trading day (and week) added to guide the eye. As revealed when zooming in on the curve describing the temporal dependence of $\lambda_1$, the jumps clearly reflect the overnight effect.

Intuitively, one would assume that the largest jumps should occur after a weekend. This follows, as the time between two trading days is longer during a weekend than during a night. However, it is clear that this is not the case, as the size of the jumps seems to be rather random. This suggests that the overnight effect is of similar strength also over weekends. In principle, this is reasonable. Even though a weekend lasts longer than a night, it is fair to assume that the rate of news released during a weekend is lower than the rate of news released during a single night. This follows, as most companies do not release news such as financial reports etc. during weekends. However, it is very common to release such news outside the opening hours of the stock exchange during weekdays.

# Chapter 8

# A random matrix approach to collective trends of the DJIA

In their first paper reporting a gain-loss asymmetry, Jensen et al. [38] considered daily closure prices from the Dow Jones Industrial Average (DJIA). Later authors trying to explain the origin of the asymmetry have therefore used similar data. As the asymmetry was speculated to originate from collective trends in the market [37, 38], Balogh et al. [40] used daily closure prices from the DJIA to investigate whether the strength of collective trends of share prices is different during stock index rising and falling periods. After performing several averaging procedures that in principle *can* affect the results, they observed a clear trend: Falling markets do indeed show stronger collective trends. However, there is a well known phrase: "You can't compare apples with pears". In other words, the results from the high-frequency data should not be directly compared to results based on daily closure prices. This follows, as the two types of data are very different by nature and catches different aspects of the market.

Even though the results from the analysis of high-frequency data cannot be directly compared to the results of Balogh et al. [40], the work was found to support their observations. This chapter presents an analysis of a similar dataset to that used by Balogh et al. [40]. Results can therefore be directly compared to their results. In addition, any potential differences between high-frequency data and daily closure prices are interesting to study. The analysis is performed following the same procedure as for the high-frequency data. Therefore, the procedure is only shortly brushed up as this is explained in detail in section 5.1 for the high-frequency dataset. Small differences arising from the size of the set of daily closure prices, however, are discussed in section 8.2.

## 8.1 The DJIA dataset

The dataset used in this chapter is similar to that used by Balogh et al. [40], and was obtained from Yahoo! Finance [21]. It consists of the adjusted[1] daily closure prices from 29 of the 30 companies that were members of the DJIA late February 2008[2]. The set covers the 18 year period May 15th 1991 to September 1th 2008, and contains 4160 daily closure prices. Company names, ticker codes and corresponding GICS-sectors of the 29 stocks are presented in appendix A.2, table A.2.

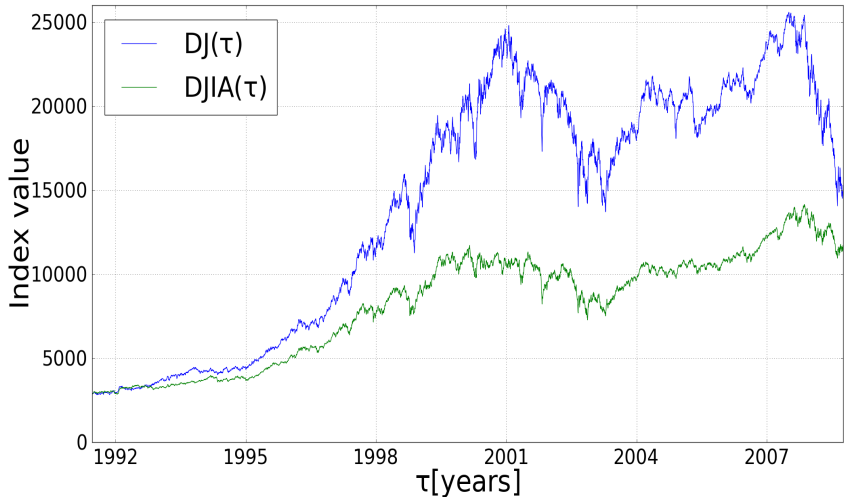



Figure 8.1: Blue line: The DJ index (the direct average of 29 of the stocks constituting the DJIA index late February 2008, according to equation (4.1)). Green line: The DJIA index. The period covered by the indexes is 15th May, 1991, to 1th September, 2008.

As the index must reflect only the 29 stocks in the dataset, one cannot use the original DJIA. Following, as the composition of the DJIA not is constant, some of the 29 stocks have not been included in the index during all 18 years covered by the dataset. An example is The Home Depot Inc., not entering the DJIA before 1999. These problems have been avoided by the construction of a new index similar to the DJIA. The index is price-weighted as discussed in section 4.2, and calculated according to equation 4.1. For simplicity, the divisor is set equal to the number of stocks for all times, $d(t) = 29$. The result

[1]The price-series are adjusted for dividends and splits.

[2]Balogh et al. [40] used all 30 components, but as GM (General Motors Co.) went through a controlled bankruptcy in 2009, its historical prices has been removed.

is a new index, denoted as the DJ index in the following discussion. Figure 8.1 presents the DJ index and the DJIA index. Note that their correlation coefficient is $\rho_{\mathrm{DJIA,DJ}} = 0.97$, indicating that the two indexes behave more or less identically, which is also revealed by inspecting the figure.

To confirm that the index exhibits a gain-loss asymmetry, inverse statistics distributions were calculated for the DJ index for a return-level $|\rho| = 5\sigma$, where $\sigma \approx 1.23\%$ corresponds to the daily standard deviation of the DJ log-returns. Expressed in percents, this equals a return-level of roughly 6%. Note that the main point is to *underline* the presence of an asymmetry in the index. The specific magnitude of such an asymmetry is not of importance. It follows that it is not necessary to detrend the index, as this only increases the gain-loss asymmetry. This follows, as the detrending shifts the optimal investment horizon for gains towards longer waiting times and oppositely the optimal investment horizon for losses towards shorter waiting times [40]. In other words, when the asymmetry is present for the not detrended index, it is surely present also for the detrended index.

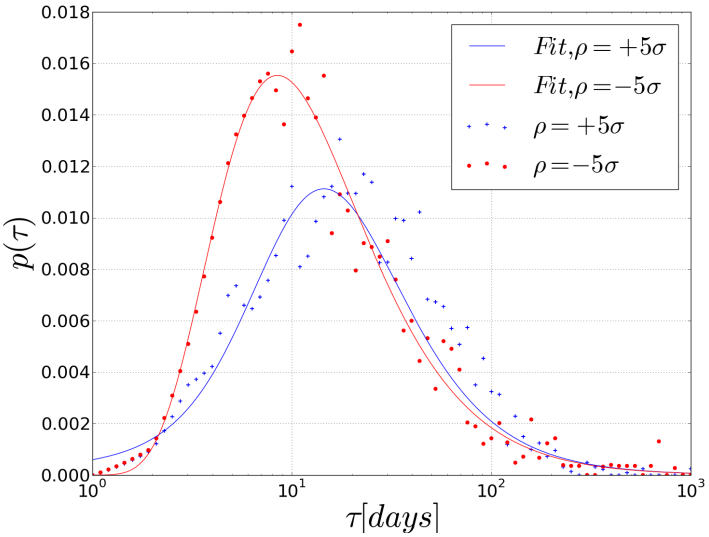



Figure 8.2: Inverse statistics distributions for the DJ index (red and blue dots). The return-level used is $|\rho| = 5\sigma$, where $\sigma \approx 1.23\%$ is the daily standard deviation of the DJ log-returns. Solid lines represent the least squares fit of equation (3.81) to the empirical data, with $\tau^*_{\pm|\rho|}$ and parameters $\nu, \alpha, \beta$ as presented in appendix B.1, table B.4. As for the high-frequency data, the tail exponent $\alpha + 1$ is not distinguishable from the "random walk value" 3/2.

The resulting inverse statistics distributions are presented in figure 8.2. An inspection leads to the conclusion that the DJ index exhibits a clear gain-loss asymmetry of magnitude about 6 days when $|\rho| = 5\sigma$. An analysis of selected individual stocks using the exact same return-level was found to show weak or no pronounced asymmetry, in accordance with the results of Balogh et al. [40]. As the analysis of individual stocks is presented in [40], it is not necessary to present the specific results here.

## 8.2 Method

The daily closure prices of the 29 stocks from the DJIA are analyzed more or less exactly as the high-frequency prices. The procedure is described in detail in section 5.1 for the high-frequency data, and consists in general of a time-window that is slid through the time-series. Each time-window leads to a correlation matrix and a corresponding density of eigenvalues.

The difference is that the DJIA dataset is much smaller than the high-frequency dataset, consisting of 4160 observations from each of the 29 stocks. To account for this, the length of the time-window has been decreased to $T = 290$ days. This gives a ratio $Q = 3$ between the number of observations $T$ and the number of stocks $N$, and was chosen to be consistent with the value $Q = 3.22$ used by Lalox et al. [24].

The result is a set of eigenvalue densities that is compared to theoretical densities based on a *null hypothesis* purely random matrix. An analysis of eigenvectors from the current dataset revealed that also for this dataset, the eigenvector corresponding to the largest eigenvalue seems to describe the market. This follows, as it has only positive components of roughly the same size. If the largest eigenvalue deviates from the theoretical density of eigenvalues, it indicates that it carry information about true correlations in the market. It follows that the largest eigenvalue can be used as an index describing the strength of collectivity in the market, similar to what was done for the high-frequency data.

## 8.3 Results and discussion

Similar to what was done for the high-frequency data in section 7.1, some selected empirical eigenvalue densities are presented and discussed in this section. As it is the largest eigenvalue that is of importance, not the densities themselves, only two densities are considered in detail. The particular time-windows these densities are based on are illustrated in figure 8.3.

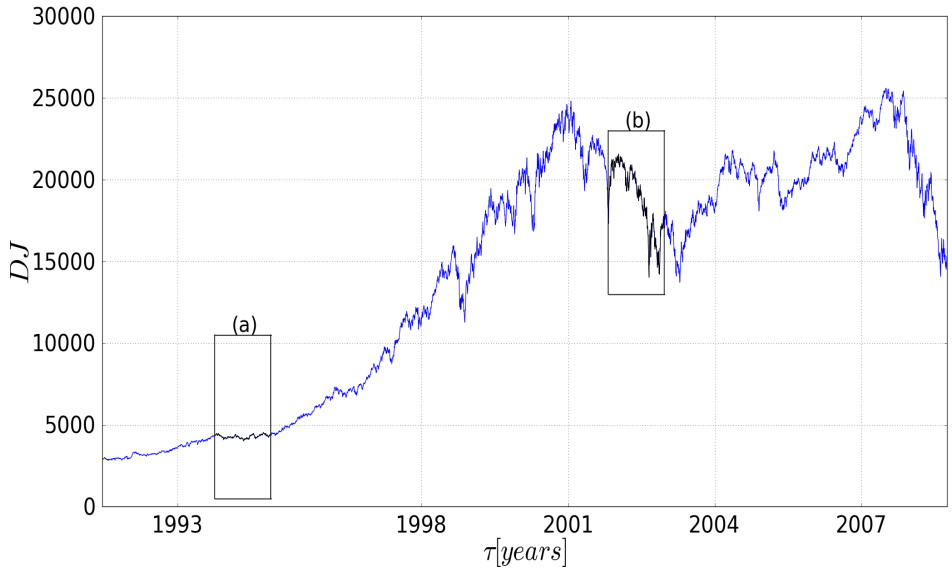
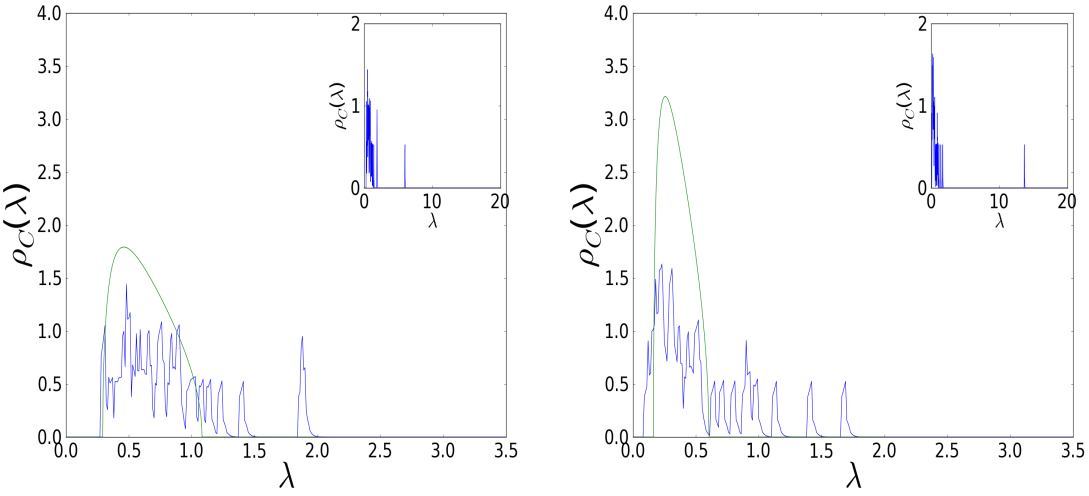



Figure 8.3: Blue line: The DJ index. Black boxes: Time-windows used to calculate the empirical densities of eigenvalues from the DJ index. Indices above the time-windows correspond to subfigure indices of figure 8.4.

The empirical densities of eigenvalues are presented in figure 8.4. Remember that the size of the correlation matrices the densities are based on is reduced from a $492 \times 492$ matrix for the high-frequency dataset to $29 \times 29$ for this dataset. This will obviously lead to larger discrepancies between theory and empirical data, as random matrix theory (RMT) in principle is based on infinitely large matrices (discussed in section 3.7.1).




(a) $\sigma^2 = 0.62$ (b) $\sigma^2 = 0.35$

Figure 8.4: Empirical densities of eigenvalues based on the DJ index, where (a) and (b) correspond to time-windows as illustrated in figure 8.3. Blue lines: Empirical densities of eigenvalues. Green lines: Theoretically predicted density of eigenvalues, based on a purely random matrix of same size with $\sigma^2$ as seen below the subplots. Inset: Same plot, but the largest eigenvalue is also included.

The first observation from figure 8.4 is that both densities fit reasonably well to the predictions from RMT. This is impressive, as they are based on a dataset containing only 29 stocks. It is also interesting to note that the densities show exact the same features as those of the high-frequency data. These features are a bulk distribution of eigenvalues somewhat lowered compared to the theoretical densities, and several eigenvalues significantly larger than the upper limit predicted from RMT. As discussed in section 7.1, the lowering of the bulk is caused by the presence of large eigenvalues outside the theoretical upper limit. These are not accounted for by the theoretical densities, and it follows that empirical densities must be somewhat lowered to ensure their normalization. Note that empirical densities from all time-windows are found to exhibit similar features, having their largest eigenvalue in the interval $\lambda_1 \in [4, 17]$.

The density presented in figure 8.4a arises from a time-window placed in a calm period of the market, as the index remains nearly unchanged during the period of 290 days covered by the time-window. A good fit to theory was obtained using $\sigma^2 = 0.62$ for the volatility accounted for by the random part

of the correlation matrix. The largest eigenvalue is observed to be $\lambda_1 \approx 5$, indicating that the market accounts for roughly 17% of the total volatility. Time-window 8.3b is placed in a volatile period where consecutive fluctuations are more or less negative. As the fluctuations also are larger during volatile periods, this causes the index to exhibit a sharp drop. The empirical density of eigenvalues fits well to theory, using $\sigma^2 = 0.35$ for the volatility of the random part. This is significantly less than the value used for period 8.3a, indicating that the largest eigenvalues account for more of the total volatility. This is also confirmed by the largest eigenvalue $\lambda_1 \approx 14$, almost triple that of the calm period. The largest eigenvalue is so large that its corresponding eigenvector (the market) accounts for roughly half of the total volatility during the period. The fact that the largest eigenvalue dominates the density suggests the presence of stronger collective trends in the market during the drop than during the calm period.

It was observed that values for $\sigma^2$ obtained from the high-frequency dataset were significantly smaller than values observed by other authors for daily closure prices. The reason for the observed discrepancy was concluded to arise from the different nature of daily closure prices and high-frequency prices. However, it is also observed that values for $\sigma^2$ based on the current dataset seem to be significantly smaller than the results of Lalox et al. [24], observing that $\sigma^2 = 0.74$ provided the best fit to their results. As Lalox et al. [24] based their work on the daily closure prices for the 500 stocks composing the S&P500 index, it is believed that the discrepancy arises from the very low number of stocks the current dataset is composed of. Note that even though individual sectors probably not are very well represented by the dataset, it is believed the market itself is represented more than good enough for the purpose of this work.

As $\lambda_1$ clearly deviates from the theoretical distribution from RMT, it is clear that it reflects information about true correlations in the market. It is therefore interesting to consider its temporal dependence, as also was done for the high-frequency data in section 7.2. Comparing this to the index can reveal whether there are any connections between the two quantities. This is presented in figure 8.5, where $\lambda_1$ has been calculated using the step-size $X = 1$ day to ensure a good resolution of $\lambda_1$'s movements. Note that $\lambda_1(t_k)$ is based on time-window $\mathbf{M}_k$ of length $T = 290$ days, covering a period $t \in [t_k - T, t_k]$, where $t_k = T + n_k X$. Note that $t_k \leq t_{\max}$, where $t_{\max}$ is the length of the stock-price series and $n_k$ an integer starting from zero.

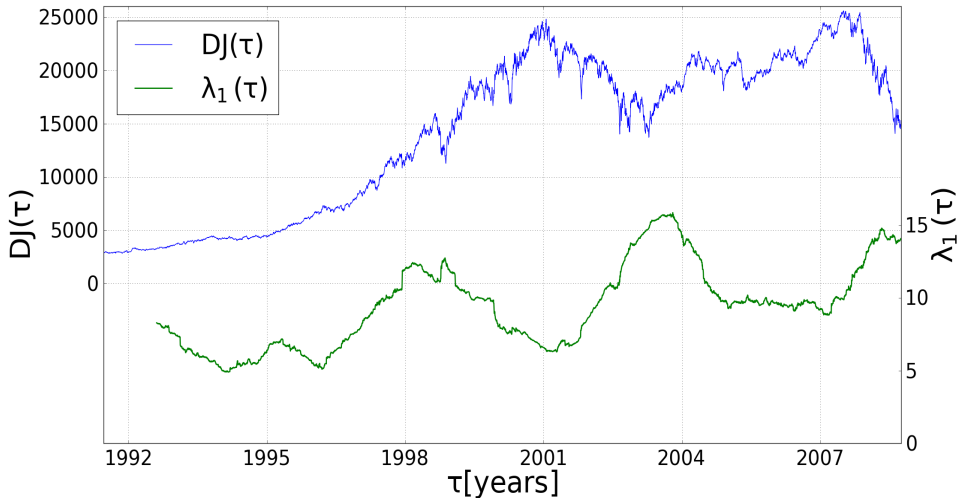



Figure 8.5: Blue line: The DJ index. Green line: The largest eigenvalue $\lambda_1$. The step-size between each time-window is set to $X = 1$ day, and the length of the time-window used to calculate $\lambda_1$ is $T = 290$ days.

Figure 8.5 reveals that the jumps observed for the high-frequency data (see figure 7.7) are not present in the daily data. If weekends causes a stronger overnight effect, one should observe periodical jumps every fifth day. However, this is not the case, supporting the speculation in section 7.2 that the effect seems to be approximately similar for weekends and single nights. It is also observed that $\lambda_1$ is significantly larger than the upper limit predicted from RMT at all times, indicating that a certain minimum of collective trends is present at all times. It is also revealed that $\lambda_1$ follows a rising trend with ascending bottoms. This indicates that the volatility and the strength of collectivity in the market, as monitored by $\lambda_1$, has increased during the nearly two decades covered by the dataset. However, it must be emphasized that it is possible that the poor statistics of the dataset only containing 29 stocks can lead to bias when analyzing the data.

A closer inspection of figure 8.5 suggests that the DJ index in general is anti-correlated to $\lambda_1$. However, also periods where the index is positively correlated to $\lambda_1$ are observed. To obtain a better overview over these correlations, it is useful to consider the running correlation between the index and $\lambda_1$. The running correlation is calculated from a time-window of length $T = 200$ days, such that $\rho_{\mathrm{I},\lambda_1}(t_k)$ is based on time-window $\mathbf{M}_k$ covering the period $t \in [t_k - T/2, t_k + T/2]$. Note that $t_k = T/2 + n_k X$ where $t_k \leq t_{\max}$, $n_k$ is an integer starting from zero and $t_{\max}$ equals the length of the series

describing $\lambda_1$. Using the smallest obtainable step-size $X = 1$ day leads to the result presented in figure 8.6.

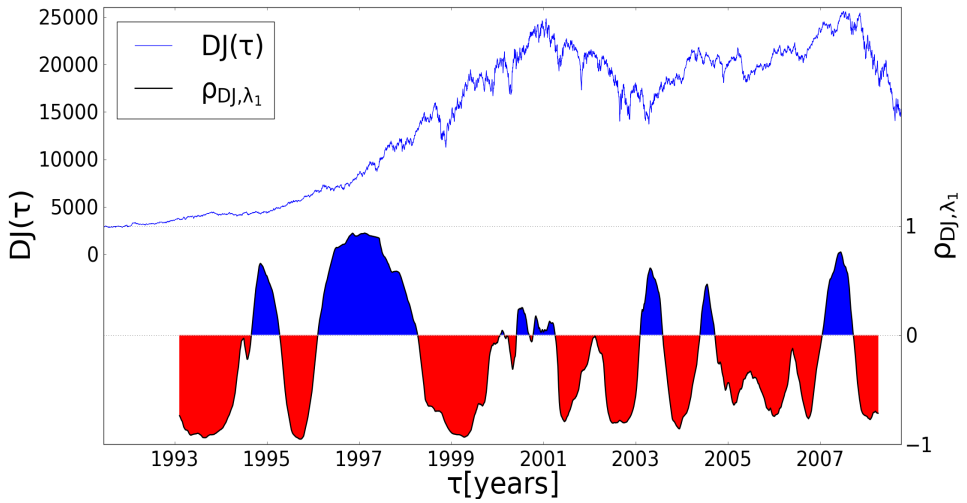



Figure 8.6: Blue line: The DJ index. Black line: The running correlation between the DJ index and the largest eigenvalue $\lambda_1$. Notice that the area below the curve is shaded red (blue) for negative (positive) correlations to guide the eye. The step-size used to calculate the movements of $\lambda_1$ was set to $X = 1$ day, and the length of the time-window used to calculate the running correlation is $T = 200$ days.

An inspection of figure 8.6 reveals that the leading trend is an anti-correlation between the index and $\lambda_1$, with the exception of a longer period lasting from roughly mid 1996 to mid 1998. During this period, the correlation between the index and $\lambda_1$ is strictly positive. The index rises during the period, indicating an increasing collectivity in the market. As discussed for the high-frequency dataset, this is very interesting and suggests the existence of the *optimism factor*, causing stocks to rise synchronously.

In general, the index and $\lambda_1$ are anti-correlated in the period after 1998, but short periods of positive correlations are observed from time to time. These positive correlations seem to arise during 'transition periods' near peaks and bottoms of the index, where the index stops following the trend that has governed the market so far. The presence of positive correlations during these periods is speculated to reflect that the market believes that the rise or drop now has come to an end. In other words, investors are ready to 'jump on the train' if the index rises, or to make use of corrections to buy 'cheap' stocks. However, the general trend is indeed an anti-correlation between the index

and $\lambda_1$. This causes the strength of collectivity in the market to increase when the market drops, and oppositely to decrease when the market rises. For the high-frequency data, one could not conclude whether the decreasing strength of collective trends was caused by the rebound that followed the drop, or that the collectivity would decrease also if the drop was followed by a calm period. The determination of this would require a dataset containing periods where the index rises sharply, but not in the vicinity of a preceding drop. Therefore, the only conclusion that could be drawn for the high-frequency data was that a sharply rising index in general was *not* connected to an increasing collectivity in the market. However, the current dataset reveals that periods with a sharply rising index in general are related to a decreasing collectivity. This follows, as periods where the index drops and rises, both characterized by an anti-correlation to $\lambda_1$, are separated by the short 'transition' periods characterized by positive correlations between the index and $\lambda_1$. As recently discussed, also optimistic periods where the index rises and is positively correlated to $\lambda_1$ are present, but these periods are rare. Large drops on the other hand seem to always correspond to a strongly increasing collectivity. These observations strongly suggests that the collective trends are stronger during falling than rising markets.

To summarize, it is observed that the index in general is anti-correlated to $\lambda_1$. As $\lambda_1$ monitors the degree of collectivity in the market, this implies that the strength of collective trends increases during periods where the index drops and oppositely decreases when the index rises. It also seems to exist a mode where the index is positively correlated to $\lambda_1$. This mode is speculated to reflect the presence of increased *optimism* in the market. During a rising market, the mode is believed to reflect investors 'jumping on the train'. Similarly, when the index exhibits corrections, the mode is believed to reflect investors using the correction to buy 'cheap' stocks and hence decrease the collectivity by deviating from existing trends. However, the general trend is an increasing collectivity during falling markets, and oppositely a decreasing collectivity during rising markets. This suggests that collective trends are stronger during periods with falling than rising markets, providing further support to the findings of Balogh et al. [40].

# Chapter 9

# Conclusion

Johansen et al. [37] speculated that the origin behind the gain-loss asymmetry was some collective motion of stocks of the index. The fear-factor model of Donangelo et al. [42] provided qualitative support for this, reproducing an asymmetry in the index but not in individual stocks. Balogh et al. [40] performed several statistical tests on daily closure prices from the Dow Jones Industrial Average (DJIA), and found that collective trends indeed were stronger during falling than rising markets. The work in this thesis has been performed in order to provide further support to their conclusion, and the current chapter will attempt to sum up the results.

The first dataset under consideration was of high-frequency, consisting of one minute stock quotes from 492 large European companies over a period of roughly 6 months. A price-weighted index based on these stocks was found to exhibit a significant gain-loss asymmetry, which has also been observed for daily closure prices of the DJIA [38]. What is interesting is that, opposite to earlier observations, the probability of the optimal investment horizon is highest for the gain distribution. Individual stocks were found to only exhibit gain-loss asymmetry for significantly increased return-levels. These observations are believed to reflect inherent properties of high-frequency stock quotes, and to the best of our knowledge this is the first inverse statistics analysis of high-frequency stock quotes.

The second dataset consisted of daily closure prices from 29 of the 30 stocks composing the DJIA late February 2008. The price-weighted index based on these stocks was observed to exhibit a clear gain-loss asymmetry, while individual stocks exhibited no or weak asymmetry for the same return-level. This is in accordance with earlier observations by Balogh et al. [40], based on a nearly identical dataset.

As both datasets demonstrated a clear gain-loss asymmetry, it was investigated whether there were any difference in strength of the collective trends during falling and rising markets. The high-frequency data revealed that collective trends got stronger during sharp index drops. It was also observed that the strength of collectivity decreased during periods where the index rises sharply. As periods where the index rises sharply only occurred after a sharp drop in the current dataset, it was not possible to conclude that the rise caused the collectivity to decrease. However, it could be concluded that the strongly rising index in general was *not* related to an increasing collectivity. These observations suggest that collective trends are stronger during falling than rising markets for the high-frequency stock quotes. An identical analysis on daily closure prices from the DJIA provided the same result for sharp index drops, but also this dataset only contained periods where the index rises sharply in the close vicinity of a drop. However, between the two regimes it was observed to be a 'transition' period. This period was characterized by an index that was positively correlated with the strength of collectivity, and separated the drop from the following rise. It could therefore be concluded that the rise caused the strength of collective trends to decrease. It follows that also this dataset suggests that the collective trends are stronger during falling than rising markets. It is believed that the reason behind both datasets showing the same features is that market factors incorporated in stocks are independent on the nature of the dataset. As both datasets indeed suggest that stronger collective trends are present during falling than rising markets, it is concluded that this work provides support to the results of Balogh et al. [40].

The rapidly increasing collectivity during sharp index drops supports the fear factor model of Donangelo et al. [42], where the key idea is that fear triggers stocks to fall synchronously causing a sharp index drop. Both datasets also contained periods where a rising index was positively correlated to the strength of collectivity in the market. However, such periods were less common and did not affect the conclusion that the collectivity appear to be stronger during falling than rising markets. What is interesting is that these periods imply that the collectivity in the market increases while the index rises, suggesting the existence of an *optimism factor* causing stocks to rise synchronously. It was also indicated that the size of the drop is connected to the amount the collectivity increases. This clearly suggests that the fear factor model [42] should incorporate not only a global fear factor affecting all stocks in the market, but also *individual* fear factors for each of the economic sectors, in addition to an optimism factor.

# Bibliography


[1] *http://nobelprize.org/nobel_prizes/economics/laureates/1997/press.html*, Nobelprice.org

[2] *Oslo stock exchange*, www.oslobors.no

[3] N. N. Taleb: *The Black Swan*, Random House Trade Paperbacks (2010)

[4] A. Chatterjee and B. K. Chakrabarti: *Econophysics of Stocks and Markets : Proceedings of the Econophys-Kolkata II*, Springer (2006)

[5] R. N. Mantegna and H. E. Stanley: *An Introduction to Econophyscs: Correlations and Complexity in Finance*, Cambridge University Press (2000)

[6] G. L. Vasconcelos: *A Guided Walk Down Wall Street: An Introduction to Econophyscs*, Brazilian Journal of Physics **34**, 3B, pp. 1039 - 1065 (2004)

[7] J. B. Rosser: *Econophysics*, James Madison University (2006)

[8] R. Dana and M. Jeanblanc: *Financial Markets in Continuous Time*, Springer (2003)

[9] B. Mandelbrot: *The Variation of Certain Speculative Prices*, Journal of Business **36**, 4, pp. 394 - 419 (1963)

[10] A. Einstein: *Investigations on the Theory of the Brownian Movement: On the Movement of Small Particles Suspended in a Stationary Liquid Demanded by the Molecular-Kinetic Theory of Heat*, Annalen der Physik **17**, pp. 549 - 560 (1905)

[11] E. F. Fama: *Efficient Capital Markets: A Review of Theory and Empirical Work*, The Journal of Finance **25**, pp. 383 - 417 (1969)

[12] M. T. Gapen, D. F. Gray, C. H. Lim, and Y. Xiao: *The Contingent Claims Approach to Corporate Vulnerability Analysis: Estimating Default Risk and Economy-Wide Risk Transfer*, International Monetary Fund (2004)

[13] M. Davis and A. Etheridge: *Louis Bachelier's Theory of Speculation*, Princeton University Press (2006)

[14] D. Chotikapanich: *Modeling Income Distributions and Lorenz Curves*, Springer (2008)

[15] L. Spadafora, G.P. Berman and F. Borgonovi: *Adiabaticity conditions for volatility smile in Black-Scholes pricing model*, European Physical Journal B **79**, pp. 47 - 53 (2011)

[16] J. Voit: *The statistical mechanics of financial markets*, Springer (2005)

[17] J. P. Bouchaud and M. Potters: *Theory of Financial Risk and Derivative Pricing*, Cambridge University Press (2003)

[18] J. M. Courtault, Y. Kabanov, B. Bru, P. Crèpel, I. Lebon and A. le Maarchand: *Louis Bachelier: On the Centenary of Thèorie de la Spèculation*, Mathematical Finance **10**, pp. 341 - 353 (2000)

[19] F. Black and M. Scholes: *The Pricing of Options and Corporate Liabilities*, The Journal of Political Economy **81**, pp. 637 - 654 (1973)

[20] J. Hull: *Options, Futures and Other Derivatives*, 7th edition, Pearson Prentice Hall (2009)

[21] *Yahoo! Finance*, http://finance.yahoo.com

[22] R. E. Walpole, R. H. Myers, S. L. Myers and K. Ye: *Probability & Statistics for Engineers & Scientists*, 8th edition, Pearson Prentice Hall (2007)

[23] B. S. Everitt: *The Cambridge Dictionary of Statistics*, 2nd edition, Cambridge University Press (2006)

[24] L. Laloux, P. Cizeau, M. Potters and J. P. Bouchaud: *Random Matrix Theory and Financial Correlations*, International Journal of Theoretical and Applied Finance **3**, pp. 391 - 397 (2000)

[25] L. Laloux, P. Cizeau, J. P. Bouchaud and M. Potters: *Noise Dressing of Financial Correlation Matrices*, Physical Review Letters **83**, pp. 1467-1470 (1999)

[26] S. Sharifi, M. Crane, A. Shamaie and H. Ruskin: *Random matrix theory for portfolio optimization: a stability approach*, Physica A **335**, pp. 629 - 643 (2004)

[27] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, T. Guhr, and H. E. Stanley: *Random matrix approach to cross correlations in financial data*, Physical Review E **65**, 066126 (2002)

[28] V. Kulkarni and N. Deo: *Correlation and volatility in an Indian stock market: A random matrix approach*, The European Physics Journal B **60**, pp. 101 - 109 (2007)

[29] A. Namakia, G.R. Jafari and R. Raei: *Comparing the structure of an emerging market with a mature one under global perturbation*, Physica A **390**, pp. 3020 - 3025 (2011)

[30] C. H. Edwards, D. E. Penny: *Elementary Linear Algebra*, Prentice Hall (2005)

[31] D. Poole: *Linear algebra: a modern introduction*, pp. 264 - 265, 2nd edition, Thomson Brooke / Cole (2006)

[32] R. A. Horn and C. R. Johnson: *Matrix Analysis*, Camebridge University Press (1990)

[33] R. Cont: *Empirical Properties of Asset Returns: Stylized Facts and Statistical Issues* Quantitative Finance **1**, pp. 223 - 236 (2001)

[34] P. T. H. Ahlgren, H. Dahl, M. H. Jensen and I. Simonsen: *Econophysics Approaches to Large-Scale Business Data and Financial Crisis* in M. Takayasu, T. Watanabe and H. Takayasu: *Econophysics Approaches to Large-Scale Business Data and Financial Crisis*, pp. 247 - 270, Springer (2010)

[35] S. Karlin and H. M. Taylor: *A first course on Stochastic processes*, 2nd edition, Academic Press New York (1975)

[36] I. Simonsen, M. H. Jensen and A. Johansen: *Optimal Investment Horizons*, The European Physics Journal B **27**, pp. 583 - 586 (2002)

[37] A. Johansen, I. Simonsen and M. H. Jensen and : *Optimal investment horizons for stocks and markets*, Physica A **370**, pp. 64 - 67 (2006)

[38] M. H. Jensen, A. Johansen and I. Simonsen: *Inverse statistics in economics: The gain-loss asymmetry*, Physica A **324**, pp. 338 - 343 (2003)

[39] W. X. Zhou, W. K. Yuan: *Inverse statistics in stock markets: Universality and idiosyncracy*, Physica A **353**, pp. 433 - 444 (2005)

[40] E. Balogh, I. Simonsen, B. Z. Nagy and Z. Nèda: *Persistent collective trend in stock markets*, Physical Review E **82**, 066113 (2010)

[41] I. Simonsen, P. T. H. Ahlgren, M. H. Jensen, R. Donangelo and K. Sneppen: *Fear and its implications for stock markets*, The European Physics Journal B **57**, pp. 153 - 158 (2007)

[42] R. Donangelo, M. H. Jensen, I. Simonsen and K. Sneppen: *Synchronization model for stock market asymmetry*, Journal of Statistical Mechanics: Theory and Experiment, L11001 (2006)

[43] J. Siven, J. Lins and J. L. Hansen: *A multiscale view on inverse statistics and gain/loss asymmetry in financial time series*, Journal of Statistical Mechanics: Theory and Experiment, P02004 (2009)

[44] P. T. H. Ahlgren, M. H. Jensen, I. Simonsen, R. Donangelo and K. Sneppen: *Frustration driven stock market dynamics: Leverage effect and asymmetry*, Physica A **383**, pp. 1 - 4 (2007)

[45] M. H. Jensen, A. Johansen, F. Petroni and I. Simonsen: *Inverse statistics in the foreign exchange market*, Physica A **340**, pp. 678 - 684 (2004)

[46] J. Wishart: *The generalised product moment distribution in samples from a normal multivariate population*, Biometrika **20** A, pp. 32 - 52 (1928)

[47] P. Wigner: *Characteristic Vectors of Bordered Matrices With Infinite Dimensions*, The Annals of Mathematics **62**, pp. 548 - 564 (1955)

[48] A. M. Tulino and S. Verdú: *Random matrix theory and wireless communications*, Publishers Inc. (2004)

[49] M. L. Metha: *Random matrices*, Elsvier (2004)

[50] J. V. Siven and J. T. Lins: *Gain/loss asymmetry in time series of individual stock prices and its relationship to the leverage effect*, Quantitative Finance Papers (2009) (available at http://arxiv.org/abs/0911.4679)

[51] J. P. Bouchaud, A. Matacz and M. Potters: *Leverage Effect in Financial Markets : The Retarded Volatility Model*, Physical Review Letters **87** (22), 228701 (2001)

[52] *Python*, http://www.python.org/

[53] W. H. Press, S. A. Teukolsky, W. T. Vetterling and B. P. Flannery: *Numerical Recipes in C - The Art of Scientific Computing*, 2nd edition, Cambridge University Press (1992)

[54] *Dow Jones Indexes*, www.djindexes.com

[55] *Standard & Poor's*, http://www.standardandpoors.com/home/en/eu

[56] *MSCI*, http://www.msci.com/

[57] C. M. S. Sutcliffe: *Stock index futures*, 3rd edition, Ashgate Publishing Company (2006)

[58] M. Müller, G. Baier, A. Galka, U. Stephani and H. Muhle: *Detection and characterization of changes of the correlation structure in multivariate time series*, Physical Review E **71**, 046116 (2005)

# Appendix A

# Companies in the datasets

## A.1 The high-frequency dataset

Table A.1 contains (sorted alphabetically after ticker code) information about the 492 companies considered in this thesis. It contains the ticker code of the companies on the various stock exchanges as well as their nationality and Global Industry Classification Standard (GICS) sector. The data have been recorded by Bloomberg, and the information about the GICS sector and country has been provided from them.

Table A.1: Information about the companies in the high-frequency dataset used in the thesis.

| Ticker | Country | GICS Sector-name |
|---|---|---|
| A2A IM Equity | Italy | Utilities |
| AAL LN Equity | Britain | Materials |
| ABBN VX Equity | Switzerland | Industrials |
| ABE SM Equity | Spain | Industrials |
| ABF LN Equity | Britain | Consumer Staples |
| ABG LN Equity | Britain | Materials |
| ABG SM Equity | Spain | Industrials |
| ABI BB Equity | Belgium | Consumer Staples |
| AC FP Equity | France | Consumer Discretionary |
| ACA FP Equity | France | Financials |
| ACE IM Equity | Italy | Utilities |
| ACS SM Equity | Spain | Industrials |
| ACX SM Equity | Spain | Materials |
| ADEN VX Equity | Switzerland | Industrials |

Continued on next page

**Table A.1 Continued from previous page**

| Ticker | Country | GICS Sector-name |
|---|---|---|
| ADM LN Equity | Britain | Financials |
| ADN LN Equity | Britain | Financials |
| ADP FP Equity | France | Industrials |
| ADS GR Equity | Germany | Consumer Discretionary |
| AF FP Equity | France | Industrials |
| AGK LN Equity | Britain | Industrials |
| AGL IM Equity | Italy | Consumer Discretionary |
| AGN NA Equity | Netherlands | Financials |
| AGS BB Equity | Belgium | Financials |
| AH NA Equity | Netherlands | Consumer Staples |
| AI FP Equity | France | Materials |
| AIXA GR Equity | Germany | Information Technology |
| AKE FP Equity | France | Materials |
| AKSO NO Equity | Norway | Energy |
| AKZA NA Equity | Netherlands | Materials |
| ALFA SS Equity | Sweden | Industrials |
| ALO FP Equity | France | Industrials |
| ALPHA GA Equity | Greece | Financials |
| ALU FP Equity | France | Information Technology |
| ALV GR Equity | Germany | Financials |
| AMEC LN Equity | Britain | Energy |
| AML LN Equity | Britain | Financials |
| ANA SM Equity | Spain | Utilities |
| ANDR AV Equity | Austria | Industrials |
| ANTO LN Equity | Britain | Materials |
| ARM LN Equity | Britain | Information Technology |
| ARYN SW Equity | Switzerland | Consumer Staples |
| ASHM LN Equity | Britain | Financials |
| ASML NA Equity | Netherlands | Information Technology |
| ASSAB SS Equity | Sweden | Industrials |
| ATCOA SS Equity | Sweden | Industrials |
| ATL IM Equity | Italy | Industrials |
| ATLN VX Equity | Switzerland | Health Care |
| ATO FP Equity | France | Information Technology |
| AU/ LN Equity | Britain | Information Technology |
| AV LN Equity | Britain | Financials |
| AZN LN Equity | Britain | Health Care |
| BA LN Equity | Britain | Industrials |
| BAB LN Equity | Britain | Industrials |

Continued on next page

**Table A.1 Continued from previous page**

| Ticker | Country | GICS Sector-name |
|---|---|---|
| BAER VX Equity | Switzerland | Financials |
| BALN VX Equity | Switzerland | Financials |
| BARC LN Equity | Britain | Financials |
| BAS GR Equity | Germany | Materials |
| BATS LN Equity | Britain | Consumer Staples |
| BAYN GR Equity | Germany | Health Care |
| BBVA SM Equity | Spain | Financials |
| BBY LN Equity | Britain | Industrials |
| BCP PL Equity | Portugal | Financials |
| BEI GR Equity | Germany | Consumer Staples |
| BELG BB Equity | Belgium | Telecommunication Services |
| BES PL Equity | Portugal | Financials |
| BG LN Equity | Britain | Energy |
| BKIR ID Equity | Ireland | Financials |
| BKT SM Equity | Spain | Financials |
| BLND LN Equity | Britain | Financials |
| BLT LN Equity | Britain | Materials |
| BMPS IM Equity | Italy | Financials |
| BMW GR Equity | Germany | Consumer Discretionary |
| BN FP Equity | France | Consumer Staples |
| BNP FP Equity | France | Financials |
| BNZL LN Equity | Britain | Industrials |
| BOKA NA Equity | Netherlands | Industrials |
| BOL SS Equity | Sweden | Materials |
| BP IM Equity | Italy | Financials |
| BP LN Equity | Britain | Energy |
| BPE IM Equity | Italy | Financials |
| BPSO IM Equity | Italy | Financials |
| BRBY LN Equity | Britain | Consumer Discretionary |
| BRI PL Equity | Portugal | Industrials |
| BSY LN Equity | Britain | Consumer Discretionary |
| BTA LN Equity | Britain | Telecommunication Services |
| BTO SM Equity | Spain | Financials |
| BUL IM Equity | Italy | Consumer Discretionary |
| BVA SM Equity | Spain | Financials |
| BVI FP Equity | France | Industrials |
| CA FP Equity | France | Consumer Staples |
| CAP FP Equity | France | Information Technology |
| CARLB DC Equity | Denmark | Consumer Staples |

Continued on next page

**Table A.1 Continued from previous page**

| Ticker | Country | GICS Sector-name |
|---|---|---|
| CBK GR Equity | Germany | Financials |
| CCL LN Equity | Britain | Consumer Discretionary |
| CDI FP Equity | France | Consumer Discretionary |
| CFR VX Equity | Switzerland | Consumer Discretionary |
| CLN VX Equity | Switzerland | Materials |
| CLS1 GR Equity | Germany | Health Care |
| CNA LN Equity | Britain | Utilities |
| CNE LN Equity | Britain | Energy |
| CNP FP Equity | France | Financials |
| CO FP Equity | France | Consumer Staples |
| COB LN Equity | Britain | Industrials |
| CON GR Equity | Germany | Consumer Discretionary |
| CORA NA Equity | Netherlands | Financials |
| CPG LN Equity | Britain | Consumer Discretionary |
| CPI LN Equity | Britain | Industrials |
| CPR IM Equity | Italy | Consumer Staples |
| CPR PL Equity | Portugal | Materials |
| CRDA LN Equity | Britain | Materials |
| CRG IM Equity | Italy | Financials |
| CRH ID Equity | Ireland | Materials |
| CRI SM Equity | Spain | Financials |
| CS FP Equity | France | Financials |
| CSCG LN Equity | Britain | Financials |
| CSGN VX Equity | Switzerland | Financials |
| CSM NA Equity | Netherlands | Consumer Staples |
| CW LN Equity | Britain | Telecommunication Services |
| CWC LN Equity | Britain | Telecommunication Services |
| DAI GR Equity | Germany | Consumer Discretionary |
| DANSKE DC Equity | Denmark | Financials |
| DB1 GR Equity | Germany | Financials |
| DBK GR Equity | Germany | Financials |
| DCO DC Equity | Denmark | Consumer Staples |
| DEC FP Equity | France | Consumer Discretionary |
| DELB BB Equity | Belgium | Consumer Staples |
| DEXB BB Equity | Belgium | Financials |
| DG FP Equity | France | Industrials |
| DGE LN Equity | Britain | Consumer Staples |
| DIA IM Equity | Italy | Health Care |
| DL NA Equity | Netherlands | Financials |

Continued on next page

**Table A.1 Continued from previous page**

| Ticker | Country | GICS Sector-name |
|---|---|---|
| DMGT LN Equity | Britain | Consumer Discretionary |
| DNBNOR NO Equity | Norway | Financials |
| DPB GR Equity | Germany | Financials |
| DPW GR Equity | Germany | Industrials |
| DRX LN Equity | Britain | Utilities |
| DSM NA Equity | Netherlands | Materials |
| DSV DC Equity | Denmark | Industrials |
| DSY FP Equity | France | Information Technology |
| DTE GR Equity | Germany | Telecommunication Services |
| EAD FP Equity | Netherlands | Industrials |
| EBRO SM Equity | Spain | Consumer Staples |
| EBS AV Equity | Austria | Financials |
| EDF FP Equity | France | Utilities |
| EDN IM Equity | Italy | Utilities |
| EDP PL Equity | Portugal | Utilities |
| EDPR PL Equity | Spain | Utilities |
| EEEK GA Equity | Greece | Consumer Staples |
| EI FP Equity | France | Health Care |
| EKTAB SS Equity | Sweden | Health Care |
| ELE SM Equity | Spain | Utilities |
| ELI1V FH Equity | Finland | Telecommunication Services |
| ELN ID Equity | Ireland | Health Care |
| ELPE GA Equity | Greece | Energy |
| ELUXB SS Equity | Sweden | Consumer Discretionary |
| EMG LN Equity | Britain | Financials |
| EN FP Equity | France | Industrials |
| ENEL IM Equity | Italy | Utilities |
| ENG SM Equity | Spain | Utilities |
| ENI IM Equity | Italy | Energy |
| ENRC LN Equity | Britain | Materials |
| EOAN GR Equity | Germany | Utilities |
| ERICB SS Equity | Sweden | Information Technology |
| ETE GA Equity | Greece | Financials |
| ETL FP Equity | France | Consumer Discretionary |
| EUROB GA Equity | Greece | Financials |
| EXO IM Equity | Italy | Financials |
| EXPN LN Equity | Ireland | Industrials |
| EZJ LN Equity | Britain | Industrials |
| F IM Equity | Italy | Consumer Discretionary |

Continued on next page

**Table A.1 Continued from previous page**

| Ticker | Country | GICS Sector-name |
|---|---|---|
| FCC SM Equity | Spain | Industrials |
| FER SM Equity | Spain | Industrials |
| FGP LN Equity | Britain | Industrials |
| FGR FP Equity | France | Industrials |
| FLS DC Equity | Denmark | Industrials |
| FME GR Equity | Germany | Health Care |
| FNC IM Equity | Italy | Industrials |
| FP FP Equity | France | Energy |
| FR FP Equity | France | Consumer Discretionary |
| FRA GR Equity | Germany | Industrials |
| FTE FP Equity | France | Telecommunication Services |
| FUM1V FH Equity | Finland | Utilities |
| FXPO LN Equity | Switzerland | Materials |
| G IM Equity | Italy | Financials |
| G1A GR Equity | Germany | Industrials |
| GA FP Equity | France | Energy |
| GALP PL Equity | Portugal | Energy |
| GAM SM Equity | Spain | Industrials |
| GAM SW Equity | Switzerland | Financials |
| GAS SM Equity | Spain | Utilities |
| GBB FP Equity | France | Energy |
| GBF GR Equity | Germany | Industrials |
| GBLB BB Equity | Belgium | Financials |
| GET FP Equity | France | Industrials |
| GETIB SS Equity | Sweden | Health Care |
| GFS LN Equity | Britain | Industrials |
| GKN LN Equity | Britain | Consumer Discretionary |
| GLE FP Equity | France | Financials |
| GRF SM Equity | Spain | Health Care |
| GSK LN Equity | Britain | Health Care |
| GSZ FP Equity | France | Utilities |
| GTO FP Equity | France | Information Technology |
| HEI GR Equity | Germany | Materials |
| HEIA NA Equity | Netherlands | Consumer Staples |
| HEIO NA Equity | Netherlands | Consumer Staples |
| HEXAB SS Equity | Sweden | Industrials |
| HHFA GR Equity | Germany | Industrials |
| HL LN Equity | Britain | Financials |
| HMB SS Equity | Sweden | Consumer Discretionary |

Continued on next page

**Table A.1 Continued from previous page**

| Ticker | Country | GICS Sector-name |
|---|---|---|
| HMSO LN Equity | Britain | Financials |
| HNR1 GR Equity | Germany | Financials |
| HO FP Equity | France | Industrials |
| HOLMB SS Equity | Sweden | Materials |
| HOLN VX Equity | Switzerland | Materials |
| HOME LN Equity | Britain | Consumer Discretionary |
| HOT GR Equity | Germany | Industrials |
| HSBA LN Equity | Britain | Financials |
| HSV LN Equity | Britain | Industrials |
| HTO GA Equity | Greece | Telecommunication Services |
| HUSQB SS Equity | Sweden | Consumer Discretionary |
| IAP LN Equity | Britain | Financials |
| IBE SM Equity | Spain | Utilities |
| IBR SM Equity | Spain | Utilities |
| IDR SM Equity | Spain | Information Technology |
| IFX GR Equity | Germany | Information Technology |
| IGG LN Equity | Britain | Financials |
| IHG LN Equity | Britain | Consumer Discretionary |
| IIA AV Equity | Austria | Financials |
| III LN Equity | Britain | Financials |
| ILD FP Equity | France | Telecommunication Services |
| IM NA Equity | Netherlands | Industrials |
| IMI LN Equity | Britain | Industrials |
| IMT LN Equity | Britain | Consumer Staples |
| INDUA SS Equity | Sweden | Financials |
| INF LN Equity | Switzerland | Consumer Discretionary |
| INVEB SS Equity | Sweden | Financials |
| INVP LN Equity | Britain | Financials |
| IPN FP Equity | France | Health Care |
| IPR LN Equity | Britain | Utilities |
| ISAT LN Equity | Britain | Telecommunication Services |
| ISP IM Equity | Italy | Financials |
| ISYS LN Equity | Britain | Industrials |
| ITRK LN Equity | Britain | Industrials |
| ITV LN Equity | Britain | Consumer Discretionary |
| ITX SM Equity | Spain | Consumer Discretionary |
| JMAT LN Equity | Britain | Materials |
| JMT PL Equity | Portugal | Consumer Staples |
| KAZ LN Equity | Britain | Materials |

Continued on next page

**Table A.1 Continued from previous page**

| Ticker | Country | GICS Sector-name |
|---|---|---|
| KBC BB Equity | Belgium | Financials |
| KD8 GR Equity | Germany | Consumer Discretionary |
| KESBV FH Equity | Finland | Consumer Staples |
| KGF LN Equity | Britain | Consumer Discretionary |
| KINVB SS Equity | Sweden | Financials |
| KN FP Equity | France | Financials |
| KNEBV FH Equity | Finland | Industrials |
| KNIN VX Equity | Switzerland | Industrials |
| KPN NA Equity | Netherlands | Telecommunication Services |
| KYG ID Equity | Ireland | Consumer Staples |
| LAND LN Equity | Britain | Financials |
| LG FP Equity | France | Materials |
| LGEN LN Equity | Britain | Financials |
| LHA GR Equity | Germany | Industrials |
| LI FP Equity | France | Financials |
| LIN GR Equity | Germany | Materials |
| LLOY LN Equity | Britain | Financials |
| LMI LN Equity | Britain | Materials |
| LOG LN Equity | Britain | Information Technology |
| LOGN VX Equity | Switzerland | Information Technology |
| LONN VX Equity | Switzerland | Health Care |
| LR FP Equity | France | Industrials |
| LSE LN Equity | Britain | Financials |
| LTO IM Equity | Italy | Consumer Discretionary |
| LUN DC Equity | Denmark | Health Care |
| LUX IM Equity | Italy | Consumer Discretionary |
| LXS GR Equity | Germany | Materials |
| MAN GR Equity | Germany | Industrials |
| MAP SM Equity | Spain | Financials |
| MB IM Equity | Italy | Financials |
| MC FP Equity | France | Consumer Discretionary |
| MED IM Equity | Italy | Financials |
| MEDAA SS Equity | Sweden | Health Care |
| MEO GR Equity | Germany | Consumer Staples |
| MEO1V FH Equity | Finland | Industrials |
| MF FP Equity | France | Industrials |
| MGGT LN Equity | Britain | Industrials |
| MHG NO Equity | Norway | Consumer Staples |
| MKS LN Equity | Britain | Consumer Discretionary |

Continued on next page

**Table A.1 Continued from previous page**

| Ticker | Country | GICS Sector-name |
|---|---|---|
| ML FP Equity | France | Consumer Discretionary |
| MMB FP Equity | France | Consumer Discretionary |
| MMT FP Equity | France | Consumer Discretionary |
| MNDI LN Equity | Britain | Materials |
| MOBB BB Equity | Belgium | Telecommunication Services |
| MRK GR Equity | Germany | Health Care |
| MRW LN Equity | Britain | Consumer Staples |
| MS IM Equity | Italy | Consumer Discretionary |
| MT NA Equity | Luxembourg | Materials |
| MTGB SS Equity | Sweden | Consumer Discretionary |
| MTX GR Equity | Germany | Industrials |
| MUV2 GR Equity | Germany | Financials |
| NDA SS Equity | Sweden | Financials |
| NEO FP Equity | France | Information Technology |
| NES1V FH Equity | Finland | Energy |
| NESN VX Equity | Switzerland | Consumer Staples |
| NG LN Equity | Britain | Utilities |
| NHY NO Equity | Norway | Materials |
| NK FP Equity | France | Materials |
| NOBN VX Equity | Switzerland | Health Care |
| NOK1V FH Equity | Finland | Information Technology |
| NOVN VX Equity | Switzerland | Health Care |
| NOVOB DC Equity | Denmark | Health Care |
| NRE1V FH Equity | Finland | Consumer Discretionary |
| NWG LN Equity | Britain | Utilities |
| NWR LN Equity | Netherlands | Materials |
| NXT LN Equity | Britain | Consumer Discretionary |
| NZYMB DC Equity | Denmark | Materials |
| OHL SM Equity | Spain | Industrials |
| OML LN Equity | Britain | Financials |
| OMV AV Equity | Austria | Energy |
| OPAP GA Equity | Greece | Consumer Discretionary |
| OR FP Equity | France | Consumer Staples |
| ORK NO Equity | Norway | Industrials |
| ORNBV FH Equity | Finland | Health Care |
| OUT1V FH Equity | Finland | Materials |
| PAJ FP Equity | France | Consumer Discretionary |
| PC IM Equity | Italy | Consumer Discretionary |
| PFC LN Equity | Britain | Energy |

Continued on next page

**Table A.1 Continued from previous page**

| Ticker | Country | GICS Sector-name |
|---|---|---|
| PHIA NA Equity | Netherlands | Industrials |
| PLT IM Equity | Italy | Consumer Staples |
| PMO LN Equity | Britain | Energy |
| PNN LN Equity | Britain | Utilities |
| POG LN Equity | Britain | Materials |
| POH1S FH Equity | Finland | Financials |
| POP SM Equity | Spain | Financials |
| PP FP Equity | France | Consumer Discretionary |
| PPC GA Equity | Greece | Utilities |
| PRU LN Equity | Britain | Financials |
| PRY IM Equity | Italy | Industrials |
| PSON LN Equity | Britain | Consumer Discretionary |
| PSPN SW Equity | Switzerland | Financials |
| PTC PL Equity | Portugal | Telecommunication Services |
| PUB FP Equity | France | Consumer Discretionary |
| PZC LN Equity | Britain | Consumer Staples |
| RAND NA Equity | Netherlands | Industrials |
| RATOB SS Equity | Sweden | Financials |
| RB LN Equity | Britain | Consumer Staples |
| RBS LN Equity | Britain | Financials |
| RDSA LN Equity | Netherlands | Energy |
| REC NO Equity | Norway | Energy |
| REE SM Equity | Spain | Utilities |
| REL LN Equity | Britain | Consumer Discretionary |
| REN NA Equity | Netherlands | Consumer Discretionary |
| REP SM Equity | Spain | Energy |
| REX LN Equity | Britain | Materials |
| RHK GR Equity | Germany | Health Care |
| RHM GR Equity | Germany | Industrials |
| RI FP Equity | France | Consumer Staples |
| RIO LN Equity | Britain | Materials |
| RNO FP Equity | France | Consumer Discretionary |
| ROG VX Equity | Switzerland | Health Care |
| RR LN Equity | Britain | Industrials |
| RRS LN Equity | Jersey | Materials |
| RSA LN Equity | Britain | Financials |
| RSL LN Equity | Guernsey | Financials |
| RTO LN Equity | Britain | Industrials |
| RTRKS FH Equity | Finland | Materials |

Continued on next page

**Table A.1 Continued from previous page**

| Ticker | Country | GICS Sector-name |
|---|---|---|
| RUKN VX Equity | Switzerland | Financials |
| RWE GR Equity | Germany | Utilities |
| RXL FP Equity | France | Industrials |
| RYA ID Equity | Ireland | Industrials |
| SAA1V FH Equity | Finland | Consumer Discretionary |
| SAB LN Equity | Britain | Consumer Staples |
| SAB SM Equity | Spain | Financials |
| SAF FP Equity | France | Industrials |
| SAMAS FH Equity | Finland | Financials |
| SAN FP Equity | France | Health Care |
| SAN SM Equity | Spain | Financials |
| SAND SS Equity | Sweden | Industrials |
| SAP GR Equity | Germany | Information Technology |
| SBMO NA Equity | Netherlands | Energy |
| SBRY LN Equity | Britain | Consumer Staples |
| SCAB SS Equity | Sweden | Materials |
| SCH NO Equity | Norway | Consumer Discretionary |
| SCHP SW Equity | Switzerland | Industrials |
| SCMN VX Equity | Switzerland | Telecommunication Services |
| SCR FP Equity | France | Financials |
| SCVB SS Equity | Sweden | Industrials |
| SDF GR Equity | Germany | Materials |
| SDR LN Equity | Britain | Financials |
| SEBA SS Equity | Sweden | Financials |
| SECUB SS Equity | Sweden | Industrials |
| SEV FP Equity | France | Utilities |
| SGE LN Equity | Britain | Information Technology |
| SGO FP Equity | France | Industrials |
| SGRO LN Equity | Britain | Financials |
| SHBA SS Equity | Sweden | Financials |
| SHP LN Equity | Ireland | Health Care |
| SIE GR Equity | Germany | Industrials |
| SK FP Equity | France | Consumer Discretionary |
| SKAB SS Equity | Sweden | Industrials |
| SKFB SS Equity | Sweden | Industrials |
| SL LN Equity | Britain | Financials |
| SLHN VX Equity | Switzerland | Financials |
| SMIN LN Equity | Britain | Industrials |
| SN LN Equity | Britain | Health Care |

Continued on next page

**Table A.1 Continued from previous page**

| Ticker | Country | GICS Sector-name |
|---|---|---|
| SOLB BB Equity | Belgium | Materials |
| SOON VX Equity | Switzerland | Health Care |
| SOW GR Equity | Germany | Information Technology |
| SPM IM Equity | Italy | Energy |
| SPSN SW Equity | Switzerland | Financials |
| SRG IM Equity | Italy | Utilities |
| SRP LN Equity | Britain | Industrials |
| SSABA SS Equity | Sweden | Materials |
| SSE LN Equity | Britain | Utilities |
| STAN LN Equity | Britain | Financials |
| STB NO Equity | Norway | Financials |
| STERV FH Equity | Finland | Materials |
| STL NO Equity | Norway | Energy |
| STM IM Equity | Switzerland | Information Technology |
| SU FP Equity | France | Industrials |
| SUN SW Equity | Switzerland | Industrials |
| SVT LN Equity | Britain | Utilities |
| SW FP Equity | France | Consumer Discretionary |
| SWEDA SS Equity | Sweden | Financials |
| SWMA SS Equity | Sweden | Consumer Staples |
| SY1 GR Equity | Germany | Materials |
| SYNN VX Equity | Switzerland | Materials |
| SZG GR Equity | Germany | Materials |
| SZU GR Equity | Germany | Consumer Staples |
| TATE LN Equity | Britain | Consumer Staples |
| TCG LN Equity | Britain | Consumer Discretionary |
| TEC FP Equity | France | Energy |
| TEF SM Equity | Spain | Telecommunication Services |
| TEL NO Equity | Norway | Telecommunication Services |
| TEL2B SS Equity | Sweden | Telecommunication Services |
| TEN IM Equity | Luxembourg | Energy |
| TFI FP Equity | France | Consumer Discretionary |
| TGM GR Equity | Germany | Industrials |
| TIT IM Equity | Italy | Telecommunication Services |
| TKA AV Equity | Austria | Telecommunication Services |
| TKA GR Equity | Germany | Materials |
| TL5 SM Equity | Spain | Consumer Discretionary |
| TLSN SS Equity | Sweden | Telecommunication Services |
| TLW LN Equity | Britain | Energy |

Continued on next page

**Table A.1 Continued from previous page**

| Ticker | Country | GICS Sector-name |
|---|---|---|
| TNET BB Equity | Belgium | Telecommunication Services |
| TNT NA Equity | Netherlands | Industrials |
| TPK LN Equity | Britain | Industrials |
| TRE SM Equity | Spain | Energy |
| TRN IM Equity | Italy | Utilities |
| TRYG DC Equity | Denmark | Financials |
| TSCO LN Equity | Britain | Consumer Staples |
| TT LN Equity | Britain | Consumer Discretionary |
| TUI1 GR Equity | Germany | Consumer Discretionary |
| UBI IM Equity | Italy | Financials |
| UBSN VX Equity | Switzerland | Financials |
| UCB BB Equity | Belgium | Health Care |
| UCG IM Equity | Italy | Financials |
| UG FP Equity | France | Consumer Discretionary |
| UHR VX Equity | Switzerland | Consumer Discretionary |
| UL FP Equity | France | Financials |
| ULVR LN Equity | Britain | Consumer Staples |
| UMI BB Equity | Belgium | Materials |
| UNI IM Equity | Italy | Financials |
| UPM1V FH Equity | Finland | Materials |
| UTDI GR Equity | Germany | Information Technology |
| UU LN Equity | Britain | Utilities |
| VED LN Equity | Britain | Materials |
| VER AV Equity | Austria | Utilities |
| VIE FP Equity | France | Utilities |
| VIG AV Equity | Austria | Financials |
| VIV FP Equity | France | Consumer Discretionary |
| VK FP Equity | France | Industrials |
| VOD LN Equity | Britain | Telecommunication Services |
| VOE AV Equity | Austria | Materials |
| VOLVB SS Equity | Sweden | Industrials |
| VOW GR Equity | Germany | Consumer Discretionary |
| VPK NA Equity | Netherlands | Industrials |
| VWS DC Equity | Denmark | Industrials |
| WCH GR Equity | Germany | Materials |
| WDH DC Equity | Denmark | Health Care |
| WEIR LN Equity | Britain | Industrials |
| WG LN Equity | Britain | Energy |
| WKL NA Equity | Netherlands | Consumer Discretionary |

Continued on next page

**Table A.1 Continued from previous page**

| Ticker | Country | GICS Sector-name |
|---|---|---|
| WOS LN Equity | Britain | Industrials |
| WPP LN Equity | Ireland | Consumer Discretionary |
| WRT1V FH Equity | Finland | Industrials |
| WTB LN Equity | Britain | Consumer Discretionary |
| XTA LN Equity | Switzerland | Materials |
| YAR NO Equity | Norway | Materials |
| YTY1V FH Equity | Finland | Industrials |
| ZC FP Equity | France | Industrials |
| ZOT SM Equity | Spain | Industrials |
| ZURN VX Equity | Switzerland | Financials |

## A.2 The dataset of daily closure prices from the DJIA

Table A.2: Information about the companies in the DJIA dataset used in the thesis.

| Company name | Ticker | GICS Sector-name |
|---|---|---|
| 3M Company | MMM | Industrials |
| Coca-Cola Company | CO | Consumer Staples |
| J.P. Morgan Chase & Company | JPM | Financials |
| Alcoa Incorporated | AA | Materials |
| DuPont | DD | Materials |
| McDonaldŠs Corporation | MCD | Consumer Discretionary |
| American Express Company | AXP | Financials |
| Exxon Mobil Corporation | XOM | Energy |
| Merck & Company, Incorporated | MRK | Health Care |
| American International Group Inc. | AIG | Financials |
| General Electric Company | GE | Industrials |
| Microsoft Corporation | MSFT | Information Technology |
| AT&T Incorporated | T | Telecommunication Services |
| Pfizer Incorporated | PFE | Health Care |
| Bank of America Corporation | BAC | Financials |
| Hewlett-Packard Company | HPQ | Information Technology |
| Procter & Gamble Company | PG | Consumer Staples |
| Boeing Company | BA | Industrials |
| United Technologies Corporation | UTX | Industrials |
| Caterpillar Incorporated | CAT | Industrials |
| Intel Corporation | INTC | Information Technology |
| Verizon Communications Inc. | VZ | Telecommunication Services |
| Chevron Corporation | CVX | Energy |
| International Business Machines | IBM | Information Technology |
| Wal-Mart Stores Incorporated | WMT | Consumer Staples |
| Citigroup Incorporated | C | Financials |
| Johnson & Johnson | JNJ | Health Care |
| Walt Disney Company | DIS | Consumer Discretionary |
| Home Depot Incorporated | HD | Consumer Discretionary |

# Appendix B

# Fit parameters

## B.1 Parameters from least squares fit to empirical data

Table B.1 contains parameters obtained from least squares fits of empirical inverse statistics distributions from the high-frequency index to the generalized Gamma distribution given by equation (3.81).

Table B.2 contains the parameters from the fit of an exponential function of form $f(x) = a + b\ \exp(cx)$ to the daily closure prices of OSEBX and DJIA.

Table B.3 contains parameters obtained from least squares fits of empirical inverse statistics distributions from individual stocks to the generalized Gamma distribution given by equation (3.81). Notice that $\sigma_I$ corresponds to the minutely standard deviation of the index log-returns, and that $\sigma$ corresponds to the minutely standard deviation of the stock log-returns.

Table B.4 contains parameters from a fit of the generalized Gamma distribution to the empirical inverse statistics distributions of the constructed DJ index.

For the fits to the general Gamma distribution, $\tau^*_{\pm|\rho|}$ are the positions of the optimal investment horizons of the fitted curves, measured in minutes (days for the DJIA dataset). The difference, $\Delta\tau = \tau^*_{+|\rho|} - \tau^*_{-|\rho|}$ corresponds to the gain-loss asymmetry.

Table B.1: Fitting parameters inverse statistics for the high-frequency dataset

| $\rho$ | $\nu$ | $t_0$ | $\beta$ | $\alpha$ | $\tau^*_{-\lvert\rho\rvert}$ [min] | $\tau^*_{+\lvert\rho\rvert}$ [min] |
|---|---|---|---|---|---|---|
| $+1\sigma$ | 3.91287671 | 1.1602312 | 1.76805272 | 0.50 | - | 2.83 |
| $-1\sigma$ | 1.15795608 | -0.4876657 | 1.68162645 | 0.50 | 2.75 | - |
| $+2\sigma$ | 1.00938287 | -0.17392243 | 2.51518366 | 0.50 | - | 4.45 |
| $-2\sigma$ | 0.84240729 | -0.34626125 | 2.79803586 | 0.50 | 4.29 | - |
| $+3\sigma$ | 0.81963987 | -0.16849346 | 3.54990392 | 0.50 | - | 6.20 |
| $-3\sigma$ | 0.71077369 | -0.27986578 | 4.05346343 | 0.50 | 6.03 | - |
| $+4\sigma$ | 0.76955469 | 0.14481775 | 4.51049712 | 0.50 | - | 8.40 |
| $-4\sigma$ | 0.65307341 | -0.08070183 | 5.32342437 | 0.50 | 8.01 | - |
| $+5\sigma$ | 0.77196338 | 0.76709621 | 5.31216124 | 0.50 | - | 11.17 |
| $-5\sigma$ | 0.61346181 | -0.00986546 | 6.62411328 | 0.50 | 10.23 | - |
| $+6\sigma$ | 0.83544607 | 2.10260194 | 5.83528002 | 0.50 | - | 14.80 |
| $-6\sigma$ | 0.61506256 | 0.46787623 | 7.59714505 | 0.50 | 13.08 | - |
| $+7\sigma$ | 0.88816518 | 3.77059072 | 6.42763744 | 0.50 | - | 19.13 |
| $-7\sigma$ | 0.59527727 | 0.79661662 | 8.92143286 | 0.50 | 16.05 | - |
| $+8\sigma$ | 0.91742943 | 5.68608258 | 7.07681041 | 0.50 | - | 23.62 |
| $-8\sigma$ | 0.60055843 | 1.30733829 | 9.76013616 | 0.50 | 19.44 | - |
| $+9\sigma$ | 0.95011941 | 7.80474714 | 7.64061073 | 0.50 | - | 28.30 |
| $-9\sigma$ | 0.62708314 | 2.40676385 | 10.24977174 | 0.50 | 23.74 | - |
| $+10\sigma$ | 0.98578777 | 9.96310677 | 8.14509521 | 0.50 | - | 33.37 |
| $-10\sigma$ | 0.59726314 | 2.53420969 | 11.75602613 | 0.50 | 27.04 | - |
| $+11\sigma$ | 0.95014032 | 11.18986862 | 8.89972947 | 0.50 | - | 37.79 |
| $-11\sigma$ | 0.59987854 | 3.4450177 | 12.67294633 | 0.50 | 31.41 | - |
| $+12\sigma$ | 1.11380744 | 17.39294364 | 9.04814431 | 0.50 | - | 45.28 |
| $-12\sigma$ | 0.61684225 | 4.56464691 | 13.15664172 | 0.50 | 36.42 | - |
| $+13\sigma$ | 1.05898085 | 18.55938313 | 9.7791211 | 0.50 | - | 50.28 |
| $-13\sigma$ | 0.65563624 | 6.75104672 | 13.26575246 | 0.50 | 43.05 | - |
| $+14\sigma$ | 1.38429849 | 30.53451917 | 9.81548633 | 0.50 | - | 60.38 |
| $-14\sigma$ | 0.64676717 | 7.86571885 | 14.40977024 | 0.50 | 48.68 | - |
| $+15\sigma$ | 1.25080367 | 29.40070989 | 10.50239402 | 0.50 | - | 65.99 |
| $-15\sigma$ | 0.61384886 | 7.7609579 | 16.11135059 | 0.50 | 52.79 | - |
| $+16\sigma$ | 1.53367501 | 42.12073971 | 10.77521352 | 0.50 | - | 75.68 |
| $-16\sigma$ | 0.65890112 | 10.64251973 | 15.75207881 | 0.50 | 60.55 | - |

Table B.2: Fitting parameters OSEBX and DJIA

| Index | a | b | c |
|---|---|---|---|
| OSEBX | $-4.292 \cdot 10^{1}$ | $6.493 \cdot 10^{1}$ | $2.81 \cdot 10^{-4}$ |
| DJIA | $2.379 \cdot 10^{2}$ | $1.352 \cdot 10^{-3}$ | $5.531 \cdot 10^{-4}$ |

Table B.3: Fitting parameters inverse statistics for stocks XTA and ABBN from the high-frequency dataset

| Stock | $\rho$ | $\nu$ | $t_0$ | $\beta$ | $\alpha$ | $\tau^*_{-\lvert\rho\rvert}$ [min] | $\tau^*_{+\lvert\rho\rvert}$ [min] |
|---|---|---|---|---|---|---|---|
| XTA | $+7\sigma_I$ | 0.87701573 | -0.46290237 | 2.38462193 | 0.50 | - | 3.55 |
| XTA | $-7\sigma_I$ | 0.85519002 | -0.47585297 | 2.41261135 | 0.50 | 3.49 | - |
| ABBN | $+7\sigma_I$ | 0.89782871 | 0.29642801 | 3.48929756 | 0.50 | - | 6.58 |
| ABBN | $-7\sigma_I$ | 0.83172851 | 0.168488 | 3.6744856 | 0.50 | 6.48 | - |
| XTA | $+7\sigma$ | 0.85150773 | 5.27118889 | 7.58201821 | 0.50 | - | 24.29 |
| XTA | $-7\sigma$ | 0.67273629 | 1.78747218 | 8.79444356 | 0.50 | 21.70 | - |
| ABBN | $+7\sigma$ | 1.36582636 | 13.32014348 | 6.85209091 | 0.50 | - | 30.52 |
| ABBN | $-7\sigma$ | 0.85774461 | 5.61562724 | 7.96266239 | 0.50 | 27.43 | - |

Table B.4: Fitting parameters inverse statistics for the DJIA dataset

| $\rho$ | $\nu$ | $t_0$ | $\beta$ | $\alpha$ | $\tau^*_{-\lvert\rho\rvert}$ [days] | $\tau^*_{+\lvert\rho\rvert}$ [days] |
|---|---|---|---|---|---|---|
| $+5\sigma$ | 2.03958571 | 7.53374162 | 4.35134857 | 0.50 | - | 14.48 |
| $-5\sigma$ | 0.77866472 | -0.51224438 | 4.2947799 | 0.50 | 8.46 | - |